Blanka Sedlačíková

HISTORIE MATEMATICKÉ LINGVISTIKY

NADACE UNIVERSITAS V BRNĚ
AKADEMICKÉ NAKLADATELSTVÍ CERM V BRNĚ
ČESKÁ MATEMATICKÁ SPOLEČNOST
2012

Text © Blanka Sedlačíková, 2012
Vydaly Nadace Universitas v Brně, Akademické nakladatelství CERM v Brně a Česká matematická společnost, 2012

Tato kniha ani jakákoli její část nesmí být přetiskována, kopírována či jiným způsobem rozšiřována bez výslovného povolení vydavatele.

ISBN 978-80-7204-815-1 (Akademické nakladatelství CERM v Brně)

Obsah

Úvod **1**

1 Matematická lingvistika **5**
1.1 Kvantitativní lingvistika 7
1.2 Algebraická lingvistika 9
1.3 Počítačová lingvistika 12

2 Historie matematické lingvistiky **15**
2.1 Stručný historický nástin 15
2.2 Starověk a středověk 18
2.3 Pojem „frekvence“ 28
2.3.1 Frekvenční a konkordanční slovníky 30
2.3.2 Těsnopis 41
2.4 Viktor Jakovlevič Bunjakovskij 48
2.5 Andrej Andrejevič Markov 50
2.5.1 Život a dílo A. A. Markova 50
2.5.2 Hlásková statistika románu Evžen Oněgin 54
2.6 George Kingsley Zipf 62
2.6.1 První Zipfův zákon 63
2.6.2 Druhý Zipfův zákon 67
2.6.3 Třetí Zipfův zákon 68
2.6.4 Ostatní výsledky 68
2.7 Textologie 69
2.7.1 Stylometrie 69
2.7.2 Určování autorství 71
2.7.3 Stemma 79
2.8 Teorie informace 92
2.9 Glottochronologie 101
2.10 Jan Amos Komenský 104
2.11 Časopis Krok a počátky využívání matematiky v lingvistice v českých zemích 109
2.11.1 Jan Svatopluk Presl a časopis Krok 109
2.11.2 Hlásková statistika z roku 1831 114
2.12 Augustin Seydler a rukopisný spor 117

2.12.1 Historie rukopisných sporů 118
2.12.2 Život a dílo Augustina Seydlera 122
2.12.3 Časopis Athenaeum a Rukopisy 128
2.12.4 Seydlerův „matematický" důkaz nepravosti Rukopisů . 130
2.12.5 Další práce týkající se RKZ 137

Závěr **143**

Jmenný rejstřík **147**

Literatura **154**

Úvod

Matematická lingvistika je vědecká disciplína, která v současné době zaznamenává značný rozvoj. Její aplikace se staly běžnou součástí našeho života (zmiňme například jazykové databáze, elektronické slovníky, překlady textů pomocí počítačů, opravy překlepů v textových editorech aj.). Matematickou lingvistikou rozumíme pomezní disciplínu, která stojí na hranici mezi matematikou a jazykovědou (nově i umělou inteligencí a informatikou). Svými výsledky se do historie matematické lingvistiky zapsali ovšem nejen lingvisté či matematici, ale i fyzikové, literární vědci a další. Interdisciplinární charakter matematické lingvistiky poměrně komplikuje snahu shromáždit informace týkající se její historie – je třeba pátrat v celé řadě oborů (jazykověda, matematika, informatika, literární věda aj.). Podrobněji se historii matematické lingvistiky u nás věnuje pouze práce [6], další informace (zpravidla jen kusé) pak nacházíme roztroušeny v různých článcích, popularizačních příručkách, monografiích apod. (viz [49], [58], [66], [77]).

Cílem této publikace, která vznikla původně jako dizertační práce doktorského studijního programu Obecné otázky matematiky a informatiky na Přírodovědecké fakultě Masarykovy univerzity v Brně, bylo jednak shromáždit tyto izolované poznatky do jedné práce, jednak doplnit neúplné nebo stručné informace o nové skutečnosti. Ačkoliv je matematická lingvistika oborem poměrně mladým a její počátky klademe až do padesátých let 20. století, lze využívání matematických metod v lingvistice vysledovat hluboko do minulosti. První aplikace matematiky v lingvistice nacházíme dokonce již ve starověku a středověku. V tomto období se formovaly například počátky těsnopisu, významnou roli hrála kabala se svým kombinatorickým přístupem k jazyku apod. Proto stranou našeho zájmu nezůstaly ani tyto první pokusy, neboť mohou přinést nový pohled na problematiku matematické lingvistiky i jiných vědních oborů s ní souvisejících (např. kombinatorika, textologie aj.). Nemalá pozornost je konečně věnována těm výsledkům, kterými se do historie (či prehistorie) matematické lingvistiky zapsali čeští původci.

Celá práce je rozčleněna na dvě kapitoly. Nejprve je v první kapitole blíže představen obor matematická lingvistika – jeho vznik, cíle, metody, aplikace. Podrobněji je čtenář rovněž seznámen s jednotlivými částmi matematické lingvistiky, na něž bývá toto odvětví tradičně děleno, a to s lingvistikou kvantitativní (tou částí matematické lingvistiky, která využívá kvantitativní matematické metody), algebraickou lingvistikou (pracující s nekvantitativními mate-

matickými metodami, jako jsou například teorie grafů, kombinatorika, algebra, logika či teorie množin) a lingvistikou počítačovou (praktickou aplikací obou předchozích částí), i když hranice mezi jednotlivými částmi matematické lingvistiky nejsou samozřejmě vždy zcela zřetelné a dochází k jejich vzájemnému propojování. Například značný rozvoj počítačové lingvistiky v dnešní době by byl jen sotva možný bez využívání výsledků zbývajících dvou teoretických částí matematické lingvistiky, tedy lingvistiky kvantitativní a lingvistiky algebraické.

Těžiště práce představuje druhá kapitola, v níž jsou podle výše uvedených kritérií podrobně představeny některé partie z historie matematické lingvistiky. Nejprve je pozornost zaměřena na starověk a středověk, v nichž se většina aplikací matematiky na jazyk pohybovala víceméně na úrovni hry či mystiky. Přesto již zde můžeme vysledovat počátky takových matematických disciplín, jako je teorie kódování či kombinatorika. Blíže je ukázáno pojetí kombinatoriky v kabale a u významného středověkého myslitele Raimunda Lulla, jehož myšlenky ovlivnily například G. W. Leibnize. Důležitým mezníkem ve vývoji matematické (přesněji kvantitativní) lingvistiky je využívání pojmu frekvence (četnost) jazykového jevu (hlásky, písmena, slova, fonémy atd.). Ta hraje důležitou roli při sestavování frekvenčních a konkordančních slovníků, z nichž některé významné jsou v práci představeny blíže. Dále je ukázána řada dalších (především praktických) aplikací pojmu frekvence, a to například její využití při tvorbě těsnopisných systémů, optimalizaci klávesnice psacího stroje, šifrování, kódování (Morseova abeceda) apod. V následujících třech částech druhé kapitoly jsou představeni tři průkopníci využívání kvantitativních matematických metod při studiu jazyků. Ruský matematik V. J. Bunjakovský již v roce 1847 vůbec jako první upozornil na možnost využívání matematických metod v lingvistice. Další ruský matematik A. A. Markov svou teorií markovských řetězců nejen významně zasáhl do historie matematické lingvistiky, ale především do samotné matematiky. A konečně si představíme práci amerického lingvisty německého původu G. K. Zipfa, který se zabýval vztahy mezi frekvencí slov a jejich rankem a po němž bývá kvantitativní ligvistika někdy nazývána lingvistikou Zipfovou. V následujícím oddíle věnovaném textologii jsou podrobně rozebrány tři důležité textologické problémy, při jejichž řešení matematika sehrála důležitou roli. Konkrétně se jedná o tzv. stylometrii (neboli měření „stylu autora"), problematiku atribuce textu a sestavování stemmatu (čili schématu zachycujícího průběh textového procesu). Následuje oddíl zaměřený na teorii informace, v němž jsou vysvětleny některé základní pojmy tohoto odvětví, jako je například entropie a redundance, jež našly v matematické lingvistice rovněž své uplatnění. Dále je ukázána metoda nazývaná glottochronologie (též lexikostatistika), zajímavá aplikace matematiky, která se používá k určování stáří jazyků a která byla inspirována tzv. rozpadovým zákonem. Poslední tři oddíly druhé kapitoly se pak zabývají přínosem Čechů k „prehistorii" využívání kvantitativních metod v lingvistice. Nejprve se zmíníme o Janu Amosi Komenském a jeho práci *Janua linguarum reserata*. Dále si ukážeme jednu z prvních hláskových statistik vůbec, která byla uveřejněna roku 1831 v časopise *Krok*, prvním českém vědeckém časopise. A na závěr se zastavíme u matematika, fyzika a astronoma Augustina Seydlera a jeho práce [57], v níž se pomocí po-

čtu pravděpodobnosti snažil dokázat, že *Rukopis královédvorský* a *zelenohorský* jsou padělky.

Text této práce vychází z dizertační práce, kterou jsem obhájila v prosinci roku 2010. Na prvním místě bych proto chtěla vyjádřit poděkování svému školiteli, doc. RNDr. Eduardu Fuchsovi, CSc., který byl nejen mým rádcem a kritikem, ale který dokázal v pravý čas nalézt slova podpory a povzbuzení. Velký dík patří rovněž oponentům, doc. RNDr. Zdeňku Karpíškovi, CSc. a doc. RNDr. Zdeňku Pospíšilovi, Dr., za pečlivé přečtení této práce a za jejich podnětné připomínky. Dále bych ráda poděkovala Nadaci Universitas, která umožnila vydání této knihy. Největší dík ovšem patří mému muži Markovi a našim třem dětem, kteří po celou dobu stáli po mém boku a byli mou největší motivací. Protože si mé poděkování zaslouží ještě celá řada dalších lidí a nerada bych některé z nich opomenula, nebudu jmenovat raději nikoho. Přesto věřím, že všichni, jichž se to týká, vědí, že velké díky patří i jim.

Blanka Sedlačíková

Kapitola 1

Matematická lingvistika

Termín *matematická lingvistika* můžeme chápat dvěma způsoby. Jednak jím lze označovat hraniční disciplínu mající stejně blízko k matematice i jazykovědě, jednak jím rozumíme tu část lingvistiky, která využívá matematických metod. Zpočátku převládalo pojetí druhé, tj. matematická lingvistika byla považována za lingvistickou disciplínu využívající matematických metod. Později se však matematická lingvistika začala chápat jako pomezní disciplína stojící mezi matematikou a lingvistikou, nověji i *umělou inteligencí* (AI – *artificial inteligence*) a *informatikou*. Cílem takto pojímané matematické lingvistiky je exaktní popis přirozeného jazyka opřený o matematické metody.

V současnosti se matematická lingvistika standardně rozděluje na tři odvětví podle toho, kterých matematických metod využívá, a to na:

1) ***lingvistiku kvantitativní*** (podle převládající metody bývá též označovaná jako *lingvistika statistická*), která pro studium přirozeného jazyka využívá kvantitativní matematické metody, jako je například matematická statistika, teorie pravděpodobnosti aj.;

2) ***lingvistiku algebraickou*** (v současnosti je rozšířenější označení *formální lingvistika*) využívající nekvantitativní matematické metody (logika, algebra, teorie množin, teorie grafů aj.) a zabývající se formálním popisem gramatik a jazyků;

3) ***lingvistiku počítačovou*** (též *komputační*, starší název *lingvistika strojová*), která ke studiu jazyků využívá počítače (dříve např. děrnoštítkové stroje) a metody informatiky. Dnes se můžeme setkat také s termínem *počítačové zpracování přirozeného jazyka* (*natural language processing* – NLP). Blízko k počítačové lingvistice má samostatně stojící *language engineering*, pod kterým si můžeme představit algoritmické techniky pro popis přirozeného jazyka a různé softwarové nástroje vzniklé jejich implementací, jež pak hrají roli v problematice např. strojového překladu, automatického rozboru, uchovávání a vyhledávání informací, tvorbě gramatických korektorů apod. Od začátku 90. let 20. století se rovněž prudce

rozvíjí tzv. *korpusová lingvistika*, tj. ta část lingvistiky, která zkoumá jazyk za pomoci rozsáhlých souborů jazykových dat (korpusů) uchovávaných zpravidla elektronicky.

O matematické lingvistice jako o samostatné disciplíně se mluví zhruba od konce 50. let a začátku 60. let minulého století. Zpravidla se jako počátek označuje rok 1957, kdy se konal VIII. mezinárodní lingvistický kongres v Oslo. Jak ale zmiňuje W. Plath v [49], můžeme aktivitu v této oblasti vysledovat již dříve, zejména v Americe a v Evropě, ale i na Dálném východě[1]. Profesor Joshua Whatmough se ve své zprávě na konferenci v Oslu zmínil o semináři matematické lingvistiky na Harvardově univerzitě v roce 1955, čímž se tento obor vůbec poprvé objevil v učebním programu vysoké školy. V dalších letech se tyto kurzy matematické lingvistiky rozšířily vedle Harvardovy univerzity i na jiné univerzity v USA (například na Massachusettskou vysokou školu technickou, univerzity v Michiganu a Pennsylvanii), v Německu na univerzitu v Bonnu a v Sovětském svazu byly kurzy matematické lingvistiky vyučovány na moskevské a leningradské univerzitě. Kurzy kvantitativní lingvistiky byly zavedeny na univerzitách v Indianě a Kalifornii v USA. V roce 1958 vydal sovětský ministr školství nařízení, aby pro studenty matematiky a filologie zavedli rektoři univerzit v Moskvě, Leningradě, Gorkém, Saratově a Tomsku volitelné přednášky o strojovém překladu a matematické lingvistice.

Od počátku se projevovala v rámci matematické lingvistiky jistá terminologická nejednotnost (např. vedle označení kvantitativní lingvistika se setkáváme rovněž s označením lingvistika statistická). Tato nejednotnost odrážela vlastně hledání ve vymezení matematické lingvistiky jako celku, ale i jejích složek. Znovu se tato problematika oživila počátkem 60. let 20. století, kdy se utvářel pojem matematické lingvistiky i její terminologie[2]. Na IX. mezinárodním kongresu lingvistů v americké Cambridgi v roce 1962 vznesl námitky proti termínu matematická lingvistika (resp. i proti pojetí matematické lingvistiky, jak se v 60. letech 20. století vyvíjelo) dánský matematický lingvista H. Spang-Hansen[3]. Navrhl rozlišovat lingvistiku kvantitativní (*arithmetical and statistical*) a nekvantitativní (*non-quantitative*), které mohou být obě jak nestrukturální, tak strukturální (kde strukturální lingvistiku chápe jako axiomatický popis kvalitativní stránky jazykových jevů).

Matematická lingvistika se nesnaží nahradit lingvistiku „klasickou". To ani není vzhledem k povaze jazyka dost možné, neboť v jazyce neexistují pouze rysy kvantitativní, ale i rysy kvalitativní. Vladimír Šmilauer v [63] uvádí názorný příklad přibližující rozdíl mezi lingvistikou kvantitativní a kvalitativní. „*Protože*" a „*poněvadž*" jsou z hlediska lingvistiky kvalitativní spojky pod-

[1] V Japonsku vznikla v prosinci roku 1956 jako první na světě společnost *The Mathematical Linguistic Society of Japan* a v roce 1957 byl založen časopis *Mathematical Linguistics*. Více viz Köhler, Reinhard – Altmann, Gabriel – Piotrowski, Raimund G.: *Quantitative linguistics*. Walter de Gruyter, 2005.

[2] Viz Novák, P.: *K vytváření terminologie matematické lingvistiky*. Čs. terminologický časopis, č. 4, 1963, s. 234–237.

[3] Viz Spang-Hansen, H.: *Mathematical linguistics – a trend in name or in fact?* Preprints of Papers for the Ninth International Congress of Linguistics, Cambridge, Mass. 1962, s. 133–138.

řadicí příčinné. Z hlediska lingvistiky kvantitativní víme, že spojka „*protože*" se z celkového počtu 1 623 527 slovních výskytů doložených ve FSČ[4] vyskytuje v počtu 1224, a to v 63 textech (ze 75 možných) a ve všech fukčních stylech. Nejvíce dokladů má v beletrii (477 výskytů), v literatuře pro mládež (263 výskytů) a v dramatech (152 výskytů), v pořadí slov podle jejich častosti je na 134. místě. Naproti tomu spojka „*poněvadž*" má jen 374 dokladů v 19 dílech (textech), a to zejména v literatuře odborné (153 výskytů) a vědecké (107 výskytů), v básních není doložena vůbec (na rozdíl od spojky „*protože*", která se v poezii vyskytuje 489krát). Lze tedy na závěr říci, že „*protože*" je spojka běžná, „*poněvadž*" je méně obvyklá a vyskytuje se především v próze naukové.

Cílem matematické lingvistiky je exaktní popis přirozeného jazyka opřený o matematické metody. Matematická lingvistika se snaží hledat nové otázky či problémy, dále exaktními metodami potvrdit výsledky, kterých dosáhla jazykověda bez užití matematických metod. Může být rovněž pomocníkem jiných oborů – např. sdělovací technika, automatizace či samotná matematika. Je potřeba se ale vyhnout nekritickému zavádění matematického aparátu, které by vedlo k nic neříkajícím výsledkům. Zjištěná data je nutno vždy převést do řeči té disciplíny, na kterou je matematický aparát aplikován, zde tedy do řeči lingvistiky. Proto by měli spolupracovat navzájem lingvisté i matematici, popř. informatici.

Často se objevují námitky, zda je vůbec možné matematickými prostředky popsat tak složité systémy, jakými přirozené jazyky bezpochyby jsou. Pravdou ale je, „*že jakmile se podaří nějaký úsek skutečnosti přesně popsat a jeho obecné zákonitosti vědecky zachytit, pak může být zpracován i matematicky. Hranice těchto možností jsou tedy dány především dosavadním stavem té které vědy (. . .)*" ([58], s. 73). Největší problémy, se kterými se setkáváme při popisu přirozeného jazyka, jsou spojeny zejména s významovou vágností ve všech jazykových rovinách a dále s velkou mírou synonymie a homonymie. Proto je jedním z hlavních cílů matematické lingvistiky v současnosti tvorba nástrojů pro automatické odstraňování víceznačných interpretací jednotek jazyka (*automatická desambiguace*). Jeho vyřešení má pak dalekosáhlý význam pro rozvoj veškerých aplikací v oblasti matematické lingvistiky.

1.1 Kvantitativní lingvistika

Tímto termínem označujeme tu část matematické lingvistiky, které využívá kvantitativních matematických metod. Jsou to například statistika, matematická statistika či teorie pravděpodobnosti. Mocným impulsem se pro rozvoj kvantitativní lingvistiky staly práce C. E. Shannona a N. Wienera z konce 40. let 20. století, které položily základy matematické *teorie informace*. Tato teorie se zabývá kvantitativními vlastnostmi sdělovacích soustav, má však také závažné praktické aspekty týkající se sdělovací techniky (úspornost kódování, odolnost kódů proti chybám, šumu apod.). Národní jazyky jako nejdůležitější

[4] Běžně užívaná zkratka pro [25].

sdělovací soustavy lidské společnosti se tedy zcela přirozeně staly centrem pozornosti matematiků.

Podle převažující metody bývá v literatuře někdy lingvistika kvantitativní nazývána *statistická lingvistika*[5]. Jak již bylo zmíněno dříve, tato terminologická nejednotnost odráží hledání předmětu matematické lingvistiky jednak jako celku, jednak jejích jednotlivých složek, tedy i lingvistiky kvantitativní. Protože je statistika nejčastější matematickou metodou uplatňovanou v rámci kvantitativní lingvistiky, můžeme se setkat rovněž s termíny jako *fonologická statistika*, *morfologická statistika*, *lexikální statistika*, *stylistická statistika* apod., které vypovídají o tom, která část lingvistiky tyto statistické metody využívá.

Kvantitativní lingvistika má ze všech tří odvětví matematické lingvistiky (vedle lingvistiky algebraické a počítačové) nejdelší tradici. Metody matematické (zejména kvantitativní) se používaly při studiu jazykových jevů již dávno, i když se výrazu matematická (kvantitativní) lingvistika neužívalo. Podrobněji se prvními aplikacemi matematiky v lingvistice budeme zabývat v následující kapitole. Nyní si představíme alespoň ty nejdůležitější. Zásadní roli v historii kvantitativní lingvistiky hraje pojem *frekvence*. Zhruba od poloviny 19. století se začíná objevovat celá řada „výzkumů" založených na tomto pojmu. Jednalo se o jednoduché počty různých jazykových jevů a zpravidla byly tyto počty vyvolány potřebami praxe (výuka cizích jazyků, vytváření těsnopisných soustav, sestavování Morseovy abecedy apod.). Koncem 19. století pak začínají vznikat první frekvenční slovníky. Autorem toho úplně prvního je německý stenograf F. W. Käding[6]. Frekvencí anglických hlásek se pravděpodobně jako první lingvista vůbec zabýval Američan W. D. Whitney (1827–1894) a bývá tak považován za předchůdce kvantitativní lingvistiky. Na možnosti využití matematických metod v lingvistice upozornil již v roce 1847 ruský matematik V. J. Bunjakovskij a polský jazykovědec Jan Baudouin de Courtenay roku 1904 zdůraznil, že v jazykovědě by bylo vhodné využít nejen matematiky elementární, ale i matematiky vyšší. V první polovině 20. století se o rozvoj kvantitativní lingvistiky zasloužili zejména ruský matematik A. A. Markov a americký lingvista G. K. Zipf. Markov vydal v roce 1913 statistickou analýzu textu veršovaného románu *Evžen Oněgin*. Na základě statistického zkoumání výskytu ruských souhlásek a samohlásek a pravděpodobnosti, s jakou po sobě následují v textu, došel k závěru, že je možné předvídat pravděpodobnost jejich výskytu („*markovův proces*"). Šlo o první důslednou aplikaci matematické statistiky v jazykovědě a toto dílo probudilo zájem o vzájemnou spolupráci lingvistů a matematiků. Lingvista G. K. Zipf pak ve dvacátých a třicátých letech 20. století zkoumal frekvenci hlásek a upozornil na některé obecně platné vztahy vyskytující se v přirozených jazycích, které jsou založeny na pojmu frekvence (tzv. *Zipfovy zákony*).

Různá nahodilá zkoumání jazykových jevů založená na pojmu frekvence se ale objevují i na české půdě. Koncepčněji se jimi zabýval Martin Hattala

[5]Toto označení uvádí např. Plath, W.; Cohen, M.: *Sur la statistique linguistique*. Extrait des Conférences de l'Institut de linguistique de l'Université de Paris IX, année 1949, Paris 1950, s. 8–16; Herdan, G.; Muller, Ch.: *Initiation à la statistique linguistique*. Paris 1968.

[6]Käding, F. W.: *Häufigkeits Wörterbuch der Deutschen Sprache*. Steglitz 1897.

v polovině 19. století, který se pokoušel o nalezení pravidelností a zákonitostí v hláskové stavbě slovanských slov. V kapitole 2.11 si blíže představíme dvě statistiky, které vyšly již roku 1831 v časopise *Krok*. V roce 1886 vychází článek [57] matematika, fyzika a astronoma Augustina Seydlera, ve kterém se pomocí počtu pravděpodobnosti snaží dokázat nepravost *Rukopisů*[7]. Za předchůdce kvantitativní lingvistiky na našem území můžeme rovněž považovat některé členy Pražského lingvistického kroužku, zejména Viléma Mathesia, který se zabýval *potencionálností* jazykových jevů. V jazyce neplatí „absolutní zákony", ale v řeči každého jednotlivce existuje jisté kolísání „v určitých mezích a s určitou tendencí". Tyto tendence jsou pak podle Mathesia statisticky postižitelné. Kvantitativní lingvistika jako samostatná disciplína se u nás rozvíjela v rámci strukturalismu a je spojena s *pražskou školou* (J. Vachek, J. Krámský, B. Trnka, M. Těšitelová aj.).

Z tohoto stručného historického přehledu je zřejmé, že kvantitativní lingvistika měla praktický význam pro celou řadu oblastí. Zpočátku se jednalo o takové oblasti jako těsnopis, výuka pravopisu, šifrování, výuka cizích jazyků apod. Vladimír Šmilauer v [63] například uvádí zajímavou myšlenku, podle níž by se pro potřeby vyučování češtiny pro cizince sestavil katalog doporučených knih, které by byly srovnány podle míry *koncentrace slovníku*[8]. Až později kvantitativní lingvistika napomohla k lepšímu poznání různých jazykových jevů. Může být rovněž velkým pomocníkem v otázkách sporného autorství, překladatelství, srovnávací lingvistiky apod. A zejména v posledních letech se výsledky kvantitativní lingvistiky využívají při sestavování různých programů pracujících s přirozeným jazykem v rámci počítačové lingvistiky.

1.2 Algebraická lingvistika

Algebraickou lingvistikou rozumíme tu část matematické lingvistiky, která využívá nekvantitativních matematických metod, jako jsou algebra, teorie grafů, matematická logika, topologie, teorie množin či kombinatorika.

Algebraická lingvistika se začala formovat v druhé polovině 50. let 20. století zejména v souvislosti s potřebami *strojového překladu*. Ukázalo se totiž, že zatím jediným možným způsobem, jak odstranit nedostatky strojového i teoretického jazyka, je jeho důsledná formalizace. Autorství termínu *algebraická lingvistika* je připisováno Y. Bar-Hillelovi. U některých matematiků a lingvistů bývalého Sovětského svazu se můžeme setkat také s označením *teorie jazykových modelů*.

Společně s lingvistikou kvantitativní tvoří algebraická lingvistika teoretické obory matematické lingvistiky. Lingvistika počítačová je pak jejich praktickou aplikací.

Základ algebraické lingvistiky tvoří zejména tyto teorie: *generativní a transformační gramatika* N. Chomského, *rekognoskativní a kategoriální gramatika* Y. Bar-Hillela, *aplikačně generativní model jazyka* S. K. Šaumjana, *analytické*

[7] Podrobněji viz kap. 2.12.

[8] Veličina vyjadřující poměr prvních 50 nejčastějších (eventuálně 10 nejčastějších) plnovýznamových slov k délce textu. Viz též kap. 2.12.5.

modely jazyka (O. S. Kulaginová, R. L. Dobrušin, A. N. Kolmogorov, S. J. Fitialov, A. V. Gladkij, I. I. Revzin, C. Crăciun, S. Marcus, L. Nebeský aj.), *závislostní gramatika*, u nás např. *funkční generativní popis jazyka* (P. Sgall a kol.).

Významným mezníkem pro rozvoj algebraické lingvistiky bylo vydání knihy amerického lingvisty Noama Chomského *Syntactic structures* (1957)[9], ve které navrhl první variantu *mluvnice* nazývané *generativní* nebo *transformační* (možné jsou i oba názvy). Generativní znamená to, že jazyk je chápán jako tvůrčí proces, ve kterém se jednotlivé věty generují podle určitých pravidel, kterých existuje (stejně jako jednotek jazyka) omezený počet. Umožňují ale generovat neomezené množství vět. Při pokusu formalizovat popis gramatiky se snažil o jeho maximální zjednodušení. Zavádí proto termín *jádrových vět* (*kernel sentences*), to znamená základních jednoduchých vět, z nichž jsou všechny ostatní věty a souvětí odvozeny pomocí *transformačních pravidel*. Odtud tedy označení *generativní transformační mluvnice*. Protože první varianta mluvnice byla čistě formální (neuvažovala jazykový obsah), přišel N. Chomsky po četných námitkách v roce 1965 a letech následujících s variantou druhou, která měla již zakomponovanou složku sémantickou. Generativní transformační gramatika N. Chomského měla celou řadu následovníků, z nichž si pozornost zaslouží zejména Jerry Alan Fodor a Jerrold Katz, kteří se roku 1963 rovněž zasloužili o zahrnutí složky sémantické. O zakomponování pragmatické složky se poprvé v roce 1971 zmiňuje Yehoshua Bar-Hillel. U nás byl průkopníkem generativní gramatiky především Petr Sgall spolu se svými spolupracovníky, kteří vypracovali tzv. *funkční generativní popis jazyka*[10].

Opakem gramatiky generativní je *gramatika rekognoskativní*. Tento typ gramatiky vychází z konkrétní věty jazyka. Větu převádí na řetěz symbolů, odhaluje strukturu této věty a dále to, zda se jedná o gramaticky správnou větu příslušného jazyka. Nejznámějším příkladem rekognoskativní gramatiky je teorie vypracovaná na základě dřívějších systémů Yehoshuou Bar-Hillelem. Podle klíčového pojmu *kategorie* bývá tato gramatika nazývaná *kategoriální gramatikou*. Protože je kategoriálních gramatik celá řada (sám Bar-Hillel navázal zejména na práce K. Ajdukiewicze a H. B. Curryho), bývá Bar-Hillelova gramatika označována někdy termínem *kategoriální mluvnice identifikačního typu*[11].

V Sovětském svazu vytvořil na začátku 60. let 20. století S. K. Šaumjan gramatickou teorii s názvem *aplikačně-generativní model*, který v sobě zahrnoval prvky strukturalismu a generativní mluvnice. V Šaumjanově gramatické teorii jsou jazykové jednotky označeny symboly a odvozují se pomocí tzv. *aplikace* (metody matematické logiky, která se týká vztahů mezi symboly). Jazykověda by neměla zkoumat pouze jevy bezprostředně pozorovatelné, ale zejména hlubší souvislosti, které jsou za nimi skryté (roviny tzv. *logických konstruktů*). Důraz by měl být kladen na takové jevy, které jsou společné všem jazykům. Systém těchto univerzálních jevů nazývá Šaumjan *genotypický jazyk*. Naproti tomu *jazykem fenotypickým* rozumí ty jevy, které se vyskytují v jednotlivých

[9] Česky ji vydalo nakladatelství Academia v Praze roku 1966.
[10] Podrobně viz [58], [6].
[11] Více viz [58], [6].

přirozených jazycích. Jeho model se tedy skládá ze dvou částí, genotypické a fenotypické. První z nich je obecná teorie jazykových univerzálií a druhá je generativní mluvnice jednotlivých přirozených jazyků.

Generativní a rekognoskativní gramatiky byly zkoumány zejména na angličtině a ukázalo se, že dobře vyhovují pro popis jazyků s pevným slovosledem a jednoduchou morfologií, což angličtina právě je. Pro jazyky flexivní (zejména jazyky slovanské) jsou však tyto gramatiky nevyhovující. Zejména sovětští gramatici na tyto nedostatky upozorňovali a snažili se vytvořit obecnější formu gramatiky vyhovující typologicky různým jazykům. Tak vytvořili tzv. *analytické modely jazyka* jako protiklad k *modelům syntetickým* (rekognoskativní a generativní mluvnice). Zatímco v syntetických modelech vytváříme soubor gramaticky správných vět jazyka či zjišťujeme, které věty do tohoto souboru patří, v analytických modelech postupujeme opačně. Výchozím pojmem je pro nás soubor gramaticky správných vět jazyka. Za zakladatelku tohoto směru je považována O. S. Kulaginová, k rozvoji této teorie přispěli v 60. letech 20. století zejména matematici R. L. Dobrušin, A. N. Kolmogorov a lingvisté S. J. Fitialov, A. V. Gladkij a I. I. Revzin. Teorie analytických modelů je založena na teorii množin a jedná se o první využití teorie množin v lingvistice. Z bývalého východního bloku se o rozvoj tohoto modelu zasloužili rumunští vědci Solomon Marcus, jeden z nejvýznamnějších teoretiků analytické metody, a C. Crăciun. U nás se analytickou metodou zabýval především L. Nebeský.

Mezi další vědce, kteří se zasloužili o rozvoj algebraické lingvistiky v Sovětském svazu, patří například O. S. Achmanova, N. D. Andrejev, R. M. Frumkina, I. A. Melčuk, J. V. Padučeva aj.

Teorie množin byla využita vedle teorie analytických modelů rovněž v celé řadě různých lingvistických oblastí. V článku [49] můžeme najít několik takových zajímavých aplikací. Například Harary a Paper (1957) využili několika pojmů z teorie relací pro popis distribuce fonémů. Tento systém vedle strukturní charakteristiky vykazuje rovněž charakteristiky kvantitativní. Autoři zavádí indexy, které navrhují aplikovat na typologický výzkum distribuce fonémů. Lingvisté *glosematické školy* (např. Louis Hjelmslev) se pokusili vytvořit tzv. *kalkul nekvantitativních funkcí*, novou algebru zaměřenou zejména na popis „humanitních" materiálů, např. jazyka. Tato *glosematická algebra* používá převážně terminologie logické teorie tříd, ale s tím důležitým rozdílem, že některé známé pojmy, jako např. funkce a negace, jsou definovány neobvyklým způsobem. Kritika Ungeheuerem ukázala, že Hjelmslev a Uldall nepoužívají glosematické terminologie vždy důsledně a vývoji jejich algebry by prospěl propracovaný logický rámec.

Algebraická teorie gramatiky je oblast, kde se matematika a lingvistika spojují nejtěsněji. Teorie gramatiky má totiž celou řadu společných znaků s matematickou *teorií automatů*. Z tohoto hlediska lze gramatiku jazyka chápat jako soustavu pravidel, které vymezují všechny správně tvořené věty daného jazyka. Pravidla takové gramatiky je možno zachytit formálně podobně jako pravidla vymezující činnost automatu. Různé typy automatů a gramatik lze srovnávat. Ukazuje se ovšem, že jednoduché typy gramatik nejsou pro lingvistiku dostatečné. Podle dosavadních výsledků nejvíce vyhovují gramatiky transformační.

Stále ale není zcela jasné, zda vůbec lze v úplnosti formálně zachytit tak složitý systém, jakým bezesporu přirozený jazyk je. Zatím se zdá, že gramatiku jazyka (fonologie, morfologie, syntax) je možno formálními prostředky celkem úspěšně popsat, i když se jedná o oblast poměrně složitou. Dosud zatím vyčerpávajícímu formálnímu popisu odolává sémantika jazyka. Tento úkol je bezesporu náročný, ale vyžaduje jej sama praxe zejména v souvislosti s rozvojem výpočetní techniky a tvorbou různých počítačových programů. Ale to už se dostáváme k dalšímu odvětví matematické lingvistiky, a to k lingvistice počítačové.

1.3 Počítačová lingvistika

Je třetím tradičně rozlišovaným odvětvím matematické lingvistiky. Název *počítačová lingvistika* nám říká, že se jedná o počítačové zpracování jazyka, jež bylo prováděno zpočátku na jednoduchých děrnoštítkových strojích (podle nich také starší označení *strojová lingvistika*), později na složitých počítačích (odtud *počítačová* nebo také někdy *komputační lingvistika*).

Počítačová lingvistika se vyvíjí od konce padesátých let minulého století, a to zejména v souvislosti s rozvojem kybernetiky, výpočetní techniky, kvantitativní a algebraické lingvistiky a jiných hraničních oborů.

Připomeňme si postavení počítačové lingvistiky v rámci matematické lingvistiky: matematická lingvistika je tvořena dvěma teoretickými obory, a to lingvistikou kvantitativní a lingvistikou algebraickou. Počítačová lingvistika je potom jejich praktickou aplikací, proto se můžeme setkat i s termínem *aplikovaná matematická lingvistika*. Důležité je v tomto případě slovo „matematická", neboť názvu aplikovaná lingvistika by odpovídala oblast podstatně širší, a to nejrůznější aplikace jazykovědy – jazykové vyučování, kultura spisovného jazyka, uplatnění jazykovědných výsledků v jiných disciplínách (např. literární věda, historie) atd.

Mezi hlavní problémy řešené v rámci počítačové lingvistiky patří strojový překlad, automatický rozbor, uchovávání a vyhledávání informací, vytváření jazyků pro automatické programování, v současnosti pak zejména tvorba nástrojů pro počítačové zpracování přirozeného jazyka. Jako součást počítačové lingvistiky je chápána i *korpusová lingvistika*, která pracuje s *korpusy*, tj. rozsáhlými soubory jazykových dat.

Mohutně rozvíjet se začala počítačová lingvistika především v souvislosti se *strojovým překladem* (rozumíme jím převedení textu ze vstupního jazyka do jazyka výstupního, cílového, pomocí stroje). Od 50. let 20. století na přípravě strojových překladů pro různé jazyky pracovaly desítky pracovišť po celém světě (první pokusy byly provedeny v roce 1954 v USA a roku 1955 v SSSR). Nakonec se ukázalo, že se jedná o úkol značně složitý. Pro účel překladu nestačí totiž jen zvládnutí mluvnické stavby, ale je třeba zakomponovat i sémantiku přirozených jazyků. Dosud nebyla sémantika zpracována natolik dostatečně, aby mohla být zachycena formálními metodami. Navíc rozsáhlá slovní zásoba a složitá stavba jazyků kladou značné nároky na paměť počítače i na dobu zpracování. Poměrně zdařilé se jeví automatické překladače určité speciální skupiny

textů (technické texty apod.), kde je poměrně úzká a stabilní terminologie a kde nezáleží na stylistických odstínech. Nejlepší výsledky zatím vykazují překladače poloautomatické, ve kterých jsou sporné jevy dodatečně redigovány člověkem.

Vyhledávání informací je úkol poměrně jednoduchý a poradily si s ním už děrnoštítkové stroje. *Automatické informování o obsahu* je záležitost mnohem náročnější. Kvalitní referát by měl být totiž jasný a srozumitelný, měl by být stručným obsahem práce, uvádět hlavní výsledky práce, její přínos a hodnocení. V zásadě se využívají dva postupy. První z nich využívá statistiky. Pro daný obor se určí soubor nejdůležitějších termínů a do referátu se automaticky přejímají ty věty, které těchto termínů obsahují nejvíce. Ze zkušenosti ale víme, že termíny nejčastěji se vyskytující nemusí být vždy nejvýznamnější a že důležité pasáže nemusí takový termín obsahovat vůbec. Druhý postup je založen na hledisku sémantickém, ale zde narážíme na ještě větší problémy, neboť bez důsledné formalizace stránky sémantické lze pracovat jen velmi obtížně. Nejnáročnějším stupněm z postupů zpracovávajících informace pomocí počítače jsou *informační jazyky*, tzn. ucelené teoretické systémy sloužící k ukládání a vyhledávání informací. Snaží se řešit řadu značně složitých úkolů – například je žádoucí, aby byl počítač schopen zodpovědět určitou otázku týkající se problematiky daného oboru na základě souboru poznatků z literatury tohoto oboru; aby odpovídal na otázky kladené v přirozených jazycích; doplňoval informace, které nejsou v textu explicitně vyjádřeny; zpracovával literaturu v různých jazycích apod.

Velkým pomocníkem nejen při tvorbě strojových překladů by se mohla stát *korpusová lingvistika*, která zaznamenává v současnosti obrovský rozvoj. Korpusovou lingvistikou rozumíme tu část počítačové lingvistiky, která se zabývá tvorbou a využitím jazykových *korpusů*, tzn. souborů jazykových dat. Tento soubor může mít podobu textů, které jsou zachyceny na papíře (nejčastěji v podobě excerpt), nebo může mít podobu elektronickou, kdy jsou jazyková data uložena v počítači. Korpus lze potom chápat jako rozsáhlý, vnitřně uspořádaný a ucelený soubor jazykových dat, která jsou elektronicky uložena, zpracována a přístupna. Ačkoliv se různých jazykových korpusů uchovaných na papíře užívalo velmi dávno (vzpomeňme na J. A. Komenského, kterému při požáru v Lešně shořel celý jeho rozsáhlý materiál na připravovaný latinsko-český a česko-latinský slovník), o skutečné korpusové lingvistice mluvíme až v souvislosti s rozvojem počítačové techniky.

Vztah počítačové lingvistiky a lingvistiky klasické je oboustranný. Na jedné straně pomáhá jazykověda ostatním oborům (zejména matematice a logice) řešit úlohy, které s lingvistikou přímo nesouvisí (např. programovací jazyky). Na straně druhé lze počítače využívat přímo v lingvistické práci (z těch jednodušších například třídění a statistické zpracování rozsáhlých souborů jazykových dat, z těch složitějších strojový překlad, automatické generování vět podle určitých pravidel, automatický rozbor větných struktur, automatické vyhledávání textových informací – rešerše aj.). Oblast počítačové lingvistiky je ale širší, zahrnuje totiž i vlastní jazykovědný výzkum, zejména z oblasti algebraické lingvistiky (počítače se využívají k prověřování různých gramatik). Spolu s lingvistikou algebraickou a kvantitativní klade počítačová lingvistika

otázky, které si tradiční lingvistika neklade, a napomáhá tak odhalovat nové stránky přirozených jazyků.

Kapitola 2

Historie matematické lingvistiky

Počátky matematické lingvistiky klademe do 50. let 20. století, jedná se tedy o vědeckou disciplínu poměrně novou. Náznaky využívání matematických metod v lingvistice ale nacházíme již mnohem dříve, první stopy můžeme vysledovat dokonce už ve starověku. V této kapitole, která je těžištěm celé knihy, budou čtenářům představeny některé zajímavé „aplikace" matematiky na jazyk z tohoto dlouhého prehistorického období (tedy z období před ustanovením matematické lingvistiky jako samostatné disciplíny). Vybrány budou takové aplikace, které v literatuře týkající se matematické lingvistiky nejsou uváděny vůbec nebo jsou zmiňovány jen velmi stručně bez bližšího objasnění. Dále bude čtenář seznámen s osobnostmi, metodami či pojmy, které sehrály důležitou roli při utváření matematické lingvistiky (a to jednak z období ještě před ustanovením matematické lingvistiky jako samostatné disciplíny či těsně po jejím vzniku). A konečně v závěru kapitoly se zmíníme o počinech, které jsou spjaty s českým kulturním prostředím a které v historii matematické lingvistiky zastávají rovněž nezanedbatelné místo.

2.1 Stručný historický nástin

V této kapitole si shrneme běžně uváděné informace z historie matematické lingvistiky.

Vůbec poprvé nacházíme zmínku o využívání matematiky v jazyce ve starověku u starých Hindů, kteří z náboženských důvodů počítali slova v textu posvátné *Rgvédy*. Ve středověku byly velice oblíbené *obrazové básně*, které kladly důraz na mystiku čísla. Od 13. století se v Evropě začínají objevovat tzv. *spekulativní gramatiky*, které vedle popisu jazyka uplatňují i postupy filozofické, ale hlavně logické, založené na aristotelovské logice (Roger Bacon). Odtud je blízko k myšlence hledání tzv. *univerzálního jazyka* (Raimundus Lullus, John

Webster, John Wilkins, Francis Lodwick, Jan Amos Komenský aj.). Tito učenci často pracují s kombinatorickými pojmy a různými mechanickými pomůckami založenými na principu soustředných otáčivých kruhů. Ve většině případů jsou ale tyto středověké práce ovlivněny magií a astrologií, hlavní roli hraje *kabala*. Ta se rovněž zasloužila o vznik *steganografie* (tj. tajných písem), podoboru kryptografie, na počátku 16. století. Od 17. století pak začíná převládat čistě kombinatorický přístup k jazyku (John Wilkins, G. W. Leibniz). Racionalismus v jazykovědě se projevil sestavováním tzv. *filozofických gramatik*, platících obecně pro libovolný jazyk. Tyto snahy významně ovlivnily moderní lingvisty, např. Noama Chomského (více viz kap. 1.2), zakladatele *generativní a transformační mluvnice*. Podrobněji se aplikacím matematiky na jazyk v období starověku a středověku věnuje kapitola 2.2.

Zhruba od 19. století se v lingvistice začíná hojně uplatňovat využívání pojmu *frekvence* (neboli *četnost*). Zpočátku se jednalo o jednoduché počty různých jazykových jevů, které byly zpravidla vyvolány potřebami praxe (sestavování kazet s písmeny pro tiskaře, optimalizace klaviatury psacího stroje, Morseova abeceda, vytvoření optimální těsnopisné soustavy, zefektivnění jazykového vyučování, tvorba frekvenčních slovníků apod.). Podrobněji je pojem *frekvence* představen v kapitole 2.3.

Prvním, kdo zdůraznil využití matematických metod v lingvistice, byl s největší pravděpodobností ruský matematik Viktor Jakovlevič Bunjakovskij v roce 1847 (více viz kap. 2.4). V literatuře o matematické lingvistice bývá jako jeden z prvních vědců, který se zabýval kvantitativními jevy v jazyce, uváděn americký jazykovědec W. D. Whitney (1827–1894). V [6] na straně 126 čteme, že se „*jako první lingvista vůbec zabýval frekvencí anglických hlásek a stal se tak předchůdcem dnešní statistické či kvantitativní lingvistiky.*" V historii kvantitativní lingvistiky ale můžeme nalézt práce mnohem starší, než jsou Whitneyho výzkumy z roku 1874. Zmiňme alespoň práce německých vědců Ernsta Förstemanna (1822–1906) a Augusta Schleichera (1821–1868). V kapitole 2.11 si představíme blíže dvě statistiky (hláskovou a lexikální), které vyšly dokonce již v roce 1831 v prvním českém vědeckém časopise *Krok*. O vážnějších pokusech zavést do lingvistiky kvantitativní hledisko můžeme hovořit až v období přelomu 19. a 20. století, což pravděpodobně souvisí se systematickým pronikáním matematiky i do jiných vědních oborů. Na konci 19. století mladogramatik Herman Paul chápe jazyk jako statistický průměr všech jazykových projevů jednotlivců a podobné kritérium zavádí i do fonetiky, kde invariantní hlásku považuje za statistický průměr jejích možných artikulací. Polský člen kazaňské školy Jan Baudouin de Courtenay (1845–1929) poukazuje na vhodnost využití v jazykovědě nejen metod elementární, ale i vyšší matematiky. Mezi osobnosti, které měly zásadní význam pro rozvoj matematické lingvistiky, patří v první polovině 20. století ruský matematik A. A. Markov (podrobně viz kap. 2.5). Ten aplikací statistických metod na text románu *Evžen Oněgin* od A. S. Puškina položil základy tzv. *markovských řetězců*. Další významnou osobností tohoto období je americký vědec G. K. Zipf, kterému je věnována kapitola 2.6 a který bývá považován za zakladatele kvantitativní lingvistiky. Jako první systematicky zkoumal a uveřejnil vztahy mezi frekvencí slov a jejich rankem

– tzv. *Zipfovy zákony*[1]. Později byly Zipfovy modely pojmově a matematicky vylepšeny a rozšířeny Benoîtem Mandelbrotem, světově uznávaným tvůrcem *fraktální geometrie*. U nás se aplikací Zipfových zákonů na češtinu zabývala Marie Těšitelová. Pro zajímavost zmiňme na závěr práci polského antropologa J. Czekanowského[2], který matematickými metodami hledal důkazy o původním slovanském osídlení v oblasti Odry a Visly a jehož myšlenky se staly základem *taxonomické metody* sestavování *stemmatu* v textologii (blíže viz kap. 2.7).

Počátky matematické lingvistiky jako samostatné vědní disciplíny klademe do 50. let 20. století, zpravidla se vymezuje její vznik rokem 1948, kdy se konal *VI. mezinárodní lingvistický kongres* v Paříži. Zde byl totiž ustaven zvláštní výbor (*Committee on Quantitative Linguistics*), který se měl věnovat problémům kvantitativní lingvistiky. Jednatelem tohoto výboru byl zvolen český lingvista B. Trnka. Kromě toho bylo stanoveno množství témat (z lexika a gramatiky), která měla být zpracována analogicky pro různé jazyky (zejména pro potřeby typologického studia). Ačkoliv se kvantitativní lingvistika v řadě zemí v tomto období sice již rozvíjela, dělo se to bohužel spíše nekoordinovaně, a proto tato snaha zůstala ne zcela vyslyšena. Na vině je možná i to, že různé jazyky mají různé předpoklady pro využívání kvantitativních metod. Do tohoto období řadíme i vznik *teorie informace* (viz kap. 2.8), která se v nově vznikající kvantitativní lingvistice taktéž výrazně uplatnila. V 50. letech 20. století vznikla tzv. *glottochronologie* (neboli *lexikostatistika*), zajímavá aplikace matematiky v jazykovědě, která slouží k určování stáří jazyků. Podrobněji bude osvětlena v kapitole 2.9.

Podíváme-li se na uplatnění matematických metod při studiu jazyka v českých zemích, a zdůrazněme, že se jednalo zpočátku pouze o metody kvantitativní, můžeme s jistou mírou volnosti říci, že jako první pravděpodobně aplikoval tyto metody při studiu slovní zásoby již náš velký pedagog J. A. Komenský, kterému je věnována kap. 2.10. Ten ve svém díle *Janua linguarum reserata* ukázal, jak je možno znalosti o počtu slov v jazyce využít k efektivnímu rozvíjení slovní zásoby u žáků[3]. Mimo to se zabýval problematikou hledání *univerzálního jazyka* a seznámil se při svém pobytu v Anglii s naukou o tajných písmech čili *steganografii*. Velkou pozornost si zaslouží ojedinělý počin matematika a fyzika Augustina Seydlera, který roku 1886 aplikoval počet pravděpodobnosti na studium jazykových jevů za účelem potvrzení nepravosti tzv. *Rukopisů*[4] (více viz kap. 2.12). Zájem o využívání kvantitativních metod v lingvistice můžeme rovněž vysledovat v pražské lingvistické škole před druhou světovou válkou (základy této školy byly položeny v 30. letech 20. století). Za připomenutí

[1] Zejména viz *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Boston 1935; *Human Behavior and the Principle of Least Effort.* Cambridge 1949.

[2] Viz *Wstęp do historii Słowian.* Lvov 1927; nové vydání *Polska synteza slawistyczna w perspektywie ilościošej.* Warszawa 1947.

[3] Podrobněji viz [65].

[4] Viz [56], [57].

stojí práce V. Mathesia[5], B. Trnky[6] (řešil rovněž i obecné otázky kvantitativní lingvistiky, včetně terminologických), dále Josefa Vachka[7] aj. I když to byli zaměřením především anglisté a pracovali tedy zejména s anglickým jazykem, výsledky své práce srovnávali s materiálem českým, popřípadě s materiálem z jiných jazyků. Jejich studium se týkalo hlavně oblasti fonologické a lexikální. Vedle lingvistů položili základy české matematické lingvistiky i pedagogové (podobně jako jinde ve světě). Tak například první frekvenční slovník češtiny [25] vznikl z podnětu pedagoga Václava Příhody a lingvisty bohemisty Vladimíra Šmilauera.

2.2 Starověk a středověk

První stopy užití kvantitativních metod v jazyce můžeme najít již u starých Hindů. Ti počítali z náboženských důvodů slova v textu posvátné *Rgvédy*, nejstarším z textů véd, který vznikal v letech 1500 až 1000 př. n. l. a byl sepsán někdy v letech 800 až 600 př. n. l. Jedná se o modlitební hymny, hymnické písně, které prostřednictvím staroindických světců (ršiů) vnukli lidem samotní bohové, aby je poučili o tom, jak je správně uctívat. Védy zpočátku předávali bráhmani v ústní podobě z pokolení na pokolení. Aby při absenci jakéhokoliv písemného záznamu nedošlo k sebemenším změnám v textu Rgvéd, byl vytvořen důmyslný a velmi složitý systém určený k jejich uchování prostřednictvím ústního podání (védská metrika a časoměrný verš), který pracoval s přesnými počty slabik. Díky tomuto propracovanému systému se nám texty véd dochovaly v nezměněné podobě až do doby, kdy mohly být písemně zaznamenány.

Ve středověku se objevují aplikace na úrovni mystiky slov, čísel a obrazců. Můžeme sem zařadit středověké *obrazové básně* (*carmen figuratum*), ke kterým má blízko například Rabelaisův kaligraf *Božská láhev* z *Gargantuy a Pantagruela* (vydáno poprvé 1564). Odtud vede vztah přes Apollinairovy *kaligramy*, český *poetismus*[8], *lettrismus* Isidora Isoua a *vizuální poezii* až do dneška (například tvorba počítačových obrazců vytvářených programátory). Ačkoliv tyto aplikace spadají do různých časových období, společná je jim právě již zmiňovaná mystika či hra.

Z předešlé kapitoly víme, že kvantitativní metody hrají významnou roli v *kabale* („*qabbalah*" lze přeložit jako „tradice"), mohutném proudu hebrejského mysticismu, který považuje stvoření světa za jazykový jev. Nesprávně byla kabala v křesťanském světě spojována s černou magií, dokud nebyla přehodnocena humanisty. Kabala vychází z tradice výkladu *Tóry* (tj. *Pentateuchu*

[5] *O potenciálnosti jevů jazykových.* Věstník Královské české společnosti nauk, třída historická, 1911; nověji ve sborníku *U základů pražské jazykovědné školy* (vyd. J. Vachek). Praha 1970, s. 5–34.

[6] Viz [75].

[7] *Poznámky k fonologii českého lexika.* LF, 67, 1940, s. 395–402.

[8] Např. báseň *Počitadlo* a *Objevy* ze sbírky *Na vlnách TSF* od Jaroslava Seiferta, optická báseň *Adié* ve sbírce *Pantomima* od Vítězslava Nezvala aj.

čili *Pěti knih Mojžíšových*) a *Talmudu* a je považována zejména za metodu četby a interpretace posvátného textu. Výchozím bodem, s kterým kabalista pracuje, je svitek *Tóry*. Za slovy psané tóry je třeba nalézt věčnou tóru, jež tu byla již před Stvořením a kterou Bůh svěřil andělům. V [11] na straně 30 čteme: „*Podle některých kabalistů byla tóra původně napsána černým ohněm na bílý oheň a v okamžiku Stvoření stála před Bohem jen řada písmen dosud nespojených do slov. Kdyby nebylo Adamova hříchu, písmena by se spojila jinak a vytvořila by jiný příběh. Proto svitek tóry neobsahuje samohlásky, interpunkční znaménka ani přízvuky, neboť tóru původně tvořila jen hromada neuspořádaných písmen. Po příchodu mesiáše Bůh nynější kombinaci písmen opět zruší nebo nás naučí číst text podle jiného uspořádání.*“

Teozofická kabala (jeden ze směrů kabalistické tradice) se snaží za slovy posvátného textu odhalit odkazy na deset *sefirot*, které představují deset hypostazí božstva. Kabalista přistupuje k textu tóry jako k symbolickému aparátu, který vypovídá o mystické a metafyzické skutečnosti. Slouží mu k tomu tři základní postupy: *notarikon*, *gematrie* a *temura*.

1) *Notarikon* je metoda akrostichu (počáteční písmena několika slov dávají slovo jiné), způsob šifrování či dešifrování textu. Tento postup byl oblíbenou hříčkou celé pozdně antické i středověké literatury. Od středověku se pak rozšiřují magické obřady nazývané *ars notoria*. Akrostich má podle kabalistů odhalit mystickou příbuznost slov.

2) *Gematrie* zase pracuje s faktem, že v hebrejštině jsou čísla označována písmeny abecedy. Každé slovo má tedy jistou numerickou hodnotu, kterou získáme součtem čísel představovaných jednotlivými písmeny. Hledají se pak slova se stejnou numerickou hodnotou a s odlišným významem a zkoumají se analogie mezi příslušnými věcmi či myšlenkami. Například numerická hodnota písmen *JHVH* (čtyři písmena označující jméno Boží) je 72, proto kabalistická tradice hledá 72 Božích jmen. Podobně Mojžíšův *had* je předobrazem *mesiáše*, neboť *had* i *mesiáš* mají numerickou hodnotu 358.

3) *Temura* je pak umění permutace, tj. záměny písmen, umění anagramu. Možností permutací je navíc v jazyce, do něhož lze libovolně vkládat samohlásky, podstatně více. Například Moše Cordovero řeší, proč se v *Deuteronomiu* (pátá kniha *Starého zákona*, poslední kniha *Tóry*) objevuje zákaz nosit smíšené oděvy z vlny a lnu. Usuzuje, že se původně tatáž písmena kombinovala jinak a tvořila výraz, který varoval Adama, aby neměnil původní světelné roucho za šat z hadí kůže, který ztělesňoval démonickou moc. Abulafija čtyři písmena tetragramu JHVH vokalizuje všemi možnými způsoby a dostává čtyři tabulky po 50 kombinacích. Eleazar Ben Jehuda z Wormsu každé písmeno tohoto tetragramu vokalizuje dvěma samohláskami (z šesti možných), čímž se počet kombinací dále zvyšuje. Temura je pro kabalisty obrovským zdrojem inspirace, neboť to není jen metoda četby, ale postup, jímž stvořil Bůh svět. Tento princip je jasně popsán v *Sefer Jecira* (*Kniha stvoření*). Podle tohoto krátkého

pojednání, které vzniklo někdy ve 2. až 6. století našeho letopočtu, Jahve stvořil svět z dvaatřiceti kamenů neboli dvaatřiceti cest moudrosti, kterým odpovídá deset sefirot a dvaadvacet písmen abecedy. „*Dva kameny staví dva domy, tři kameny staví šest domů, čtyři kameny staví dvacet čtyři domů, pět kamenů staví sto dvacet domů, šest kamenů staví sedm set dvacet domů, sedm kamenů vytvoří pět tisíc čtyři sta domů. Počítáš-li tak stále dál, dojdeš k číslům, která již ústa nejsou s to vyslovit a ucho je nedokáže pochytit.*“[9]

Kniha stvoření (*Sefer Jecira*) hovoří tedy o faktoriálním počtu. Můžeme říci, že kabala naznačuje, že z konečné abecedy lze sestavit závratný počet kombinací. Nejdále toto kombinatorické umění dovedl ve 13. století Abraham Abulafia (1240–1295) ve své *Kabale jmen*. V *Kabale jmen* se předříkávají Boží jména ukrytá v textu *Tóry* a pracuje se s nejrůznějšími kombinacemi hebrejské abecedy. „*Četba prostřednictvím permutací má extatické účinky.*“[10] Odtud také označení *extatická kabala*. Podle *extatické kabaly* je jazyk univerzem o sobě a struktura jazyka odpovídá struktuře skutečnosti. Podle Abulafii je hebrejština prajazykem lidského rodu. Dvaadvacet písmen hebrejské abecedy představuje ideální hlásky, z kterých vzniklo všech sedmdesát existujících jazyků. Pokud se v jiných jazycích vyskytne větší počet samohlásek, je to způsobeno drobnými změnami ve výslovnosti dvaadvaceti základních písmen (v lingvistické teorii řečeno: jedná se o alofony základních fonémů). Abulafia rovněž tvrdí, že 22 písmen hebrejské abecedy představuje všechny zvuky přirozeně vydávané artikulačními orgány. Zrození jednotlivých jazyků umožňují různé kombinace písmen – známe-li zákony kombinatoriky, máme klíč k formování všech jazyků.

V období humanismu a renesance, typickém svým okouzlením minulostí, převažovalo magicko-astrologické pojetí kosmu. Podle něj nebeská tělesa ovlivňují pozemské věci. Poznáním planetárních zákonů lze tyto pozemské jevy předvídat a usměrňovat. Mezi makrokosmem univerza a mikrokosmem člověka existuje tedy příbuznost a na toto silové pole je možno působit astrální magií. Magické praktiky se pak provádí prostřednictvím slov či jiných forem znaků, existuje tedy jazyk, kterým můžeme poroučet hvězdám. Pod vlivem *kabaly* se věřilo, že shoda, která vládne mezi pozemskými předměty a předměty nebeskými, panuje také mezi jmény. Podle Heinricha Cornelia Agrippy z Nettesheimu (1486–1535) dával Adam věcem jména s ohledem na tyto vlivy, a proto „*tato jména obsahují zázračné síly označovaných věcí*“ (viz [11]). Písmo Židů musí být tedy považováno za nejposvátnější ze všech vzhledem k dokonalé shodě mezi písmeny, věcmi a čísly.

Další náznaky matematického přístupu v lingvistice nacházíme v Evropě ve 13. století, a to v souvislosti s tzv. *aristotelovským racionalismem*. Ten byl vyvolán zásluhou Tomáše Akvinského (1225–1274), na jehož podnět bylo přeloženo Aristotelovo (384–322 př. n. l.) dílo *Organon* do latiny. *Organon* (v překladu „nástroj“) je vlastně soubor spisů zabývajících se pravidly („nástroji“) správ-

[9] Viz [11].
[10] Viz [11], str. 33.

ného myšlení, usuzování a argumentace. Aristoteles zde položil základy tzv. *aristotelovské logiky*, z níž formalizováním později vznikla *logika predikátová*. V jazykovědě vzniká v tomto období celá řada gramatik, které se nespokojovaly s pouhým popisem jazyka, ale uplatňovaly postupy obecně filozofické a logické (tzv. *spekulativní gramatiky*). Jednou z nich je například *Summa grammatica* Rogera Bacona (1214–1294), podle něhož je gramatika u všech jazyků stejná, existují jen nepodstatné (akcidentální) variace.

Zde můžeme vysledovat myšlenku tzv. *univerzálního jazyka*, která byla blízká i Baconově současníkovi Ramónu Llullovi (polatinštěno na Raimundus Lullus). Tento Katalánec se narodil na Mallorce roku 1232 (nebo 1235) a zemřel v roce 1316. V té době byla Mallorka místem, kde se střetávala kultura křesťanská, židovská a islámská, což mělo velký vliv na Lullovu tvorbu, neboť většinu ze svých 280 děl napsal arabsky a katalánsky. Po bouřlivém mládí a mystické krizi vstoupil jako terciář do františkánského řádu. Zde vytvořil kolem roku 1275 svůj projekt *Ars magna* („velké umění"), systém dokonalého filozofického jazyka, kterým by bylo možno obrátit nevěřící na pravou víru. Idea svornosti mezi lidmi různých ras a vyznání byla totiž důležitou součástí františkánského myšlení. Tento jazyk měl být jazykem univerzálním, neboť univerzální byla jednak kombinatorika, strukturující jeho výrazový plán, jednak systém myšlenek společných všem národům, které Lullus zahrnul do plánu obsahového. Ale *Ars magna* je univerzální rovněž proto, že využívá písmen abecedy a obrazců, takže je srozumitelná i analfabetům. Jen pro zpestření uveďme, že podle legendy byl Lullus umučen Saracény, když k nim přišel vyzbrojený svou *Ars magna* jako spolehlivým přesvědčovacím prostředkem[11].



Podívejme se nyní na jeho dílo podrobněji, neboť ačkoliv bývá zpravidla zařazováno do dějin logiky, jeho myšlenky kombinatorické sehrály důležitou roli v historii matematiky a značně ovlivnily další učence (zejména G. W. Leibnize v jeho spise *Ars combinatoria*).

Lullova *Ars magna* pracuje s abecedou o devíti písmenech (B, C, D, E, F, G, H, I, K) a čtyřmi obrazci (viz dále). Lullus vytváří seznam šesti celků o devíti prvcích (*tabula generalis*), představujících obsahy, které lze v určitém pořadí přiřadit devíti písmenům (viz tab. 2.1). Tyto celky jsou:

1) absolutní principy (též „božské atributy"),

2) relativní principy,

3) typy otázek,

[11] Viz též Mačák, K.: *Poznámky k formování kombinatoriky v 16. a 17. století*. In: Matematika v 16. a 17. století (ed. Bečvář, J. – Fuchs, E.), Prometheus, Praha 1999.

4) subjekty,

5) ctnosti,

6) neřesti.

	PRINCIPIA ABSOLUTA	PRINCIPIA RELATIVA	QUESTIONES	SUBJECTA	VIRTUTES	VITIA
B	Bonitas	Differentia	Utrum?	Deus	Iustitia	Avaritia
C	Magnitudo	Concordantia	Quid?	Angelus	Prudentia	Gula
D	Aeternitas	Contrarietas	De quo?	Coelum	Fortitudo	Luxuria
E	Potestas	Principium	Quare?	Homo	Temperantia	Superbia
F	Sapientia	Medium	Quantum?	Imaginatio	Fides	Acidia
G	Voluntas	Finis	Quale?	Sensitiva	Spes	Invidia
H	Virtus	Majoritas	Quando?	Vegetativa	Charitas	Ira
I	Veritas	Aequalitas	Ubi?	Elementativa	Patientia	Mendacium
K	Gloria	Minoritas	Quomodo? Cum quo?	Instrumentativa	Pietas	Inconstantia

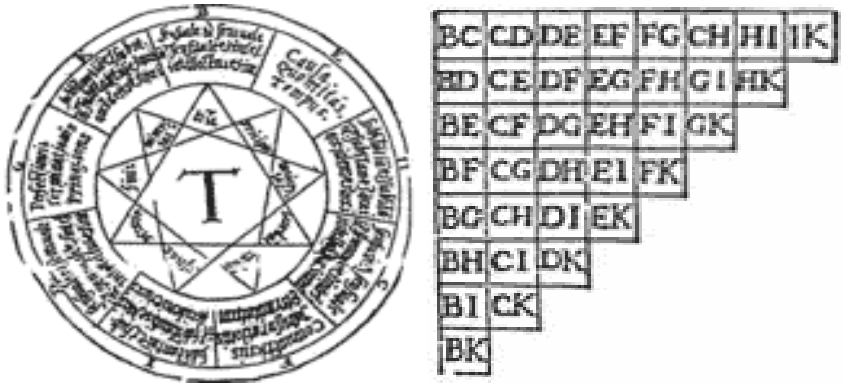
Tabulka 2.1: Tabula generalis (tabulka principů podle Raimonda Lulla)

Lullus upřesňuje, že první skupina (principia absoluta, tj. božské atributy) představuje subjekty predikace, dalších pět skupin označuje predikáty. To vysvětluje to, proč jeho kombinatorika v mnoha případech nepovoluje změnu pořadí.



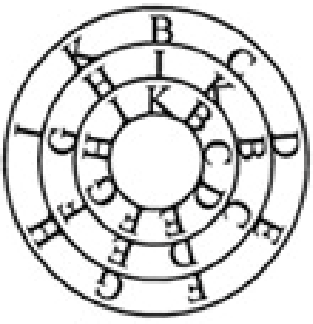
První obrazec. Zde přiřadil Lullus devíti písmenům abecedy božské atributy spolu s příslušnými adjektivy a kreslí všechny možné spojovací čáry mezi těmito principy. Vznikají tak věty typu: *Bonitas magna (est). Dobro je veliké. Magnitudo gloriosa (est). Velikost je slavná.* Tyto principy mají formu substantiv, jestliže plní funkci podmětu, a formu adjektiv, vystupují-li jako přísudek věty. Každá čára tohoto obrazce se může číst tedy v obou směrech, takže vedle věty „*Bonitas magna.*" (*Dobro je veliké.*) vznikne například věta „*Magnitudo bona.*" (*Velikost je dobrá.*). Lullův obrazec by měl umožňovat tvorbu pravidelných sylogismů. Kombinace typu BB nebo CC neuvažuje, neboť neumožňují nalézt střední termín sylogismu. Na obrázku je zakresleno celkem 36 čar, které lze číst dvěma způsoby, získáváme tedy (řečeno dnešní terminologií) celkem 72 dvoučlenných variací bez opakování z devíti prvků.

Druhý obrazec. Nepoužívá kombinatoriku, slouží pouze jako vizuálně-mnemotechnická pomůcka k zapamatování stálých vztahů mezi různými typy relací a mezi různými druhy jsoucen.



Třetí obrazec. Zde Lullus uvažuje všechny možné způsoby, jak seřadit písmena do dvojic. Získává 36 dvojic písmen v 36 *komorách* (jak nazývá kombinace). Ve skutečnosti předpokládá i obrácené pořadí písmen (písmeno může fungovat jako subjekt i jako predikát, jak je to objasněno na prvním obrazci), možností je tedy celkem 72. Po těchto kombinatorických operacích Lullus přistupuje k tzv. *vyprazdňování komor*. Komora BC čtená podle prvního obrazce dává *bonitas* a *magnitudo*, podle obrazce druhého *differentia* a *concordia*. Tímto způsobem dostaneme 12 výpovědí: *Dobro je velké. Dobro je rozdílné. Dobro je shodné. Velikost je dobrá. Velikost je rozdílná. Velikost je shodná. Rozdíl je dobrý. Rozdíl je velký. Rozdíl je shodný. Shoda je dobrá. Shoda je velká. Shoda je rozdílná.* Jedná se tedy o počet dvoučlenných variací ze čtyř prvků bez opakování. Vrátíme-li se k tabulce 2.1 (*tabula generalis*) a přiřadíme-li písmenům B a C příslušné otázky (*utrum* a *quid*), dostaneme z těchto 12 výpovědí 24 otázek (otázky jsou typu *Je-li dobro velké? Co je velké dobro?*). Teoreticky je tedy možné pro třetí obrazec, který je tvořený 72 komorami, vytvořit 432 oznamovacích vět a 864 otázek.

Čtvrtý obrazec. Je nejznámější a v průběhu historie se setkává s největším úspěchem. Pracuje s trojicemi, které se vytváří z devíti prvků. Jedná se o pohyblivý mechanismus složený ze tří soustředných kruhů různé velikosti, které jsou vloženy do sebe a ve středu upevněny zadrhnutým lankem. Zde můžeme spatřovat vliv kabaly na Lulla, s jejíž tradicí se pravděpodobně seznámil na Iberském poloostrově, kde žil. *Sefer Jecira* přirovnává totiž božskou kombinatoriku ke kolu. Devět prvků ve skupinách po třech dává 84 kombinací. Jinde ve svém díle (*Ars brevis*) uvádí Lullus rovněž počet 252 kombinací, protože každé trojici lze přiřadit ještě tři otázky odpovídající každému z písmen trojice. Tyto trojice vložením písmene T rozšiřuje na čtveřice (získává tak například čtveřice typu $BCDT$, $BCTB$, $BTBC$ apod.) tak, že každá z 84 trojic generuje sloupec o 20 kombinacích. Písmenko T v tomto případě není součást kombinatoriky, ale mnemotechnická pomůcka, která říká, že písmena před T je třeba číst jako absolutní principy prvního ob-



razce, písmena za T jako relativní principy obrazce druhého (viz tabulku 2.1). Například čtveřici $BCTC$ chápeme takto: B = *bonitas*, C = *magnitudo*, druhé C = *concordantia*. Všechny vzorce začínající písmenem B odpovídají podle tabulky principů první otázce, vzorce začínající písmenem C druhé otázce apod. Výraz $BCTC$ tedy čteme: „*Je-li dobro velké, když obsahuje shodné věci?*"

bdkt	beft	begt	beht	beft	bekt	bfgt	baht	bift	bfkt	baht	bgit
bdtb	betb	betb	betb	betb	betb	bftb	bftb	bftb	bftb	bgtb	bgtb
bdtd	bete	bete	bete	bete	bete	bftf	bftf	bftf	bftf	bgtg	bgtg
bdtk	betf	betg	beth	beti	betk	bftg	bfth	bfti	bftk	bgth	bgti
bktb	bftb	bgtb	bhtb	bitb	bktb	bgtb	bhth	bitb	bktb	bhtb	bitg
bktd	bfte	bgte	bhte	bite	bkte	bgtf	bhtf	bitf	bktf	bhtg	bitg
bktk	bftf	bgtg	bhth	biti	bktk	bgtg	bhth	biti	bktk	bhth	biti
btbd	btbe	btbe	btbe	btbe	btbe	btbf	btbf	btbf	btbf	btbg	btbg
btbk	btbf	bteg	btbh	btbi	btbk	btbg	btbh	btbi	btbk	btbh	btbi
btdk	btef	betg	bteh	btei	btek	btfg	brfh	btfi	btfk	btgh	btgi
dktb	eftb	egtb	ehtb	eitb	ektb	fgtb	fhtb	fitb	fktb	ghtb	gitb
dktd	efte	egte	ehte	eite	ekte	fgtf	fhtf	fitf	fktf	ghtg	gitg
dktk	eftf	egtg	ehth	eiti	ektk	fgtg	fhth	fiti	fktk	ghth	giti
dtbd	etbe	etbe	etbe	etbe	etbe	ftbf	ftbf	ftbf	ftbf	gtbg	gtbg
dtbk	etbf	etbg	etbh	etbi	etbk	ftbg	ftbh	ftbi	ftbk	gtbh	gtbi
drdk	etef	eteg	eteh	etei	etek	ftfg	ftfh	ftfi	ftfk	gtgh	gtgi
ktbd	ftbe	gtbe	htbe	itbe	ktbe	gtbf	htbf	itbf	ktbf	htbg	itbg
ktbk	ftbf	gtbg	htbh	itbi	ktbk	gtbg	htbh	itbi	ktbk	htbh	itbi
ktdk	ftef	gteg	hteh	itei	ktek	gtfg	htfh	itfi	ktfk	htgh	itgi
tbdk	tbef	tbeg	tbeh	tbei	tbek	tbfg	tbfh	tbfi	tbgk	tbgh	ibgi

Tabulka 2.2: Kombinace čtyř prvků (podle štrasburského vydání z r. 1598)

Na první pohled jsou tyto čtveřice matoucí, neboť se v nich zdánlivě opakují písmena. Pokud bychom totiž uvažovali opakování písmen, nebylo by trojic 84, ale 729. Řešení je následující. Protože každé písmeno může označovat jak *principia*, tak *relationes* podle toho, jestli stojí před písmenem T nebo za ním, má vlastně každé písmeno dvě hodnoty, a proto v každém z 84 sloupců Lullus nekombinuje tři písmena, ale písmen šest. Rozdělíme-li těchto šest různých prvků do trojic, dostaneme přesně 20 kombinací, které se nám objevují v každém sloupci. Celkem nám tedy 84 sloupců o 20 čtveřicích dává 1 680 kombinací.

A zde se již dostáváme k omezení Lullovy kombinatoriky, neboť všech 1 680 čtveřic nevede k platné argumentaci. Jak jsme již uvedli dříve, je jednak Lullova kombinatorika svázána zákony sylogistiky (může vést k objevům pouze tehdy, existuje-li střední termín), jednak je omezena uspořádáním kosmu. Lullus některá tvrzení odmítá, i když jsou po formální stránce v pořádku, neboť neodpovídají skutečnosti. Například podle sylogistiky můžeme říci, že „*Lakota*

je odlišná od dobroty, Bůh je lakomý, a tedy Bůh je odlišný od dobroty", ale Lullus toto tvrzení odmítá, neboť není podle něj reálné.

Z posledního příkladu je zřejmé, že jeho metoda není pouze formální. Leibniz si dokonce (ve své *Dissertatio de arte combinatoria* z roku 1666) kladl otázku, proč se Lullus spokojil s tak omezeným počtem prvků, proč jejich počet nenechal otevřený. Jako by Lullus vůbec neuvažoval o volném kombinování výrazových prostředků bez vztahu k obsahu. Svou *Ars* Lullus totiž chápal jako prostředek k přesvědčování na pravou víru. Proto studoval židovské a muslimské náboženství a sám uvádí, že některé termíny z *Ars* si vypůjčil od Arabů. Hledal dokonce takové základní pojmy, které by byly společné i bezvěrcům, a proto nakonec omezil počet principů na devět (desátý princip označený písmenem A, který představoval božskou dokonalost a jednotu, nakonec vyloučil). Mezi první a poslední verzí své práce urazil obrovskou cestu a *Ars* se tak stala spíše nástrojem k pojednání o celé encyklopedii vědění.

Často se rovněž zdůrazňovala analogie mezi Lullovou kombinatorikou a kabalou. Kabalistické myšlení ale kombinováním písmen skutečnost vytváří, zatímco Lullova kombinatorika je jakýmsi rétorickým nástrojem, kterým má být dokázáno to, co už je známé.

Dílo Raimunda Lulla si našlo celou řadu pokračovatelů. Jako první se o Lullovi zmiňuje Pico della Mirandola v práci *Apologia* z roku 1487. Všiml si jistých analogií mezi kabalistickou *temurou* (zde ji nazývá „*revolutio alphabetaria*") a Lullovou kombinatorikou, ale zároveň si uvědomuje, že se jedná o dvě rozdílné věci. Vliv Lulla je rovněž patrný u Johna Deea, který patřil k významným vědcům, ale i alchymistům své doby. V práci *Monas Hieroglyphica* z roku 1564, která je považována právě za dílo alchymistické, se pokouší o objasnění kosmických vztahů, které vychází z pozorování a vysvětlení základního symbolu tvořeného kruhem a přímkou. Pomocí nich generuje všechny aritmetické veličiny.

Myšlenku univerzálního jazyka můžeme vysledovat v 17. století například u Johna Webstera, který kritizuje ve své práci z roku 1654 *Academiarum examen* akademický svět za to, že se problematice univerzálního jazyka nevěnuje dostatečně. Seth Ward ve stejném roce vystupuje proti Websterovi na obranu akademického světa prací *Vindiciae Academiarum*, v níž kritizuje mystické sklony svého protivníka. Mezi další zastánce univerzálního jazyka pak patří například Angličan John Wilkins, Dán Francis Lodwick či Čech Jan Amos Komenský (v práci *Via lucis* z roku 1668 uvádí pravidla umělého univerzálního jazyka).

Lullovy myšlenky o univerzálním jazyce a o kombinacích pojmů měly vliv také na G. W. Leibnize. Ten hovořil o „*ars combinatoria*" a díky tomuto počinu jej dnes považujeme za jednoho ze zakladatelů teorie kombinatoriky. Lullův vliv můžeme vysledovat i u Giordana Bruna, představitele myšlenky nekonečného vesmíru, který nemá nikde obvod a střed má všude. Nejvíce je tento vliv patrný v jeho mnemotechnickém traktátu *De lampade combinatoria Liliana ad infinitas propositiones et media invenienda…* z roku 1586 a v práci *De umbrisidearum* (1582), v níž Bruno navrhuje podobně jako Lullus pohyblivé soustředné kotouče rozdělené na 150 oddílů. G. P. Harsdörffer uvádí ve své

práci *Mathematische und philosophische Erquickstunden* z roku 1651 hru, kterou lze rovněž spojovat s Lullem či Brunem. Na 5 kotoučích se rozmístí 264 jednotek (předpony, přípony, písmena a slabiky), které mohou kombinatoricky vytvořit 97 209 600 německých slov včetně slov neexistujících, použitelných k poeticko-kreativním účelům. Autor uvažuje o tom, že pokud lze sestavit něco takového pro němčinu, je to možné i pro ostatní jazyky. V roce 1670 Chrisoph Clavius ve své práci *In spheram Ioannis de Sacro Bosco* kombinuje 4 prvotní vlastnosti (teplo, zima, sucho, vlhko). Dochází k tomu, že lze těchto kombinací vytvořit celkem 6, ale protože některé z nich jsou neslučitelné (teplo x zima, sucho x vlhko), fakticky možné jsou jen 4 kombinace, které vystihují přírodní živly (země – studená a suchá, oheň – teplý a suchý, vzduch – teplý a vlhký, voda – studená a vlhká). Řeší zde tedy stejný problém jako Lullus, tj. omezení skrytou kosmologií. Uvažuje dále o počtu termínů sestavených z 23 písmen abecedy, zkombinuje-li je po 2, 3,...23. Nabízí různé vzorce, ale nedokončuje je, neboť je zcela ohromen obrovským množstvím těchto kombinací, zvláště je-li bráno v úvahu i opakování. Paul Guldin roku 1622 v *Problema arithmeticum de rerum combinationibus* vypočítává počet „*dictiones*" z 23 písmen bez opakování bez ohledu na smysl. Dochází k závěru, že počet slov o délce od 2 do 23 písmen převyšuje „*70 000 miliard miliard*" a k jejich zapsání by bylo třeba přes „*milion miliard miliard písmen*". Jen pro představu uvádí příklad se seznamem. Má-li seznam 1 000 stran o sto řádcích po 60 znacích, pak by bylo potřeba „*257 milionů miliard*" takových seznamů a 8 052 122 350 knihoven, které by měly krychlovou konstrukci o straně 432 stop a jež by obsáhly 32 milionů svazků. Vypočítáme-li využitelnou plochu na celé planetě, mohli bychom na ní umístit jen 7 575 213 799 těchto knihoven. Podobnými úvahami se zabývá i francouzský páter Marin Mersenne v práci *Harmonie universelle* z roku 1636, který ale bere v potaz vedle „*dictiones*" i generovatelné „*zpěvy*" (tedy hudební sekvence). Generuje permutace bez opakování 22 tónů, tj. 3 oktáv.

Závěrem můžeme tedy konstatovat, že výše uvedení autoři (Mersenne, Clavius, Guldin aj.) již tenkrát objevili kouzlo kombinatoriky – nepracovali totiž s pojmy (jako například Lullus), ale jen s výrazovými prostředky.

Kabala a lullismus měly vliv na rozvoj *steganografie*, tj. tajných písem. Za zakladatele je považován opat Trithemius (1462–1516), který vychází spíše z kabaly a velice často pracuje s různými mnemotechnickými pomůckami k dešifrování či šifrování. Podobně jako Lullus pracuje s pohyblivými soustřednými kruhy (na nichž jsou zobrazena písmena), které ovšem na rozdíl od něj nepoužívá k hledání a odhalování vztahů, ale pro snadnější vznik různých šifrovacích klíčů či jako pomůcky k dešifrování. Důležité je říci to, že ačkoliv steganografie vychází z kabaly či lullismu, zachází dále, neboť steganografy nezajímá obsah (pravdivost) kombinací, ale vytváří kombinace čistě formální – snaží se původní symboly nahrazovat symboly novými vždy jinak, nevyzpytatelně. S určitostí nelze říci, zda Trithemius znal Lulla, nejspíše ale díky zájmu o kabalu znal temuru. Ostatní steganografové už se ale na R. Lulla odkazují. Giambattista della Porta (též Giovanni Battista Della Porta či John Baptist Porta) v 1. vydání své kryptografické práce *De furtivis literarum notis* z roku 1563 uvádí tabulky 400

dvojic písmen vzniklých kombinací 20 písmen abecedy, dále uvádí kombinace trojic písmen a tyto kombinatorické tabulky doplňuje seznamy tajných abeced, ať už vymyšlených či přejatých z jazyků Středního východu. Gustavus Selenus ve své práci z roku 1624 *Cryptomenytices et Cryptographiae libri IX* sestrojuje kotouč s 25 soustřednými kruhy, v němž kombinuje 25 řad o 24 dvojicích a vzápětí uvádí sérii tabulek, které obsahují na 30 000 trojic. Na závěr ještě zmiňme Heinricha Hillera, který ve své práci *Mysterium artis steganographicae novissimum* z roku 1682 podává návod k dešifrování jakékoli zprávy, a to nejen, když je napsána šifrou, ale také latinsky, německy, italsky nebo francouzsky, a to na základě statistického výskytu písmen a dvojhlásek v jednotlivých jazycích. Zdůrazněme, že frekvence jazykových jevů se k různým účelům začala prosazovat zejména během 19. století (viz kap. 2.3), zde ale nacházíme doklad využití frekvence o celých sto padesát let dříve.

Steganografie má již velmi blízko ke *kombinatorice*. Jedním z nejvýznamnějších reprezentantů kombinatorického přístupu k jazyku byl v 17. století John Wilkins. Tato osobnost ovlivnila například významného spisovatele Jorge Luis Borgese k napsání eseje *The Analytical Language of John Wilkins*. Jak jsme již uvedli dříve, na přelomu 17. a 18. století se kombinatorikou zabýval G. W. Leibniz.

Důraz na matematiku a prosazování přísně příčinného výkladů jevů je typické pro René Descarta (1596–1650). Tyto jeho myšlenky vedly k zformování tzv. *racionalismu*. V jazykovědě se Descartův vliv projevil vydáváním tzv. *filozofických gramatik* (racionalisticky pojatých gramatik), z nichž nejproslulejší byla tzv. „*Gramatika Port-Royal*“ vydaná roku 1660 v Paříži pod dlouhým názvem *Grammaire générale et raisonnée contenant les fondemens de l'art de parler, expliqués d'une manière claire et naturelle*, který bývá zpravidla zkracován na *Grammaire générale et raisonnée de Port-Royal*. Byla napsána předními představiteli opatství a školy Port-Royal (zejména A. Arnauldem a C. Lancelotem) ležící nedaleko Paříže. Snahou bylo vytvořit obecně pojatou gramatiku (platící pro libovolný jazyk) a podat rozumové zdůvodnění gramatických jevů. Gramatika Port-Royal měla značný vliv na rozvoj lingvistiky zejména v 18. století, ale odkazovali se na ni i lingvisté moderní, například Ferdinand de Saussure a zakladatel generační a transformační mluvnice Noam Chomsky, který v této gramatice dokonce viděl svého předchůdce.

Na našem území se touto problematikou vedle J. A. Komenského (více viz kap. 2.10) zabýval v 17. století logik, matematik a lingvista Jan Caramuel z Lobkovic, který se v několika spisech (*Steganographia*, *Apparatus philosophicus*, *Theologia rationalis*, *Leptotatus*) věnoval otázkám konstrukce umělého jazyka a jeho spekulativní mluvnice, které by překonaly nedostatky přirozených jazyků.

2.3 Pojem „frekvence“

Přelom ve vztahu matematika versus lingvistika nastává s využíváním pojmu *frekvence* (*četnost*), jednoho z nejdůležitějších pojmů kvantitativní lingvistiky. Frekvence vyjadřuje počet určitého jevu v celku (viz např. tabulku 2.3). Jazykové jednotky ve všech rovinách (slova, slabiky, fonémy, hlásky, písmena, slovesné časy, pády, typy vět apod.) se vyskytují v různém počtu, mají různou četnost. Využívání znalosti frekvence určitých jazykových jevů bylo vyvoláno zejména potřebami praxe.

Pořadí	*Slovo*	*Frekvence*
1.	a	67 122
2.	být	43 148
3.	ten	37 280
4.	v(e)	33 679
5.	on	32 496
6.	na	27 753
7.	že	18 092
8.	s(e)	14 951
9.	z(e)	13 408
10.	který	11 692

Tabulka 2.3: Pořadí prvních 10 nejčastějších slov podle FSČ[12]

Například tiskaři při sestavování kazet se zásobami písmen zjišťovali, že pro svou práci potřebují mnohem více takových písmen, která mají v daném jazyce frekvenci vysokou, než písmen s frekvencí nízkou (například v češtině by v takové kazetě musela být mnohem větší zásoba písmene „*a*“ než například písmene „*ů*“).

Rovněž stenografové při sestavování těsnopisných systémů přiřazovali z ekonomických důvodů nejjednodušší znaky těm jednotkám, které mají v daném jazyce nejvyšší frekvenci (podrobněji viz samostatná kapitola 2.3.2).

Frekvence hrála důležitou roli v pedagogické praxi – např. při hledání nejefektivnějších metod vyučování cizím jazykům, zlepšování pravopisného systému apod. Již J. A. Komenský ve své učebnici *Janua linguarum reserata* (vydané poprvé roku 1631 v Lešně) zdůrazňuje znalost výskytu slov daného jazyka pro efektivnost vyučování, i když žádné přesné statistiky neuvádí.

[12] Běžně užívaná zkratka pro [25]. Tento slovník byl sestaven na základě 1 623 527 slovních výskytů ze 75 literárních děl a tiskovin různého charakteru z 8 různých slohově druhových skupin (krásná próza, poezie, literatura pro mládež, drama, odborná literatura, žurnalistika, vědecká literatura, mluvené projevy). Podrobněji popsáno v kapitole 2.3.1.

Písmeno (znak)	*Kód*	*Frekvence v angličtině (%)*	*Pomocná slova*
A	.-	8.167	akát
B	-...	1.492	blýskavice
C	-.-.	2.782	cílovníci
D	-..	4.253	dálava
E	.	12.702	erb
F	..-.	2.228	Filipíny; filiálka
G	--.	2.015	Grónská zem; gnómóny
H	....	6.094	hrachovina
CH	----		chvátá k nám sám; chléb nám dává
I	..	6.966	ibis; Ivan
J	.---	0.153	jasmín bílý; junácká hůl
K	-.-	0.772	krákorá; království
L	.-..	4.025	lupíneček
M	--	2.406	mává; mlází
N	-.	6.749	národ; nástup; Náchod
O	---	7.507	ó náš pán
P	.--.	1.929	papírníci
Q	--.-	0.095	kvílí orkán
R	.-.	5.987	rarášek
S	...	6.327	sekera; sobota
T	-	9.056	trám; tón; tůň
U	..-	2.758	učený; uličník
V	...-	0.978	vyučený; vyvolený; vinobraní
W	.--	2.360	Waltrův vůz; wagón klád; dvojté vé
X	-..-	0.150	Xénokratés; Xénie má
Y	-.--	1.974	ý se krátí; ý se ztrácí; Ykar mává
Z	--..	0.074	zpíváme jen; známá žena; zrádná žena
1	.----		
2	..---		
3	...--		
4	....-		
5	.....		
6	-....		
7	--...		
8	---..		
9	----.		
0	-----		

Tabulka 2.4: Symboly pro kódování písmen a číslic v Morseově abecedě

Sestavování *Morseovy abecedy*, složené z teček a čárek, bylo rovněž založeno na pojmu frekvence. Toto kódování abecedy vymyslel americký vynálezce Samuel Morse a první telegrafické spojení uskutečnil roku 1844 mezi Washing-

tonem a Baltimorem. V roce 1918 byla původní abeceda dále zdokonalena. Konkrétně byla Morseova abeceda postavena na frekvenci písmen v angličtině. Písmenům nejčastěji se v angličtině vyskytujícím byla přidělena nejjednodušší sekvence teček a čárek. Proto například nejfrekventovanějšímu písmenu v angličtině „e" odpovídá v Morseově abecedě znak „·" (tečka), písmenu „t" odpovídá znak „–" (čárka) atd. Podrobněji viz tab. 2.4.

Podobně i na klaviatuře psacího stroje jsou nejčastější písmena na místech nejsnadněji dostupných nejobratnějšími prsty, tj. ukazováčkem a prostředníkem. I tato aplikace je ale optimálně sestavena pro angličtinu. Klaviatura českých psacích strojů vychází z této klaviatury anglické, je pouze doplněna o písmena v angličtině se nevyskytující a dále je provedena záměna písmen „z" a „y". Ale například polská matematická lingvistika v 60. letech 20. století provedla optimalizaci polských psacích strojů.

Frekvence písmen se využívala také při *dešifraci* textu. Vzhledem k tomu, že šifrováním a dešifrováním informací se zabývá disciplína nazývaná *teorie kódování* a není v našich silách popsat ani základní problematiku této disciplíny, omezíme se pouze na jeden zajímavý příklad z české literární historie. Básník Karel Hynek Mácha (1810–1836) si šifroval některé části svých intimních deníkových záznamů. Šifrování prováděl tak, že si sestavil vlastní abecedu, většinou odvozenou z řeckých písmen, a každé písmeno textu zaměnil symbolem z takto vytvořené abecedy. Situaci Mácha sice ještě poněkud ztížil tím, že češtinu kombinoval s německými slovy a větami a dále že každý druhý řádek (s jedinou výjimkou) psal pozpátku, ale při dešifraci pak v podstatě stačilo (zjednodušeně řečeno) sestavit tabulku výskytů Máchových symbolů a porovnat ji s obdobnou tabulkou sestavenou pro českou abecedu. Jednalo se vlastně o jednoduchou substituční šifru, kdy je jeden symbol nahrazen symbolem jiným. Řešení jednoduchých substitučních šifer bylo v arabském světě známo již v 9. století – Abú-Yūsuf Ya'qūb ibn Ishāq al-Kindī (801–873). Pro Evropu bylo řešení substitučních šifer objeveno mnohem později, a to někdy na začátku století čtrnáctého. Už tehdy se proto začala hledat vylepšení tohoto systému, která by umožnila zakrytí statistických závislostí v šifrovaném textu.

Všechny výše uvedené aplikace pojmu četnost se pohybovaly na úrovni pouhé deskripce, tj. něco se počítalo. Jako první na nedostatky pouhé deskripce a na vhodnost využívání vyšší matematiky upozornil roku 1847 V. Ja. Bunjakovskij (viz kap. 2.4) . Nyní se ale u některých aplikací matematiky v lingvistice využívajících pojmu *frekvence* zastavme podrobněji. Nejprve si tu představíme problematiku frekvenčních a konkordančních slovníků a dále si ukážeme využívání frekvence jazykových jednotek v těsnopise.

2.3.1 Frekvenční a konkordanční slovníky

S pojmem *frekvence* úzce souvisí *frekvenční* a *konkordanční* slovníky, které začaly hojně vznikat od konce 19. století. *Frekvenčním slovníkem* rozumíme víceúčelový seznam slov informující o jejich frekvenci, která bývá zjištěna na základě dostatečně rozsáhlého reprezentativně zvoleného materiálu. Zpravidla bývá v takovém slovníku uvedeno několik seznamů slov sestavených podle ur-

čitých hledisek, nejčastěji je to hledisko frekvenční (slova jsou seřazena podle klesající frekvence) a abecední (slova jsou uspořádána podle abecedy). Jednotlivá slova jsou zpravidla uváděna v základním tvaru (u substantiv, adjektiv, zájmen a číslovek to je 1. pád singuláru, u sloves infinitiv). Někdy bývají slova doplněna dalšími statistickými údaji. Pokud vytváříme pouze jednoúčelové seznamy slov na základě studia méně rozsáhlého materiálu (například dílo autora apod.), hovoříme o tzv. *frekvenčních seznamech*. Je-li ve frekvenčních slovnících navíc u slova uvedena informace, na jakém místě v excerpovaném textu se toto slovo nachází, mluvíme o *slovnících konkordančních*. Frekvenční seznamy, které bývají doplněny informací o tom, na jakém místě v textu se slovo nachází, bývají nazývány *indexy* (též *konkordance*).

První frekvenční slovníky vznikaly pro dobové praktické účely, například pro potřeby tiskařů[13], konstrukci ekonomického těsnopisného systému[14], zlepšení pravopisného systému[15], tvorbu metod slepeckého čtení[16], efektivnost vyučování cizím jazykům[17] apod. Až poměrně pozdě se frekvenční slovníky začaly užívat jako důležitý nástroj k poznání systému a struktury slovní zásoby jednotlivých děl, autorů, funkčních stylů apod., dále pro účely strojového překladu, anotování a indexování textu z hlediska statistického. Zpočátku se frekvenční slovníky zpracovávaly ručně a uchovávaly pomocí excerpčních lístků. Aby byly frekvenční slovníky dostatečně vypovídající a odrážely skutečnou jazykovou situaci, je třeba zpracovat rozsáhlý materiál, což je ovšem ručně velmi pracné a hlavně zdlouhavé. Dnes tyto nedostatky umožňují úspěšně překonat počítače. V rámci počítačové lingvistiky se tak vytvořila samostatná disciplína, kterou nazýváme *korpusová lingvistika* a která se zabývá právě zpracováním rozsáhlých souborů jazykových dat (tzv. *korpusů*). V současné době existuje celá řada různých frekvenčních slovníků v různých jazycích, neboť tyto slovníky mohou být zajímavým zdrojem informací nejen pro lingvisty, ale i pro matematiky, informatiky, psychology, filozofy, pedagogy, metodiky, stenografy, kryptology apod.

Frekvenční slovníky se od sebe odlišují nejčastěji rozsahem zpracovávaného materiálu (ty novější mají rozsah od přibližně 500 000 slov až po několik milionů slov), výběrem excerpovaných textů (mluvené projevy, beletrie, poezie, dílo jednoho autora, díla určité literární skupiny apod.) či zpracováním (ruční, počítačové – dříve pomocí děrnoštítkových strojů). Vznikají různé speciální frekvenční slovníky (seznamy) – např. slovník děl (díla) určitého autora[18], re-

[13] Gamble, W.: *Two lists of selected characters containing all in the Bible and twenty-seven other books*. Shanghai 1861.

[14] Käding, F. W.: *Häufigkeits Wörterbuch der Deutschen Sprache*. Steglitz 1897.

[15] Ayers, L. P.: *A Measuring Scale for Ability in Spelling*. NY 1915.

[16] Knowles, J.: *The London Point System of Reading for Blind*. London 1904.

[17] Eldridge, R. C.: *Six Thousand Common English Words: Their Comparative Frequency and What Can Be Done With Them*. Niagara Falls, NY 1911.

[18] *Slovar jazyka Puškina*. Moskva 1957; *Materialy k častotnomu slovarju jazyka Puškina*. Moskva 1963; *Slovnyk movy Ševčenka*. Kyjiv 1964; *Słownik języka Adama Mickiewicza*. Warszawa 1965; Sambor, J.: *Badanie statystyczne nad słownictwem (Na materiale Pana Tadeusza)*. Wrocław etc. 1969; *Goethe-Wörterbuch*. Berlin 1966; *Rečnik na ezika na Christo Botev I*. Sofija 1960; L. Benkö: *Juhász Gyula Költöi nyelvének szotári feldolgozása*. Szeged 1957.

trográdní slovníky[19], dialektologické slovníky[20], slovník literární skupiny apod.

Při zpracovávání frekvenčních slovníků se můžeme setkat s jistými problémy. Ty nastávají již při definování pojmu „*slovo*". Nejjednodušší (a při počítačovém zpracování nejčastější) je chápání slova jako grafické jednotky, tzn. jako skupiny písmen vyskytující se mezi dvěma mezerami. Komplikace způsobují při tomto pojetí slova zejména složené slovesné tvary (*bude žehlit*), sdružená pojmenování (*vlčí mák*, *paví oko*), spřežky (často je lze psát dohromady i zvlášť – *vpodvečer* i *v podvečer*), stupňování přídavných jmen a příslovcí, číslovky (brát jako jednu lexikální jednotku celou číslovku či její části – *pět set*) či částice (dost dlouho se s tímto slovním druhem nepracovalo, zpravidla byly částice zařazovány k příslovcím, k přehodnocení dochází až v SSČ[21]), rozpoznání rozdílu mezi trpným rodem slovesným a spojením „*být*" s přídavným jménem (*chlapec byl vychován* x *chlapec je nemocen*) apod. V těchto případech je třeba držet se zásady M. Těšitelové, a to té, že jednou přijatý způsob je třeba pokud možno všemi a všude dodržovat, aby bylo možno získané výsledky navzájem porovnávat.

Nyní si podrobněji představíme několik frekvenčních slovníků, které hrály významnou roli v historii kvantitativní lingvistiky. Nejpodrobněji se zastavíme u slovníků sestavených pro němčinu, neboť právě pro německý jazyk vznikl vůbec první frekvenční slovník na světě. Na tomto přehledu si ukážeme, jak se měnil obsah jednotlivých slovníků podle momentálních požadavků oboru. Orientačně se zmíníme o některých frekvenčních slovnících sestavených pro angličtinu, ruštinu, francouzštinu, španělštinu a rovněž pro slovenštinu a češtinu.

Němčina

Za vůbec první frekvenční slovník považujeme dvoudílný německý *Häufigkeitswörterbuch der deutschen Sprache* („Slovník četnosti výskytu německého jazyka")[22] stenografa F. W. Kädinga (1843–1928). Původně byl tento slovník vytvořen k těsnopisným účelům, podnět k jeho sestavení dal pracovní výbor německých těsnopisců v roce 1891. Výzkum probíhal více než 5 let na 110 pracovištích a byl zpracován materiál o rozsahu 20 milionů slabik (10 910 777 slov). Tento slovník se pak stal cennou pomůckou při kodifikaci těsnopisné soustavy Stolzeovy-Schreyovy v roce 1897. Byl založen na promyšlených obecných zásadách, z kterých později vycházely i další frekvenční slovníky. Slovník je složen ze dvou částí: v části první se uvádí údaje týkající se frekvence slov a slabik, v části druhé údaje o frekvenci písmen (ve skutečnosti se ale jedná o hlásky). V úvodu práce se správně předvídá, že tyto údaje mohou být důležité nejen pro těsnopisce, ale i pro lingvisty. Při tvorbě tohoto slovníku byl kladen důraz na „*zákon velkých čísel*", tzn. na dostatečně velký rozsah materiálu, který by

[19] Jsou to slovníky, v nichž jsou slova uspořádana v abecedním pořadí od posledního písmene k prvnímu. Využívány bývají například ve slovotvorbě - výskyt různých sufixů, prefixů apod.

[20] Jsou to slovníky zachycující slovní zásobu různých nářečí.

[21] *Slovník spisovné češtiny pro školu a veřejnost.* Praha 1978. Další vydání: 2. opravené a doplněné vydání – 1994, 3. opravené vydání – 2003, 4. vydání – 2005).

[22] Steglitz 1897, 643 s.

eliminoval počet slov s nízkou frekvencí, nadměrný výskyt slov závislých na tématu apod. Touto problematikou se později zabývala ve 20. století kvantitativní lingvistika (jednotka kvantitativní analýzy, výběr a rozsah materiálu apod.). Základní jednotkou je v Kädingově slovníku slovo, přesněji tvar slova. Autor rozlišuje slova formální (řadí sem zájmena, adjektiva kvantitativní a číslovky, členy, pomocná a modální slovesa, nekvalitativní adverbia, předložky, spojky a citoslovce) a plnovýznamová (sem zařazuje substantiva, kvalitativní adjektiva a adverbia, konkrétní slovesa). Pro němčinu typické členy mají za následek nárůst počtu formálních slov, která jsou ale obecně nejčastější slova v různých jazycích.

Vedle svého prvenství je tento slovník významný i obrovským rozsahem ručně zpracovaného materiálu (10 910 777 slov – přesněji tvarů slov, z toho 258 173 různých tvarů slov). Dnes si zpracování tak rozsáhlého souboru bez pomocí počítačů vůbec nedovedeme představit. Pro potřeby stenografie byla slova dále členěna na slabiky (celkem asi 20 000 000 slabik) a na hlásky. K těsnopisným účelům bylo nutno slova rozkládat i z hlediska morfematického, tzn. na prefixy, kmeny, sufixy a koncovky. Značné úsilí bylo věnováno problematice složených slov – např. části složených slov a oddělitelné předpony byly zapisovány na různě barevné lístky. Všechny tyto slovotvorné prvky byly zaznamenány zvlášť a potom řazeny podle frekvence, popřípadě podle abecedy. To znamená, že vedle frekvence slov jsou ve slovníku obsaženy i seznamy prefixů, sufixů a koncovek, výskyt samohlásek a souhlásek v prefixech, kmenech, sufixech a v koncovkách apod. Pozornost není věnována slovům o frekvenci 1 až 3, neboť pro těsnopisné účely nejsou podstatná. Je zde ale správně upozorněno, že pro některá jazyková studia mohou mít tato slova zvláštní význam.

Za zmínku stojí i výběr materiálu, neboť se jednalo o souvislé texty z těch oborů, v nichž se nejčastěji v tehdejší době užívalo těsnopisu. Celkem se jednalo o 16 tematických oblastí (právo, obchod, náboženství, medicína, historie, časopisy a knihy, vojenství, dopisy aj.). Vhodnost tohoto výběru se ukazuje, srovnáme-li tento slovník například s výběrem textů pro těsnopisné účely, který plánoval v letech 1940–43 První pražský spolek stenografů pod vedením B. Trnky, neboť tento český výběr se s tematikou textů v Kädingově slovníku v podstatě shoduje, jen rozsah je značně menší.

Obecné zásady uplatněné při tvorbě Kädingova slovníku se staly součástí řady pozdějších prací z kvantitativní lingvistiky (zejména v první polovině 20. století).

Na závěr si uveďme ještě několik zajímavých výsledků. Tak například prvních 15 nejfrekventovanějších výrazů představuje asi čtvrtinu celého slovního materiálu běžného německého textu, k pokrytí celé poloviny textu postačuje prvních 66 slov a prvních 320 výrazů pokrývá již téměř tři čtvrtiny textu. K běžnému dorozumění je ovšem potřeba znát podle výsledků různých vědců slov přibližně 850.

Kromě toho, že se jednalo o první frekvenční slovník vůbec, jednalo se rovněž o první dílo věnované lexikální a gramatické statistice. Ke Kädingově slovníku se v pozdějších letech vracela celá řada lingvistů (např. Meier) a pedagogů (např. B. Q. Morgan). Ti však využili zpravidla jen část rozsáhlé slovní zásoby.

Zastavme se u některých pokračovatelů Kädinga podrobněji.

B. Q. Morgan[23] vypracoval v USA svůj frekvenční slovník němčiny pro potřebu jazykového vyučování, a to v době, kdy americká jazyková pedagogika opřela výuku mateřského jazyka, ale zejména jazyka cizího, o studium frekvence slov či jiných jazykových jednotek. Ve svém slovníku uvádí Morgan na rozdíl od Kädinga jen 2 400 nejčastějších německých slov. Protože podle něj Kädingův slovník nesl rysy svého původního určení, vyřadil ze seznamu uspořádaného podle klesající frekvence asi 320 slov, aby se dostal ke slovům, která jsou podle jeho mínění v němčině frekventovanější. Neuváděl rovněž další seznamy uvedené v Kädingově práci (frekvence hlásek, údaje o tvoření slov aj.). Lze pochopit, že Morgan pro potřeby jazykového vyučování využil jen některé ze seznamů, neakceptovatelné jsou ovšem subjektivní zásahy do původního seznamu.

Této chybě se chtěl vyvarovat H. Bakonyi[24], který sestavil rovněž pro pedagogické účely slovník nejužívanějších slov. Metodicky vycházel z Kädingova slovníku. Bakonyův slovník obsahuje celkem 5 874 slov uspořádaných na základě frekvenčních údajů (absolutní a relativní frekvence) a k těmto slovům je připojen i tzv. *stupeň důležitosti*. Pro potřeby didaktické totiž nestačí někdy jen znalost slov nejčastějších, ale je třeba znát i slova nejznámější (tedy nejdůležitější). Známost slov se snažil zjistit na základě slovní zásoby dětského slovníku, kterým rozšířil materiál Kädingova slovníku. Na principu slovníku H. Bakonyiho u nás vznikly rovněž 2 slovníky, a to frekvenční slovník němčiny a angličtiny[25].

H. Meier[26] se pokusil o podrobné statistické roztřídění Kädingova materiálu na 12 stupňů častosti (*Häufigkeitsstufen*). Tyto stupně jsou dány frekvenčními intervaly, a to od frekvencí nejvyšších do frekvence nejnižší. Toto podrobné třídění ale vede k jisté nepřehlednosti celé práce.

Mezi další autory frekvenčních slovníků němčiny, kteří už ovšem nevycházeli ze slovníku Kädingova, patří H.-H. Wängler[27], který vydal pro němčinu frekvenční slovník hovorové horní němčiny. Slovník byl vytvořen na základě 80 266 slov ze severoněmecké oblasti zachycených v nenucené zábavě a pro srovnání byl týž počet slov získán excerpcí různých textů z denního tisku a ze zábavných časopisů. Každé slovo pak bylo doplněno informací o frekvenci jednak v němčině hovorové, jednak v němčině psané.

Za vedení L. Hoffmanna vznikla v DDR na lipské univerzitě série frekvenčních slovníků odborných disciplín[28]. Jedná se o speciální druh frekvenčního slovníku, který má obdobu v některých pracích sovětských[29]. V podstatě jsou

[23] Morgan, B. Q.: *German Frequency Word Book*. New York 1928.

[24] *Die gebräuchlichsten Wörter der deutschen Sprache*. München 1934.

[25] Kumprecht, K.: *Nejužívanější slova v němčině*. Praha 1936; Kumprecht, K. – Schulhofová, D.: *Nejužívanější slova v angličtině*. Praha 1947.

[26] *Deutsche Sprachstatistik*. Hildesheim 1964.

[27] *Rangwörterbuch hochdeutscher Umgangssprache*. Marburg 1963.

[28] *Fachwortschatz Medizin. Häufigkeitswörterbuch. Russisch, Englisch, Francösisch*. Leipzig 1970; *Fachwortschatz Physik. Häufigkeitswörterbuch. Russisch, Englisch, Francösisch*. Leipzig 1970; *Fachwortschatz Chemie. Häufigkeitswörterbuch. Russisch, Englisch, Francösisch*. Leipzig 1973.

[29] Piotrovskij, R. G. (red.): *Statistika teksta I*. Minsk 1969; Kiselevskij, A. I. (red.): *Statistika teksta II*. Minsk 1970.

to slovníky dvojjazyčné – například frekvenční slovník fyziky uvádí frekvenční seznam rusko-německý, anglicko-německý, francouzsko-německý. Seznamy obsahují kolem 1 100 slov a u jednotlivých lexikálních jednotek se uvádí jejich relativní frekvence. Slovníky mají sloužit jako metodická pomůcka pro žáky i učitele cizích jazyků. Za redakce L. Hoffmanna vyšel rovněž speciální sborník[30], v němž je obsažena problematika sestavování frekvenčních slovníků jazyka odborného stylu. Autor se v něm na základě zkušeností z práce na výše uvedených slovnících zabývá například metodologickými otázkami statistického zpracování vědeckých a technických odborných textů. Vedle toho jsou zde stati analyzující například slovník sovětských novin *Izvestija*, zvláštnosti ruských textů z oblasti marxisticko-leninské filozofie či z literární vědy.

I. Rosengrenová sestavila frekvenční slovník ze západoněmeckých novin[31]. Korpus byl získán reprezentativním výběrem z těchto novin v období jednoho roku (od 1. listopadu 1966 do 30. října 1967). Materiál z novin *Die Welt* má rozsah 6 545 600 slov, z novin *Süddeutsche Zeitung* 5 855 478 slov. Oba podkorpusy se člení na pět kategorií, v podstatě novinových rubrik: názory, politika, kultura, ekonomie a různé.

V roce 1972 vznikl první díl západoněmeckého frekvenčního slovníku německých odborných textů, jehož autorem je H. Erk[32]. Slovník přináší frekvenci sloves a jejich vazeb. Je vytvořen na základě materiálu o rozsahu 250 000 slov z 34 různých oborů. Ve slovníku se uvádějí pouze slova s výskytem nejméně 10. Jednotlivé lexémy jsou v seznamu doplněny *koeficientem distribuce* a *frekvence*. V roce 1975 vyšel druhý díl tohoto slovníku zaměřený na substantiva[33] a v roce 1982 díl třetí, který se zaměřuje na adjektiva, adverbia a ostatní slovní druhy[34].

Angličtina

K nejstarším anglickým frekvenčním slovníkům patří slovník L. P. Ayrese[35] z roku 1915, který vznikl na základě excerpce 368 000 slov z obchodních a soukromých dopisů.

Pro účely vyučování angličtiny (a to jak mateřskému jazyku, tak jazyku cizímu) byl vytvořen jeden z nejznámějších anglických frekvenčních slovníků, na němž se podílel E. L. Thorndike[36]. Slovník tvoří 10 000 nejfrekventovanějších slov, která byla získána z materiálu o rozsahu 4 565 000 slov, pořízeného především z literatury pro mládež, z učebnic, rozmanitých příruček apod. Postupným rozšířením o dalších 5 milionů slov vznikla druhá verze slovníku[37]

[30] *Fachsprachen und Sprachstatistik*. Berlin 1975.

[31] *Ein Frequenzwörterbuch der deutschen Zeitungssprache Die Welt, Süddeutsche Zeitung*. Lund 1972.

[32] *Zur Lexik wissenschaftlicher Fachtexte. Verben – Frequenz und Verwendungsweise*. München 1972.

[33] *Zur Lexik wissenschaftlicher Fachtexte. Substantive – Frequenz und Verwendungsweise*. München 1975.

[34] *Zur Lexik wissenschaftlicher Fachtexte. Adjektive, Adverbien und andere Wortarten. Frequenz und Verwendungsweise*. München 1982.

[35] *A Measuring Scale for Ability in Spelling*. New York 1915.

[36] *The Teachers Word Book*. New York 1921.

[37] *A Teachers Word Book of Twenty Thousands Words*. New York 1931–1932.

a ve spolupráci s I. Lorgem rozšířením o nových 4,5 milionů slov verze třetí[38]. Podobně jako Kädingův německý frekvenční slovník i slovník Thorndikův byl zdrojem pro další práce, zejména z oblasti didaktické a lingvistické. L. Faucett a Itsu Maki vydali slovník opírající se o jazykový materiál 10 milionů slov[39], který vznikl spojením Thorndikova slovníku z roku 1921 a slovníku E. Horna[40], jenž byl sestaven ze soukromé a obchodní korespondence. Tento Hornův slovník se stal základem další práce, a to slovníku A. Hooda Robertse[41], jehož cílem bylo zjištění frekvence v běžné americké angličtině.

Pozornost si zaslouží frekvenční slovník G. Deweye[42] vytvořený na základě korpusu o rozsahu 100 000 slov zejména z publicistiky. Protože byl tento slovník určen k těsnopisným účelům, zjišťoval rovněž frekvenci slabik a hlásek.

Ojedinělým počinem bylo vytvoření prvního sémantického frekvenčního slovníku M. Westem[43] v roce 1953. Frekvence významů jednotlivých slov se uvádí v procentech, přičemž významy se rozlišují ve shodě se slovníkem *The Oxford English Dictionary*. Na Westův slovník navazovali rovněž další autoři. Například D. Wolf[44] vyšel z údajů ve slovníku *Shorter Oxford English Dictionary* (obsahuje 80 000 slov, která jsou opatřena údaji o počtu významů, druhu slov, o prvním dokladu apod.) a výsledky své analýzy srovnává právě se slovníkem Westovým (2 000 nejfrekventovanějších slov) a se slovníkem *Advanced Learners Dictionary* (28 000 slov). Sleduje zejména počet významů slov se zřetelem k prvnímu výskytu slova. Dokazuje tak celkem známý fakt, že čím je slovo starší, tím více má významů.

H. Kučera a W. N. Francis jsou autory frekvenčního slovníku současné americké angličtiny[45], který byl vytvořen na základě materiálu 1 014 232 slov z 500 textů (každý o rozsahu cca 2 000 slov), jež byly publikovány během jednoho roku (1963–1964). Při sestavování tohoto slovníku bylo již ke zpracování použito počítačů (stejně jako u slovníků Robertse a Wolfa).

Další anglický frekvenční slovník vznikl na území bývalého Sovětského svazu v letech 1971–1972. Jednalo se o třísvazkový frekvenční slovník slovních spojení v angličtině[46]. Materiál byl získán ze 124 děl umělecké prózy (cca 7 milionů slov), 9 divadelních her (300 000 slov), 12 pramenů publicistických (300 000 slov) a 15 textů vědeckotechnických (500 000 slov). Východiskem při sestavování slovníku byl slovník Thorndikův-Lorgeovův. Slovník slovních spojení

[38] *The Teachers Word Book of 30 000 Words*. New York 1944.

[39] *A Study of English Word-Values Statistically Determined from Latest Extensive Word-Counts*. Tokyo 1932.

[40] *A Basic Writing Vocabulary*. Iowa City 1926.

[41] *A Statistical Linguistic Analysis of American English*. The Hague 1965.

[42] *Relativ Frequency of English Speach Sounds*. Cambridge 1923.

[43] *A General Service List of English Words with Semantic Frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology*. London 1953.

[44] *Bedeutungshäufigkeit und ihr statistisches Verhalten*. Beiträge zur Linguistik und Informationsverarbeitung 1972, č. 22, s. 33–44.

[45] *Computational Analysis of Present-Day American English*. Providence, Rhode Island 1967.

[46] Volkova, N. O. – Ginzburg, R. Z. – Perebejnos, V. I. – Sankin, A. A. – Chidekel, S. S.: *Častotnyj slovar sočetajemosti sovremennogo anglijskogo jazyka*. 1. část, Moskva 1971; 2. část ve dvou svazcích, Moskva 1972.

v angličtině má význam nejen pro studium lexika, ale i gramatiky, stylistiky apod.

Ruština

Rovněž frekvenční slovníky ruštiny vznikaly jednak na území bývalého Sovětského svazu, jednak mimo něj.

Pro didaktické účely se zřetelem k vyučování ruštiny na českých základních školách zkoumala frekvenci slov v ruském jazyce A. Střížová[47]. Excerpcí sovětských učebnic autorka stanovila 2 060 slov označených jako nejčastější slova v ruštině. K témuž účelu zjišťoval frekvenci slov v ruštině F. Malíř[48]. Materiál získal excerpcí sovětského tisku.

V USA byl ručně zpracován velký ruský frekvenční slovník H. H. Josselsona[49], který je založen asi na jednom milionu slov z psané ruštiny od roku 1830 do roku 1950. Podobně jako FSČ uvádí nejen frekvenci slov, ale i některých gramatických kategorií. Zřetel k frekvenci gramatických jevů, resp. jejich kategorií, je významným rysem frekvenčních slovníků vznikajících zhruba od 50. let 20. století.

Na půdě SSSR byl pořízen slovník E. A. Štejnfeldtové[50]. Jedná se o seznam 2 500 nejfrekventovanějších slov, která byla zjištěna na základě materiálu 400 000 slov z 350 textů. Přestože byl určen pro potřeby školské praxe, má význam i pro lingvistiku. Byl přeložen do jiných jazyků, např. do němčiny[51].

V tehdejším Leningradě vznikl frekvenční slovník současné ruštiny L. N. Zasorinové[52]. Materiál tvoří texty ze 4 stylistických skupin (po 25 %): umělecká próza, drama, populárně vědecké texty, publicistika. Vznik textů spadá do časového rozpětí 60 let.

Vedle toho vzniká v SSSR množství frekvenčních seznamů (či slovníků) pro speciální účely jednotlivých vysokoškolských učebních oborů, např. radioelektroniky, stavebních materiálů, fyzikální chemie, pro účely publicistické. Ukazují to například sborníky, které vyšly v Minsku za redakce R. G. Piotrovského[53] a A. I. Kiselevského[54].

Francouzština

První prací týkající se frekvence slov ve francouzštině byl *Le Vocabulaire d'un journal*, kterou pořídil Bony v letech 1920–1921 na základě jednoho čísla novin *Le Temps*.

[47] *Nejužívanější slova v ruštině.* Pedagogická revue 2, 1948, s. 335–359.
[48] *Rusko-český slovník nejdůležitějších slov pro četbu sovětského tisku.* Praha 1952.
[49] *The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literatury Russian.* Detroit 1953.
[50] *Častotnyj slovar sovremennogo russkogo jazyka.* Tallin 1963.
[51] Štejnfeldt, E.: *Häufigkeitswörterbuch der russischen Sprache.* Moskau 1965.
[52] *Častotnyj slovar russkogo jazyka.* Moskva 1977. Více informací o jeho koncepci nalezneme ve zprávě autorky (*Avtomatizacija i statistika v leksikografii.* Leningrad 1966.).
[53] Piotrovskij, R. G. (red.): *Statistika teksta I.* Minsk 1969.
[54] Kiselevskij, A. I. (red.): *Statistika teksta II.* Minsk 1970.

Pokus o sestavení frekvenčního slovníku francouzštiny učinil V. A. C. Henmon, který sestavil frekvenční seznam 4 000 slov na základě materiálu o rozsahu 400 000 slov. Tento seznam pak rozšířil a přepracoval G. E. Vander Beke[55]. Vander Bekův slovník je založen na materiálu 1 147 748 slov a u jednotlivých slov uvádí jednak *frekvenční koeficient*, jednak tzv. *distribuční koeficient*, tj. v kolika textech se slovo vyskytuje. Tento slovník byl podobně jako Thorndikovy slovníky určen pro didaktické účely, zejména pro vyučování francouzštině jako cizímu jazyku. Na podkladu Vander Bekova slovníku vznikly práce další – u nás například pomůcka pro výuku francouzštiny od K. Kumprechta a A. Formánkové-Kubínové[56].

Frekvenční slovník současné francouzštiny vydal roku 1970 Al. Juilland jako třetí v sérii frekvenčních slovníků románských jazyků (vedle španělštiny a rumunštiny). Jeho spoluautory byli D. Brodin a C. Davidovitch[57]. Korpus tvoří (podobně jako ve zmiňovaných frekvenčních slovnících španělštiny a rumunštiny) 500 000 slov z 5 žánrů (drama, beletrie, eseje, technická literatura a periodika). Slovník obsahuje u jednotlivých slov tyto koeficienty – *frekvenci*, *disperzi*[58] a *užití*[59]. V obsáhlém úvodu jsou zmíněny nejen otázky týkající se problematiky této práce, ale i základní otázky lexikální statistiky vůbec (struktura korpusu, vymezení pojmu slova, koeficienty, interpretace výsledků aj.).

Zvláštní místo má frekvenční slovník projevů bývalého francouzského prezidenta Charlese de Gaullea[60], který je založen na 46 televizních politických projevech a obsahuje 6 195 slov (z toho asi 4 000 lexémů). Analyzují se dva typy televizních projevů, a to tzv. provolání a úvahové projevy.

Španělština

Vedle již zmíněného Juillandova slovníku vzniklo pro španělštinu několik dalších frekvenčních slovníků. Podle vzoru Vander Beka a Thorndika sestavil svůj frekvenční slovník M. A. Buchanan[61]. Opírá se o 1 200 000 slov ze 40 textů.

V 50. letech 20. století vyšly další dva frekvenční slovníky španělštiny. Slov-

[55] *French Word Book*. New York 1930.

[56] *Přehled nejčastěji užívaných slov ve franštině*. Praha 1935.

[57] *Frequency Dictionary of French Words*. The Hague 1970.

[58] *Koeficient disperze* (D), též *rozptýlení*, zavedl A. Juilland a postihuje rozložení frekvence jednotky (slova) v jednotlivých textech korpusu, eventuálně v jejich částech. Lze jej vyjádřit vztahem

$$D = \frac{\delta}{m \cdot \sqrt{n-1}},$$

kde δ je směrodatná odchylka, m aritmetický průměr, n je suma všech frekvencí. Koeficient disperze se pohybuje v rozmezí 0–1. Při rovnoměrném rozdělení frekvence se blíží 0, při nerovnoměrném rozdělení se blíží 1.

[59] *Koeficient užití* (U) zavedl rovněž A. Juilland a lze jej vyjádřit vzorcem

$$U = \frac{F \cdot D}{100},$$

kde F je frekvence a D disperze.

[60] Cotteret, J. M. – Moreau, R.: *Le vocabulaire du général de Gaulle*. Paris 1969.

[61] *A Graded Spanish Word Book*. Toronto 1927.

ník V. G. Hoze[62], který obsahuje materiál 400 000 slov získaných ze soukromých dopisů, publicistiky a beletrie, je určen pro výuku španělštiny jako mateřskému jazyku. Stejný účel má i druhý slovník, jehož autorem je J. Rodriguez Bou[63]. Tento slovník se opírá dokonce o materiál 7 066 637 slov.

Al. Juilland[64] vydal svůj frekvenční slovník španělštiny společně s matematikem E. Chang-Rodriguezem. Stavba slovníku je stejná jako u výše uvedeného frekvenčního slovníku francouzštiny, rovněž obsahuje u jednotlivých slov *koeficienty frekvence*, *užití* a *disperze*. Zpracován byl pomocí počítačů.

Slovenština

Rovněž první frekvenční slovník slovenštiny[65] (tradičně zkracováno na FSS) vznikl k těsnopiseckým účelům. J. Mistrík jej sestavil na základě materiálu 1 milionu slov, která byla získána excerpcí 60 celých textů 5 stylových skupin (drama, umělecká próza, poezie, žurnalistika, naučná literatura). Frekvenční seznam obsahuje celkem 9 568 lexikálních jednotek uspořádaných podle klesající frekvence, a to až do frekvence 3. Abecední seznam je tvořen 21 823 nejfrekventovanějšími slovy v slovenštině.

Čeština

Různé frekvenční statistiky se pro češtinu vytvářely již během 19. století. Jednalo se zpravidla o různé seznamy hlásek či písmen určené k různým praktickým účelům. Vůbec první jednoduchou hláskovou statistiku nacházíme roku 1831 v časopise *Krok*. Tato statistika si kladla za cíl vyvrátit tvrzení, že čeština je jazyk tvrdý s obtížnou výslovností, neboť je v něm obsaženo velké množství souhlásek. Podrobněji se o této statistice zmíníme v kap. 2.11. Statistické výzkumy lexika byly prováděny až později.

V roce 1911 byl založen Lexikografický archív Ústavu pro jazyk český. V tomto archívu bylo shromážděno více než 9 miliónů excerpčních lístků. Dnes tyto lístky tvoří jeden abecední celek, ale původně těchto celků bylo asi 2 200. Každý tento celek byl samostatným dokumentem (dokumentem jazyka některého z našich předních spisovatelů, jistého časového úseku například v publicistice nebo odborné literatuře, tvorby určitého období apod.). Když se na základě tohoto materiálu začal zpracovávat slovník, bylo nutné všech těchto 2 200 abecedních celků sloučit v jeden, čímž byl ovšem ztracen materiálový podklad pro specializované práce. Na tomto příkladu se ukazují právě výhody strojního (dnes již počítačového) zpracování, neboť pomocí nich lze snadno pracovat pouze s částí jakéhokoliv celku (vyjmout například slovní zásobu autorů spřízněných lokálně, tematicky, generačně), tyto části dále zpracovávat (například uspořádat slovník podle stáří dokladů, což může být výhodné pro studium historie jazyka apod.), třídit podle různých kritérií (například podle

[62] *Vocabulairo usual, comun y fundamental*. Madrid 1952.
[63] *Recuento de Vocabulario Español*. Puerto Rico 1952.
[64] *Frequency Dictionary of Spanish Words*. The Hague 1964.
[65] Mistrík, J.: *Frekvencia slov v slovenčine*. Bratislava 1969.

charakteru pramenů – próza, poezie, publicistika), provádět gramatická studia (třídění podle slovních druhů, gramatických kategorií atd.). Navíc všechny tyto seznamy by bylo možno velmi jednoduše doplnit o frekvenční údaje a vytvořit tak frekvenční slovník. Snadné by bylo rovněž sestavení např. slovníku retrográdního, v němž jsou slova abecedně uspořádána podle pořadí písmen od konce.

Od roku 1924 se provádějí první kvantitativní výzkumy v oblasti českého lexika, a to zejména pro účely didaktické či těsnopisné. Pro účely těsnopisné studoval na základě frekvence slov frekvenci hlásek ve slovenštině a v češtině A. Holas, dále například J. Sedláček[66]. Z prací vzniklých k účelům didaktickým si zmínku zaslouží nepublikovaný *Slovník šestiletých dětí*, který z podnětu V. Příhody sestavili Th. Hamerský (na základě výzkumu ve Zlíně roku 1944) a M. Bartušková (na základě výzkumu v Praze v roce 1945).

Roku 1938 vznikla první kvantitativní práce z oblasti jazyka publicistiky[67], a to na základě excerpce jednoho čísla deníku *Národní osvobození* (celkem 42 448 slov). Marie Těšitelová[68] roku 1948 provedla kvantitativní rozbor knihy Karla Čapka *Život a dílo skladatele Foltýna*.

První frekvenční slovník češtiny byl vydán roku 1961 v Praze pod názvem *Frekvence slov, slovních druhů a tvarů v českém jazyce* (standardně se označuje zkratkou FSČ). Autory byli J. Jelínek, J. V. Bečka a M. Těšitelová. Původně byl slovník určen pro účely didaktické, ale toto své určení daleko překročil, takže slouží i ke studiu kvantitativních poměrů v češtině vůbec. Slovník vznikal ručně po dobu více než 20 let na základě 1 623 527 slov získaných ze 75 textů. Celkem je v něm obsaženo 54 486 slov různých. Kdyby autoři neměli k dispozici četné pomocníky, dílo by vůbec nemohlo vzniknout. FSČ rozlišoval osm funkčních stylů – čtyři umělecké (beletrie, básně, literatura pro mládež, dramata), čtyři odborné (populárně vědecká literatura, noviny, vědecká literatura, mluvené projevy). Zastoupení bylo ovšem nerovnoměrné: beletrie zaujímala 30 % dokladů, mluvené projevy jen 6 %. Vedle frekvence slov se ve FSČ uvádí například i frekvence některých gramatických kategorií (např. slovních druhů, pádů substantiv apod.).

Ve stejném roce vznikl spoluprací Oddělení pro teorii strojového překladu Filosofické fakulty Karlovy univerzity a strojně početní stanice Vysoké školy ekonomické v Praze frekvenční slovník matematických textů. Slovník obsahoval 51 341 slov představujících asi 3 050 různých lexikálních jednotek. Byl zpracován pomocí děrných štítků. Pro účely strojového překladu byl na stejných pracovištích vytvořen frekvenční slovník anglických elektrotechnických textů. Rovněž byl sestaven pomocí děrných štítků. Na štítek se vedle anglického slova uváděl i slovní druh a český ekvivalent (překlad), třídit slova pak bylo možno jak podle slov anglických, tak podle českých, což je jistě významná pomůcka při řešení homonymie a synonymie.

[66] *Poměry slov a jejich částí v češtině, získané rozborem 18 000 slov.* Těsnopisné rozhledy 2, 1924, s. 65–66.

[67] Korejs, J. – Korejsová, H.: *Slovník našich novin.* Praha 1938.

[68] *Frekvence slov a tvarů ve spise „Život a dílo skladatele Foltýna" od Karla Čapka.* Naše řeč 32, 1948, s. 126–130.

Z dalších prací podobného charakteru jmenujme alespoň v roce 1969 v Praze vydaný *Konkordanční a frekvenční index k Slezským písním Petra Bezruče*, i ve světovém kontextu ojedinělé dílo *Biblická konkordance*[69] a další práci J. V. Bečky, který zkoumal složení českých odborných textů technického zaměření na materiálu 175 000 slov ze 7 vědních oborů: lékařství, technická chemie, zemědělství, strojírenství, báňský a hutní průmysl, elektrotechnika, stavitelství[70]. Vedle úvodu, v němž se probírají základní otázky techniky práce i některé výsledky, obsahuje práce frekvenční seznam slov a slovních skupin.

2.3.2 Těsnopis

Jednou z oblastí, která z praktických důvodů využívala výzkumy frekvence jazykových jednotek, byl již od 19. století *těsnopis*.

Těsnopis (též *stenografie*, *rychlopis*, *tachygrafie*) je písmo umožňující psát rychlostí mluvené řeči. Podle B. Trnky se těsnopis liší od obvyklé grafické reprodukce řeči časovou a prostorovou úsporností, „*má tudíž k obyčejnému písmu poměr písma pomocného, prostředkujícího mezi ním a mluveným, po případě myšleným slovem.*“ ([76], str. 7)

Novodobé rukopisné těsnopisné soustavy se většinou rozdělují na dvě skupiny:

1) *soustavy geometrické*, ve kterých jako znaky vystupují většinou geometricky nejjednodušší tvary (přímka v různých polohách, kružnice a její části, smyčka, klička, tečka apod.) a tvary vzniklé jejich spojováním; tyto soustavy začaly vznikat na přelomu 16. a 17. století v Anglii.

2) *soustavy kurzívní*, které se snaží co nejvíce přiblížit obyčejnému psacímu písmu pravidelným střídáním pohybů shora dolů a zdola nahoru, používají i znaků geometricky složitějších. Tyto systémy začaly vznikat na počátku 19. století zejména v Německu. Důraz byl kladen na to, aby tyto systémy byly co nejplynulejší.

Vedle rukopisných těsnopisných systémů vznikly v moderní době i soustavy strojové (klávesnicové), které jsou rozšířeny zejména v USA a západní Evropě.

Za předstupeň těsnopisu můžeme považovat tzv. *akropolský kámen* z doby kolem 350 př. n. l. (objeven roku 1883), na němž byly nalezeny zásady pro tvorbu zkrácených znaků. Kolem roku 300 př. n. l. vznikly tzv. *delfské desky*, na kterých neznámý autor uvádí seznam znaků nahrazujících skupiny dvou až tří hlásek.

Za objevitele těsnopisu je považován Marcus Tullius Tiro (103–4 př. n. l.), nejdříve otrok, později Ciceronův tajemník, který k označení jednotlivých slov užíval tzv. *not* (odtud název „*tironské noty*“). Tiro vyšel ze starší soustavy

[69] Bič, M. – Souček., J. B.: *Biblická konkordance*. Ústřední církevní nakladatelství Praha, 1954–1967; též Bič, M. – Souček., J. B.: *Biblická konkordance (3 sv.)*. Kalich, Praha 1961, 838 s.; 1963, 777 s.; 1967, 820 s.

[70] Bečka, J. V.: *Lexikální složení českých odborných textů technického zaměření I.* Praha 1973; Týž: *Lexikální složení českých odborných textů technického zaměření II.* Praha 1975.

značek, kterými se zkracovala slova v obyčejném písmu. Poprvé byl tento těsnopisný systém použit v Římě 5. prosince 63 př. n. l. k zapsání Catonovy obžalovací řeči proti Catilinovi. Těsnopis se užíval i ve vojsku ve funkci tajného písma. Ve starověkém Římě se zpracovávaly stenografické slovníky, učebnice a prováděla se různá zlepšení. Po rozpadu Římské říše roku 476 n. l. se těsnopisný systém přenesl i do Galské a Francké říše. Užíval se zejména v klášterech, ale jeho složitost (velké množství not) způsobila, že se v průběhu 10. a 11. století užívat přestal. V Itálii se vytvořil slabikový těsnopis (6. – 8. století, novější verze 8. – 11. století). Ačkoliv byl kratší než obyčejné písmo, byl méně rychlý než slovní písmo, což bylo důvodem k jeho zániku.

Ve středověku se pak výrazně uplatnila *zkrácená latina*. Za autora tohoto zkráceného písma je považován mnich John of Tilbury, který v 11. století na podnět arcibiskupa z Canterbury Thomase Becketa zkoumal systém *tironských not*. Ačkoliv vyzdvihl systém tohoto zkráceného písma, konstatoval obtížnost těchto not. Sestavil tedy nové latinské slovní písmo, v němž kmeny označoval hlavními notami, koncovky notami pomocnými.

Na vývoj novodobého těsnopisu mělo velký vliv tajné písmo (*steganografie*). Autorem jednoho z nejpropracovanějších systémů byl londýnský lékař, později venkovský farář Timothy Bright (1551–1615), který roku 1588 vydal *Characterie* (tj. umění krátce, rychle a tajně psát pomocí charakterií), první knihu anglického rychlopisu. Podobnou soustavu o dva roky později vypracoval Peter Bales (1547–1610?) v knize *The Art of Brachygraphy*. Brzy je ale zastínil John Willis (1575–1625). Tento londýnský farář vydal roku 1602 knížku o 100 stranách malého formátu *Art of Stenographie* (jen do roku 1647 byla vydána celkem 14krát). Autor zde také poprvé použil označení „*těsnopis*“[71]. Zavedl pevnou abecedu, jejíž znaky byly vytvořeny z přímky a oblouků v různé poloze, tečky, kruhu a smyčky (první geometrická těsnopisná soustava). Při volbě znaků přihlížel údajně i k frekvenci.

Při své návštěvě Anglie se s Willisovou těsnopisnou soustavou seznámil i Jan Amos Komenský. V jeho latinsky psaném dopise *Ad amicos Lesznae in Polonia agentes* z Londýna 8. října 1641 pak nacházíme nejstarší zmínku o těsnopise u českého autora (viz [31]): „*Valná část mladších i starších mužů zaznamenává si kázání perem, a to doslova. Bylo tu totiž před 30 lety (za Jakuba) vynalezeno a již i na venkov proniklo rychlopisné umění, jež oni nazývají steganografie, jímž (ne pomoci písmen, nýbrž značek značících celá slova) rychlost jazyka rukou napodobují. Učí se pak onomu umění ve městech skoro všichni, jakmile se naučili obyčejnému písmu ve škole, přidávajíce asi rok k naučení steganografie.*“ Komenský se tu patrně seznámil s Brightovou, popř. Balesovou soustavou tajného písma, neboť v textu užívá pojem steganografie (tajnopis), což by zdůvodňovalo i existenci zvláštních značek pro celá slova. Ve svém spise *De rerum humanarum emendatione consultatio catholica* pak Komenský uvádí nejen návrh na jednotný mezinárodní dorozumívací jazyk (tzv. *univerzální jazyk*), ale je rovněž autorem skici těsnopisné soustavy pro tento jazyk.

Od počátku 19. století se začíná významně uplatňovat v těsnopisu pojem *frekvence*. Těsnopisci si začali uvědomovat důležitost jazykového popisu a právě

[71] Viz [48].

těsnopis byl podnětem pro různé jazykové rozbory. Například zakladatel německého kurzívního těsnopisu Franz Xaver Gabelsberger (1789–1849) získával při sestavování své těsnopisné soustavy podněty jednak studiem těsnopisné literatury, jednak praktickou stenografickou činností. Při volbě znaků pak přihlížel především k frekvenci – přiděloval častým hláskám znaky jednodušší a psatelnější[72]. Těsnopis byl inspirací pro první frekvenční slovník na světě, a to pro frekvenční slovník F. W. Kädinga z roku 1897[73]. Statistická šetření publikovaná v tomto slovníku a týkající se frekvence slov, slabik i hlásek v německé řeči se stala cennou pomůckou při kodifikaci Stolzeovy-Schreyovy soustavy v roce 1897.

Kvantitativní vlastnosti jazyka za účelem sestavení optimální těsnopisné soustavy zkoumal rovněž Francouz Jean-Baptiste Estoup (1868–1950). Ve Francii založil *Institut International de Métagraphie Duployé* (odnož těsnopisné školy Duployé, jedné z nejrozšířenějších těsnopisných soustav ve Francii). Při vyučování těsnopisu se uplatňují jeho cvičení, která obsahují kombinace vysoce frekventovaných slov (tzv. *Estoupovy stenografické gamy*). S jeho jménem je spjat rovněž tzv. *Zipfův-Estoupův zákon* (více viz kap. 2.6).

Soustava GESS (= Gosudarstvenaja jedinaja sistema stenografii) N. N. Sokolova, která byla 10. června 1933 sovětskou vládou zavedena jako jediná státní těsnopisná soustava, se opírala rovněž o rozsáhlý vědecký výzkum v oblasti biomechaniky, jazykovědy, matematiky apod.[74]

Na Slovensku přední místo zaujímají práce Jozefa Mistríka. Ten publikoval celou řadu stenografických prací a jeho dílo je dalším příkladem toho, jak se původně stenografický výzkum[75] stal základem práce s daleko širším záběrem, a to frekvenčního slovníku slovenštiny[76]. Spoluprací J. Mistríka (později Ladislava Bojnanského) s českými těsnopisci K. Matouškem a M. Matulou bylo do slovenského těsnopisu z češtiny přeneseno tzv. krácení II. stupně.

Na základě požadavků stenografů připravili velký frekvenční slovník polští jazykovědci a rovněž například v Bulharsku vznikla řada prací z oblasti statistiky písmen vyvolaná konkurzem na novou bulharskou těsnopisnou soustavu (viz [31]).

Pomineme-li zmínku o steganografii z pera Jana Amose Komenského z roku 1641, první pojednání o těsnopise na našem území roku 1834 publikuje Jan Dalibor Kopecký v pražském časopise *Květy*. Jedná se o stručné seznámení s těsnopisem vůbec a zvláště s Gabelsbergerovou soustavou používanou v Německu, ve kterém vyslovuje přání, aby toto umění bylo přeneseno a pěstováno i u nás.

[72] Zda tak činil na základě nějakého konkrétního statistického výzkumu, se v literatuře neuvádí.

[73] *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz 1897.

[74] Zásady vyložil Sokolov v díle *Těoretičeskije osnovy gosudarstvennoj jedinoj systemy stenografii*. Moskva 1949. Tato kniha patří k jedněm z nejpropracovanějších děl věnovaných základním principům těsnopisu.

[75] *Slovenský stenograf VIII*. 1954, s. 85–86 a 105–107; *Stenografia I*. Bratislava 1955, s. 195–223.

[76] J. Mistrík: *Frekvence slov v slovenčine*. Bratislava 1969.

První českou těsnopisnou soustavu vypracoval Hynek Jakub Heger (1808–1854) ve 40. letech 19. století jako převod německé soustavy F. X. Gabelsbergera, a to nejen pro češtinu, ale i pro hlavní slovanské jazyky (tzv. *těsnopisný panslavismus*). Ačkoliv měl Heger nemalý podíl na rozšíření těsnopisu do Čech, neměla jeho soustava dlouhého trvání. Heger do původní soustavy Gabelsbergerovy z úcty před autorem vnášel totiž jen málo změn a výběr znaků ne úplně vyhovoval češtině.

Roku 1859 vznikl pod vedením Jindřicha Fügnera *První pražský spolek stenografů*, který roku 1861 vypsal soutěž na nejlepší převod Gabelsbergovy soustavy do češtiny (vítěz měl získat 25 dukátů ve zlatě). Z pěti návrhů ale porota žádný nedoporučila. Práce na převodu přesto pokračovaly a na Nový rok 1863 vyšla malá publikace *Těsnopis český* jako výsledek práce sedmi členů spolku stenografů (Eduard Novotný, František Brzobohatý, Josef Faigl, A. Konrad, Josef Novotný, Alois Stáně a František Tauer). Autoři přihlédli k některým zvláštnostem českého jazyka, i když ne úplně důkladně. Chyběli podklady o frekvenci hlásek a jejich kombinací. Ačkoliv se tato práce setkala nejprve s kritikou německých odborných časopisů, byla později znovu vydávána a přepracovávána, a to až do roku 1921, kdy vyšlo 14. vydání této práce.

Autorem první původní soustavy českého těsnopisu byl kněz František Gába (1829–1896), který tvrdil, že český těsnopisný systém by měl být založen na osobitých rysech slovanských jazyků (zejména češtiny) a nemůže být jen převodem německé soustavy. Zdůrazňoval tyto dva požadavky:

1) Písmena by měla být v rámci možností obrazem zvuku, tzn. artikulačně (foneticky) podobné hlásky by měly být zachyceny graficky podobnými znaky. Tento princip se uplatnil později v Trnkově těsnopisu, který byl spjat s vývojem funkčního hláskosloví ve 20. a 30. letech 20. století.

2) Tvorbě soustavy by měl předcházet rozsáhlý jazykový rozbor, zaměřený zejména na poznání kvantitativních vlastností jazyka (zvláště slovníku a grafematické roviny). Počátkem 50. let 19. stol. sám Gába údajně prováděl kvantitativní výzkum, na jehož základě se pokoušel sestavit takovou soustavu, ve které by graficky nejjednodušší znaky odpovídaly nejčastěji se vyskytujícím písmenům.

I když se v praxi nadále osvědčoval převod soustavy Gabelsbergerovy, Gábova myšlenka na původní českou těsnopisnou soustavu nezanikla. Roku 1864 český lékař Jan Bareš vydal spisek *Základy slovanského rychlopisu*, ve kterém se pokoušel vytvořit takovou těsnopisnou soustavu, kterou by bylo možno převést do ostatních slovanských jazyků. „*Konstrukci slovanské abecedy předcházel u Bareše výpočet frekvenčního poměru hlásek v češtině, ovšem na velmi omezeném materiálu.*“ ([48], str. 91) V soustavě se ale vyskytovala řada nedostatků způsobených autorovými neznalostmi slovanských jazyků i nedostatkem praktických zkušeností, byla proto komisí pražského spolku zamítnuta.

Antonín Krondl ve svém referátu *Kterak vyhovuje převod soustavy Gabelsbergerovy hláskoslovné povaze jazyka českého* na třetím sjezdu českých stenografů (konaném 20. a 21. května 1888 v Brně) ukázal, že při sestavování

původní české těsnopisné soustavy bude poznání kvantitativních charakteristik jazyka nezbytné a bude třeba opírat se o přesně zjištěnou *iteraci* (= opakování) a zvláště kombinaci českých hlásek. Sám provedl řadu frekvenčních výzkumů pro práci na své původní české těsnopisné soustavě.

Opomenutá zůstala rukopisná práce Josefa Sedláčka z let 1913–1917 *Základní studie k českému těsnopisu*, která obsahuje jednak frekvenční seznam nejčastějších slov sestavený ze souboru 100 000 slovních výskytů, jednak frekvenci hlásek, hláskových skupin a slabik ze souboru 18 000 slov[77]. Přesněji se jedná o kvantitativní charakteristiku psaných textů stylu veřejných projevů ze začátku 20. století[78].

Statistické výzkumy prováděl i František Hradilík (1871–1939), lektor těsnopisu na brněnské technice. Zúčastnil se veřejné soutěže vyhlášené ministerstvem školství dne 8. července 1920 na původní soustavu československého těsnopisu a jako přípravné práce k ní konal frekvenční výzkumy hlásek a hláskových skupin na 10 000 českých slovech. Iterace a kombinace byly rovněž vědeckým základem soustavy inženýra Svojmíra Mikulíka a PhDr. Aloise Herouta, která se stala vítězem této soutěže na původní český těsnopis (15. června 1921 se porota jednomyslně usnesla na jejím vítězství z celkem 14 soustav). Tato soustava (tzv. *soustava Heroutova-Mikulíkova*) byla jmenována oficiálním českým těsnopisem. 23. června 1921 ministr školství schválil zavádění nové soustavy do škol, přičemž zkoušet podle ní se mělo při učitelských státních zkouškách od letního termínu zkušebního roku 1922. K jejímu rozšiřování a zdokonalování byla zřízena 6. prosince 1921 státní instituce *Státní ústav těsnopisný*.

Soutěže na původní českou těsnopisnou soustavu se zúčastnil mimo jiné i profesor anglické filologie Karlovy univerzity v Praze PhDr. Bohumil Trnka, jeden z nejvýznačnějších světových průkopníků v oblasti funkční fonologie. Trnka se pokusil o vybudování důsledné a lingvisticky odůvodněné těsnopisné soustavy. Byla založena na soustavě distinktivních rysů fonologického systému a jednotlivým hláskám byly přiřazeny grafické znaky podle hláskové iterace (výskytu) a kombinace. To znamená, že se Trnka opíral nejen o kvalitativní výzkum jazyka (reprezentovaný strukturalismem, zejména pak N. S. Trubeckým, V. Mathesiem, R. Jakobsonem, B. Havránkem a jinými členy Pražského lingvistického kroužku), ale i o výzkum stránky kvantitativní. S prvním návrhem[79] své soustavy vystoupil již v roce 1914, přepracovanou ji předložil do ministerské soutěže. Bohužel Trnkova těsnopisná soustava nenašla většího ohlasu, neboť se její návrh objevil až po převládnutí systému Heroutova-Mikulíkova, který je s některými změnami oficiálním těsnopisným systémem dodnes. Vý-

[77] Rozbor práce je zahrnut v knize M. Matuly: *Teorie a praxe těsnopisu I.* Praha 1958, s. 116n.; též poznámka u B. Trnky: *Pokus o vědeckou teorii a praktickou reformu těsnopisu.* Praha 1937, s. 33n.

[78] Pro srovnání s dnešními poměry viz M. Ludvíková – J. Kraus: *Kvantitativní vlastnosti soustavy českých fonémů.* SaS 27, 1966, s. 334n. Sedláček se od nich mírně odlišuje vymezením sledovaných jednotek, např. nerespektováním ztráty znělosti na konci slova, rozlišováním i/y apod.

[79] Podrobně svou koncepci vyložil v [76]. V předmluvě Trnka vyzdvihuje těsnopis jako důležitou dovednost a dále oceňuje jeho vliv na vývoj fonetických teorií či jeho důležitou roli v literární historii.

znam Trnkova návrhu původní těsnopisné soustavy ale spatřujeme v propojení strukturní fonologie, těsnopisu a kvantitativní lingvistiky.

Je tedy vidět, že pro potřeby stenografie byla provedena řada různých jazykových statistik. Přestože tyto statistiky zpracovávaly jen omezený materiál a neodpovídaly dnešním požadavkům na jazykový rozbor, jsou spolu s problematikou sporného autorství považovány za nejstarší zdroje matematické, přesněji kvantitativní lingvistiky.

Pro novější český těsnopisný výzkum je typická spolupráce stenografů s vědci z příbuzných oborů – jazykověda, matematická lingvistika, fyziologie, psychologie psaní apod. Zmiňme zde alespoň Miloše Matulu[80]. Tento vědecky aktivně pracující matematik se zájmem o otázky jazykové je rovněž vynikající praktik v oblasti těsnopisu – je držitelem českého rekordu, kdy v roce 1957 dosáhl rychlosti 200 slov za minutu v pětiminutovém diktátu s odstupňovanou rychlostí. Jen pro představu uveďme, že pro zápis běžné řeči stačí rychlost zápisu 70 slov za minutu, pro rychlejší přednes či například televizní zprávy je to 80 až 100 slov za minutu. Píšeme-li běžným písmem, bývá rychlost zápisu okolo 20 až 40 slov za minutu. Největší Matulův význam ale spatřujeme v rozpracování *teorie krácení slov* (oproti dosavadnímu hledání nejvýhodnější abecedy založené na studiu frekvence). Společně s K. Matouškem vytvořil tzv. *silné krácení* (také *krácení II. stupně*), založené na slovotvorných vlastnostech jazyka, které stenografům umožňuje psát vyšší rychlostí. V první fázi prací na krácení II. stupně posloužila autorům neprávem opomenutá práce Josefa Sedláčka *Základní studie k českému těsnopisu*, později se jako potřebné ukázaly frekvenční statistiky slov a slovních skupin v projevech hospodářských a politických[81]. Na závěr ještě připomeňme, že vedle *krácení II. stupně* se M. Matula zabýval analýzou a měřením procesu stenografování, při němž využíval statistického průzkumu a matematických metod. Z českých lingvistů, kteří se věnovali problematice těsnopisu, jmenujme alespoň J. Krause.

Různé frekvenční výzkumy se prováděly i v pozdějších letech, zejména pod záštitou Státního ústavu těsnopisného, který se v teoretické oblasti zaměřoval především na usnadnění naučitelnosti těsnopisné soustavy Heroutovy-Mikulíkovy. Výsledkem této snahy byla reforma Heroutovy-Mikulíkovy soustavy v roce 1961, která byla založena na několik let trvajících výzkumech frekvenčních, srovnávacích a grafických. Ústav měl rovněž svůj podíl při položení základů *krácení II. stupně*.

V souvislosti s kritickým hodnocením Heroutovy-Mikulíkovy těsnopisné soustavy a s novou vlnou zájmu o otázky frekvence slov, písmen a jejich kombinací prováděl při práci na své nové těsnopisné soustavě STENOP rozsáhlé statistiky například Josef Horák (vedle podrobného studia fyziologie a psychologie psaní)[82]. Ačkoliv v druhé polovině 20. století vznikly nové návrhy na modernější

[80] Viz Matula, M.: *Teorie a praxe těsnopisu*. Praha 1958, 439 s.; *Aplikace matematiky na studium psaní*. Pokroky matematiky, fysiky, astronomie 3, 1958, s. 245n. a 393n. aj.

[81] Viz Čáp, J. – Matoušek, K. – Matula, M. – Petrásek, J.: *Frekvence slov v stenografické praxi*. Praha 1961; Matula, M.: *Frekvence kořenů slov*. Praha 1963.

[82] Viz *Metodika vyučování těsnopisu*. Praha 1967; *Využití statistiky fonosémů pro optimalizaci těsnopisné teorie*. Zprávy SÚT, 1969/70, č. 1, s. 11n.; *Máme těsnopisnou soustavu plně vyhovující duchu českého jazyka?* Zprávy SÚT 1968/69, č. 3, s. 65n.

a vhodnější soustavy českého těsnopisu, než je soustava Heroutova-Mikulíkova, po zkušenostech se zaváděním nového systému a velkou náročností přechodu, který by trval zhruba 10 let, nebyly tyto návrhy využity.

Přestože frekvenční těsnopisné slovníky a výzkumy kladou důraz více na formu slov, než na jeho význam (slučují tedy hesla, která by v lingvisticky pojatých slovnících či seznamech byla rozdělena), je srovnání stenografických a lingvistických statistik možné i užitečné současně.

Jak již bylo uvedeno na začátku, lze těsnopisné systémy rozdělit na ruční a strojové. K nahrazení ručního stenografického písma písmem strojovým slouží *těsnopisný psací stroj*. Strojový záznam zpravidla užívá tzv. *akordů* – současného stisku více kláves. Písmo se otiskuje na úzký proužek papíru, který je navinut na cívce. Po každém úhozu se papír posune o řádku zdola nahoru. Počet kláves na stroji je co nejvíce zredukován. Práce stroje musí být nehlučná, aby nerušila okolí. Těsnopisné stroje se používaly zejména ve Spojených státech a západní Evropě (Anglie, Německo) a začaly se vyrábět i v Sovětském svazu. V USA i západní Evropě je použití těchto strojů běžné, i když dnes bývají nahrazovány počítačem, který pomocí speciálního editoru může softwarově převádět tento záznam do běžného písma. U nás začal na soustavě českého strojového těsnopisu pracovat ke konci 30. let 20. století Jaroslav Vrátný. Navrhoval klávesnici s 24 klávesami, na něž se snažil jednotlivá písmena umisťovat tak, aby to bylo ve shodě s charakterem českého jazyka. Konal proto rozsáhlá frekvenční statistická šetření českých hlásek asi na 20 000 slovech. Vypočítal, že jeho soustavou by bylo možno napsat v průměru 6 slov 7 akordy. Prakticky se výcvikem na Vrátného klávesnici zabýval Jaroslav Junek. K výrobě těsnopisných psacích strojů podle Vrátného prototypu však u nás nedošlo a dodnes nenašly těsnopisné psací stroje v rámci české těsnopisné praxe uplatnění. Převládá ruční těsnopisný zápis v kombinaci s moderní audiotechnikou. Pokusy vyvinout pro češtinu strojovou těsnopisnou soustavu zatím skončily neúspěchem.

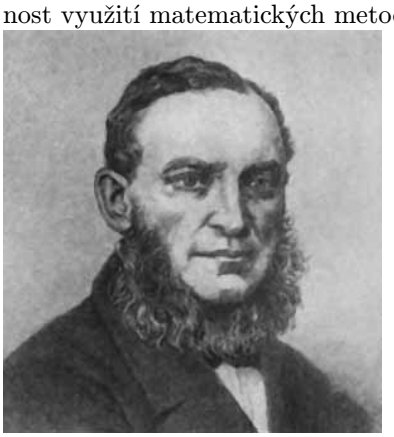
Ačkoliv se může zdát, že těsnopis je v dnešní době v důsledku zdokonalování moderních technologií na ústupu, není to docela pravda. Aktivně se těsnopis dosud používá při veškerých jednáních Parlamentu České republiky (například při jednáních Poslanecké sněmovny Parlamentu České republiky bývá přítomno 10 stenografů). Stenografický zápis z jednání je zakotven přímo v jednacím řádu Parlamentu ČR, a to z toho důvodu, aby byly zachovány zápisy ze všech parlamentních jednání. Bohužel se těsnopis dnes již na žádné střední škole mezi povinnými předměty nevyučuje (lze si jej zvolit jako kroužek na Obchodní akademii v Dušní ulici v Praze pod vedením Heleny Janovské). Zájemci o výuku těsnopisu jsou tak odkázáni buď na samostudium z knih, nebo mohou absolvovat tříměsíční kurz Základy těsnopisu či pracovat s programem Těsnopis MINI, z nichž poslední dvě zmiňované možnosti jsou zaštítěny Státním těsnopisným ústavem.

V současnosti je těsnopisná problematika u nás náplní práce dvou institucí. První z nich je *Státní těsnopisný ústav*, který byl založen roku 1921 za účelem rozšiřování a zdokonalování oficiální těsnopisné soustavy Heroutovy-Mikulíkovy. Od roku 1999 je součástí Národního ústavu odborného vzdělávání v Praze. Koordinuje výuku předmětů technika administrativy, obchodní kore-

spondence a těsnopis na středních školách. Podílí se na mezinárodní spolupráci v rámci federace *Intersteno* (*Mezinárodní federace zpracování informací*) založené roku 1887. Naše republika je jejím členem od roku 1960 (tehdy ještě jako součást Československa). Interstemo pořádá každé dva roky mezinárodní kongresy, v jejichž rámci se koná mistrovství světa v psaní na stroji (od roku 1955) a mezinárodní soutěže v těsnopise (od roku 1961). Soutěží se vždy odděleně v jednotlivých jazykových skupinách. Prezidentem Interstema je pro dvouleté období vždy příslušník toho státu, v němž se má konat příští kongres. Vedle Státního těsnopisného ústavu v naší republice pracuje *Český těsnopisný spolek*[83], který byl ustaven 5. března 2001 v Praze a navázal na přerušenou činnost *Prvního pražského spolku stenografů* založeného roku 1859 na podnět Jindřicha Fügnera.

2.4 Viktor Jakovlevič Bunjakovskij

Tento významný ruský matematik zasáhl rovněž do prehistorie matematické lingvistiky. Pravděpodobně jako první totiž upozornil v roce 1847 na možnost využití matematických metod v lingvistice.



Viktor Jakovlevič Bunjakovskij se narodil 16. 12. 1804 v Baru (dnešní Ukrajina) a zemřel 12. 12. 1889 v Petrohradu (Sankt Petěrburg). Základní vzdělání získal Bunjakovskij v Moskvě v domě hraběte A. P. Tormasova, jenž byl přítelem jeho otce. Společně s druhým synem hraběte studoval Bunjakovskij od roku 1820 v cizině – nejprve v Coburgu, potom v Lausanne a nakonec v Paříži (zde studoval na Sorbonně a Collège de France, a měl tak možnost seznámit se s přednáškami například Laplace, Fouriera či Poissona). V roce 1824 získal Bunjakovskij titul bakaláře, 1825 byl jmenován doktorem matematiky na Přírodovědecké fakultě (Collège de France). Po návratu do Ruska v roce 1826 začal vyučovat matematiku v Petrohradě na různých školách. V roce 1828 byl zvolen aspirantem Akademie věd, roku 1830 byl zvolen mimořádným členem Akademie a 1836 řádným členem Akademie věd. V letech 1846–1859 přednášel na petrohradské univerzitě (analytická mechanika, diferenciální a integrální počet či teorii pravděpodobnosti). Od roku 1864 až do doby krátce před svou smrtí byl viceprezidentem Akademie věd. Roku 1883 Bunjakovskij, který se převážně zabýval oblastí teorie čísel a teorií pravděpodobnosti, publikoval seznam svého vědeckého díla[84] čítající 108 vědeckých prací.

Roku 1847 publikoval Bunjakovskij v časopise *Sovremennik*[85] svůj článek s názvem *O vozmožnosti vvedenija opredelitelnych mer doverija k rezultatam*

[83] Více viz na internetových stránkách www.tesnopis.cz.
[84] *Liste des travaux mathématiques de Victor Bouniakowsky.* SPb 1883.
[85] Vol. 3, part II.

nekotorych nauk nabljudatelnych, i preimuščestvenno statistiki (angl. *On the possibility to apply determining measures of confidence to the results of some observing science, particularly statistics*). Opomineme-li význam, který má tento článek z historického hlediska na rozvoj statistiky, je to nejspíše jedna z prvních zmínek, ve které je uváděna možnost a důvody využití statistiky mimo jiné k řešení filologických otázek.



О ВОЗМОЖНОСТИ

ВВЕДЕНІЯ ОПРЕДѢЛИТЕЛЬНЫХЪ МѢРЪ ДОВѢРІЯ КЪ РЕЗУЛЬТАТАМЪ НѢКОТОРЫХЪ НАУКЪ НАБЛЮДАТЕЛЬНЫХЪ, И ПРЕИМУЩЕСТВЕННО СТАТИСТИКИ.

Obrázek 2.1: Článek V. J. Bunjakovského týkající se matematické lingvistiky

V samém závěru článku totiž Bunjakovskij píše, že ačkoliv by na tomto místě mohl být jeho výklad ukončen, přece jen by rád dodal pár slov týkajících se jisté aplikace pravděpodobnostní analýzy, na kterou zřejmě ještě nikdo neupozornil. Tato nová aplikace zahrnuje gramatické a etymologické studium jazyka či srovnávací filologii. Jakkoliv zvláštní se mohou tyto oblasti pro aplikaci matematických metod na první pohled zdát, je Bunjakovskij zcela přesvědčen, že se tu otevírá široký prostor pro jejich užití. Své tvrzení nezakládá na nejistých dohadech a domněnkách, ale na kritickém zhodnocení objektů, na řadě pokusů, které již údajně provedl, a na analytických formulích, které zavedl za účelem definovat numerickou pravděpodobnost jednotlivých slovních formací. Zdůrazňuje ale, že zde nechce představovat detaily věci, ale pouze na tuto možnost upozornit.

Dále uvádí, že k tomu, aby mohla být využita aplikovaná matematika, je třeba nejprve provést detailní aritmetické popisy a statistiky jazyka, např. početní tvrzení o úplném inventáři slov daného jazyka, o jejich rozlišení na jednotlivé slovní druhy, statistiky počtu písmen, počty počátečních písmen apod. Musí být tedy pojmenována hlavní pravidla, výjimky různého druhu, slova přejatá z cizích jazyků apod. Teprve na tomto numerickém materiálu lze provádět přesnou analýzu, od níž zcela přirozeně požadujeme matematickou vážnost. Podobným způsobem zpracovaná statistická data týkající se různých jazyků pak lze srovnávat z různých hledisek a získané výsledky tak mohou získat respekt, který filologové při současném stavu vědy dosud nemají.

Upozorňuje rovněž na to, že vytvoření takové statistiky jazyka si vyžádá ohromné množství práce, které možná mohou někteří filologové považovat za zbytečné, neboť podle některých z nich předpokládaný zisk v podobě přesnosti

závěrů o jazyce nevyváží tuto časovou náročnost. Do jaké míry je tato námitka oprávněná, nechává Bunjakovskij k samostatnému posouzení[86].

2.5 Andrej Andrejevič Markov



Další osobností, která se významnou měrou zasloužila o rozvoj matematické lingvistiky, byl Andrej Andrejevič Markov (1856–1922), ruský matematik, známý především díky svým pracem z oblasti teorie pravděpodobnosti (kde zobecněním nezávislých pokusů dospěl k tzv. *markovským řetězcům*), dále teorie čísel a matematické analýzy. Do dějin matematické lingvistiky se zapsal průkopnickou statí *Primer statističeskogo issledovanija nad tekstom „Jevgenija Onegina", illjustrirujuščij svjaz ispytanij v cep* z roku 1913, která je považována za vůbec první důslednou aplikaci matematických poznatků v lingvistice. V této práci Markov na ruském textu *Evžena Oněgina* statisticky zkoumal výskyt souhlásek a samohlásek a na zjištěné hodnoty následně aplikoval svou teorii markovských řetězců. Mluvení lze totiž chápat jako „*proces, který spočívá v tom, že k jednotlivým jazykovým jednotkám už vysloveným neustále postupně připojujeme jednotky nové, a to podle pravidel jejich relativní frekvence, která jsou po daný jazyk závazná*" ([6]). Odtud je blízko nejen k matematické statistice a teorii pravděpodobnosti, ale i k teorii informace (napodobování textu uplatněním pravděpodobnosti výskytu písmen, jak si ukážeme dále, souvisí s jedním ze základních pojmů teorie informace, a to s pojmem *entropie*). Podstatné je ovšem to, že Markovova práce vůbec poprvé podnítila zájem matematiků o jazykovědnou problematiku (a nepochybně i lingvistů o otázky matematické) a významně se tak zasloužila o rozvoj matematické lingvistiky. Ukazuje se totiž, že nejlepších výsledků je možno dosáhnout jen vzájemnou spoluprací odborníků příslušných hraničních disciplín.

2.5.1 Život a dílo A. A. Markova

Andrej Andrejevič Markov se narodil 14. června 1856 v Rjazani v Rusku. Jeho otec Andrej Grigorjevič Markov pracoval jako státní úředník a později správce majetku. Byl dvakrát ženatý. Z prvního manželství s Naděždou Petrovnou měl 6 dětí[87] (mezi nimi i Andreje), z druhého manželství s Annou Josifovnou měl tři další děti[88]. Vedle Andreje se proslavil ještě jeho nevlastní bratr Vladimír (1871–1897), který ačkoliv zemřel velmi mlád na tuberkulózu,

[86] Bunjakovského článek se později stal předmětem rozboru v referátu P. B. Struveho předneseném na zasedání sekce historických nauk a filologie 18. 5. 1918. Viz [62].

[87] Petr, Pavel (umřel v dětském věku), Marie, Jevgenie, Andrej a Michail.

[88] Vladimír, Lýdie a Jekatěrina.

byl rovněž uznávaným matematikem. Starší sestra Jevgenie byla jednou z prvních ruských lékařek (specializace v oboru psychiatrie) a lékařkou se stala také jeho nevlastní sestra Jekatěrina. Povolání středoškolské učitelky vykonávala sestra Lýdie.

Od dětství se A. A. Markov potýkal se zdravotními problémy – tuberkulóza kolenního kloubu způsobila nepohyblivost kolene jedné nohy. Tento handicap se nepodařilo zcela odstranit ani úspěšnou operací, kterou podstoupil v deseti letech. Ačkoliv Markov po zbytek života na tuto nohu lehce napadal, dlouhé procházky patřily k jeho oblíbeným činnostem.

Počátkem 60. let 19. století se Markovův otec stal správcem majetku Jekatěriny Alexandrovny Valvatjevové a celá rodina se přestěhovala do Petrohradu. Zde v letech 1866–1874 navštěvoval Andrej 5. petrohradské gymnázium. Na toto období nevzpomínal příliš rád. Nevyhovoval mu styl výuky (přísnost, biflování) a ve většině předmětů se špatně učil – pouze v matematice měl samé jedničky. Začal sám matematiku hlouběji studovat. V době gymnaziálních studií objevil jistý způsob integrování obyčejných lineárních diferenciálních rovnic s konstantními koeficienty, o němž se domníval, že je zcela nový. Tento svůj objev představil významným ruským matematikům tehdejší doby: Bunjakovskému, Zolotarevovi a Korkinovi. První z nich na dopis gymnazisty Markova neodpověděl, ale druzí dva mu podrobně vysvětlili, proč tento způsob není ve skutečnosti nový. Tak se Andrej seznámil s profesory petrohradské univerzity A. N. Korkinem a E. I. Zolotarevem.

V roce 1874 začal studovat na Fyzikální a matematické fakultě petrohradské univerzity. Zapsal se do semináře vedeného právě Korkinem a Zolotarevem. Navštěvoval rovněž jejich semináře pro nadané studenty, kde bez problémů řešil úlohy zde předkládané. Právě diskuze s Korkinem ho inspirovaly k řadě jeho samostatných prací. Velký vliv, který spatřujeme v celé jeho pozdější odborné práci, měl na Markova Pafnutij Lvovič Čebyšev, tehdejší vedoucí katedry matematiky. 31. května 1878 získal Markov titul kandidát (první stupeň vědecké hodnosti) a v témže roce mu byla udělena zlatá medaile za nejlepší studentskou práci s názvem *O integraci diferenciálních rovnic pomocí řetězových zlomků* (*Ob integrirovaniji differencialnych uravnenij pri pomošči neprerývnych drobej*). 25. dubna 1880 obhájil dizertační práci *O kladných binárních kvadratických formách*, Čebyševem vysoce ceněnou, která Markova zařadila mezi nejlepší ruské matematiky a která představuje jeden ze znamenitých úspěchů petrohradské školy teorie čísel. Pro zajímavost si přibližme, jak hluboko Markov v této práci do problematiky racionálních aproximací pronikl; i přestože byla práce prakticky ihned publikována francouzsky v *Mathematische Annalen* a touto problematikou se v tehdejší době zabývali nejpřednější teoretici čísel z Francie i Německa, mezi západními matematiky byla problematika této práce zvládnuta až v letech 1910–1920 díky berlínským matematikům F. G. Frobeniovi a Robertu Remakovi.

Po získání doktorského titulu v roce 1880 začal Markov vyučovat na petrohradské univerzitě jako soukromý docent (v letech 1880/81 a 1881/82 vedl kurzy diferenciálního a integrálního počtu; v roce 1883 převzal po Ju. V. Sochockém a K. A. Possem úvod do analýzy; po odchodu Čebyševa v roce 1883

poprvé vyučoval kurz teorie pravděpodobnosti, který od školního roku 1885/86 již vyučoval každý rok). 9. února 1885 obhájil svoji docentskou práci a v následujícím roce byl jmenován externím profesorem petrohradské univerzity, v roce 1893 pak profesorem řádným. Roku 1905 odešel Markov do důchodu s titulem zasloužilý profesor, nicméně ve výuce kurzu teorie pravděpodobnosti pokračoval téměř do konce života. Jako vyučující měl na své studenty poměrně značné nároky – ve svých kurzech žákům představoval mnoho nových vědeckých poznatků, často na úkor tradiční problematiky. Vedle toho byl Markov členem petrohradské akademie věd (25. prosince 1886 byl na návrh Čebyševa zvolen aspirantem petrohradské akademie, 11. února 1890 externím akademikem a 14. března 1896 akademikem řádným).

V roce 1883 se A. A. Markov oženil s Marií Ivanovnou Valvatjevovou (1860–1942), jednou ze dvou dcer Jekatěriny Alexandrovny Valvatjevové, u níž byl Markovův otec zaměstnán jako správce. Zpočátku nebyla Jekatěrina Alexandrovna tomuto vztahu nakloněna, neboť se jí Andrej Andrejevič nezdál dostatečně finančně zabezpečený. Když se ale stal soukromým docentem na univerzitě a dokončoval svou doktorskou dizertaci (obdoba docentury, zisk 1885) a perspektivní se jevila i profesura, dala Jekatěrina Alexandrovna souhlas. Manželé Markovovi neměli dlouho své vlastní děti, proto přijali do rodiny tři děti z příbuzenstva Andreje Andrejeviče, kteří ztratili rodiče. 22. září 1903 (to bylo Marii Ivanovně již 43 let) se konečně manželům Markovovým narodil syn, který byl pojmenován po otci Andrej Andrejevič a který se stal rovněž uznávaným matematikem.

Ke koníčkům Andreje Andrejeviče Markova patřilo fotografování a šachy. Patřil k nejlepším ruským šachistům své doby a dosáhl v této oblasti značných úspěchů. Například v roce 1890 se stal vůbec prvním vítězem speciálního šachového turnaje, organizovaného M. I. Čigorinem, zakladatelem ruské šachové školy. Tento turnaj měl velmi přísná pravidla: hrálo jej 12 hráčů losem rozdělených do dvou skupin po šesti, každý hráč z jedné skupiny hrál dvě utkání (jednou s černými figurkami a podruhé s bílými) se všemi hráči druhé skupiny, každý tah musel být odehrán do dvou dnů bez možnosti čas kumulovat. V Markovově pozůstalosti se dochovalo okolo tisíce dopisů s 45 vynikajícími šachisty tehdejší doby, a to nejen z Ruska, ale i z různých míst Evropy, na jejichž základě bylo zrekonstruováno přes 100 neznámých šachových partií.

Podle pamětníků byl Markov člověk čestný a věrný svému přesvědčení, také ale upřímný a odvážný. Celý život důsledně bojoval s hloupostí, a to často bez ohledu na možné důsledky. Když byl například v roce 1902 zvolen čestným členem Akademie věd literát A. M. Gorkij a jeho zvolení bylo vzápětí na základě nařízení cara Mikuláše II. z politických důvodů zrušeno, Markov na tuto situaci ihned reagoval dopisem Akademii věd. Protože ale tehdejší sekretář akademie záměrně dopis založil, aniž by byl na shromáždění akademie veřejně čten, požádal Markov rovněž o vyloučení z akademie. Jeho žádost nebyla ovšem přijata a on tak mohl pokračovat v přípravě vydání Čebyševových sebraných spisů (za což byl velmi rád, neboť jako jeho žák a nejbližší pokračovatel měl obavy svěřit toto dílo do cizích rukou). O rok později ale odmítl v souvislosti s vyloučením Gorkého přijmout vyznamenání udělované carskou vládou. Po-

dobně v únoru 1912 na protest proti vyloučení velkého ruského spisovatele L. N. Tolstého z pravoslavné církve, zažádal Markov pravoslavný synod rovněž o své vyloučení z církve. Tentokrát bylo jeho žádosti po řadě humorných peripetií vyhověno (jako dostatečný důvod pro vyloučení uvádí svou knihu *Počet pravděpodobnosti*, v níž vyjadřuje svůj záporný postoj k příběhům, tvořícím základ hebrejského a křesťanského náboženství). A na závěr uveďme třetí příklad, kdy v roce 1913 na protest proti oslavám třístého jubilea vlády rodu Romanovců připravil Markov jubileum „naučné", konkrétně dvousté výročí zákona velkých čísel.

Velkou pozornost věnoval Markov také problematice výuky matematiky na středních školách. Razantně vystupoval proti různým experimentům v této oblasti. Ve školním roce 1917/18 dokonce sám na střední škole vyučoval. Když byl totiž na návštěvě příbuzných ve městě Zarajsk (rjazaňská gubernie), zůstaly neplánovaně vyšší ročníky zdejšího reálného učiliště bez matematika. Markov požádal Akademii věd o umožnění studijního pobytu a celý rok strávil s rodinou v Zarajsku, kde bezplatně vyučoval na této škole. Jeho syn se tak stal jeho oficiálním žákem, neboť tehdy právě navštěvoval šestý ročník.

Na podzim roku 1918 se rodina vrátila zpět do Petrohradu. Zhoršující se těžké dědičné onemocnění oka si vyžádalo operační zákrok, po kterém se zrak zlepšil a Markov znovu začal přednášet na petrohradské univerzitě. Jeho celkový zdravotní stav již ale nebyl zdaleka uspokojivý (a přejezd ze Zarajska do Petrohradu se na něm rovněž negativně odrazil). Ve školním roce 1920/21 ho na jeho přednášky z teorie pravděpodobnosti doprovázel syn, neboť Markov se sotva udržel na nohou. V tuto dobu s vypětím sil pracoval na čtvrtém vydání své knihy *Počet pravděpodobnosti*, které se výrazně odlišuje od vydání předchozích. Bohužel vyšlo až posmrtně v roce 1924. Na podzim roku 1921 Markova upoutal na lůžko těžký zánět míšních nervů, ke kterému se na jaře následujícího roku přidalo aneuryzma v noze doprovázené krvácením. Protože toužil po přírodě, na radu ošetřujícího lékaře se rozhodl pro pobyt v sanatoriu, ale náročná cesta vlakem měla na jeho zdravotní stav neblahý vliv. Krvácení zesílilo a Markov se musel neprodleně vrátit do Petrohradu kvůli operaci k oddálení aneuryzmy. Tato operace zpočátku pomohla, ale za několik dní došlo k prudkému zhoršení způsobenému celkovou infekcí krve. 20. července 1922 v 22 hodin Markov v Petrohradě umírá.

Rané práce akademika Markova, jednoho z nejvýraznějších představitelů tzv. *petrohradské matematické školy* (dále například P. L. Čebyšev, A. M. Ljapunov), se týkaly především *matematické analýzy* (např. limity integrálů, teorie aproximací a konvergence řad, konstruktivní teorie funkcí, diferenciální rovnice) a *teorie čísel*. Ačkoliv je jeho prací z teorie čísel poměrně málo – 15, měly v tomto oboru značný význam. Zabýval se v nich především teorií neurčitých kvadratických forem, důkazy transcendentnosti čísel e a π a konečně studiu čistě kubických polí.

Nejzásadnější je ale jeho přínos do *teorie pravděpodobnosti*. Z přibližně 120 původních vědeckých prací, jichž je Markov autorem, se více než jedna třetina týká právě teorie pravděpodobnosti. Stal se v nich pokračovatelem myšlenek

svého učitele P. L. Čebyševa (například po roce 1900 Markov aplikoval metodu řetězových zlomků, jejímž průkopníkem byl právě Čebyšev, na teorii pravděpodobnosti). V korespondenci A. A. Markova s A. A. Čuprovym (několik desítek dopisů věnovaných obecným otázkám matematické statistiky) se můžeme dočíst, jak se postupně utvářel postoj k této matematické discíplíně od negativního až ke kladnému. Jeho práce z teorie pravděpodobnosti se dají rozčlenit do tří hlavních skupin, a to:

1) *zákon velkých čísel*,

2) *centrální limitní věty*,

3) *teorie markovských řetězců*.

Teorie markovských řetězců je zcela nové odvětví teorie pravděpodobnosti, k jehož vzniku Markova vedly úvahy o závislých náhodných veličinách. Později se tato teorie rozvinula v rozsáhlou a velmi důležitou oblast – *teorii stochastických procesů*. Markovským řetězcem rozumíme posloupnost náhodných proměnných, v nichž je budoucí proměnná určena proměnnou současnou bez ohledu na způsob, jakým přítomný stav ze stavu předchozího vznikl. V roce 1923 se Norbert Wiener stal prvním, kdo přesně a souvisle pojednával o markovských procesech. Vznik obecné teorie je připisován Andreji Nikolajeviči Kolmogorovi v průběhu 30. let 20. století. Název *markovské řetězce* byl údajně poprvé použit francouzským vědcem Jacquesem Hadamardem (1865–1963) a toto označení se rozšířilo do odborné literatury. Vůbec poprvé Markov podstatu řetězové závislosti objasňuje ve své práci *Rozšíření zákona velkých čísel na veličiny navzájem závislé* (*Rasprostranenije zakona bolšich čisel na veličiny, zavisjaščije drug ot druga*) z roku 1907, samozřejmě ještě bez užití termínu *řetězec*. Jednoduše a srozumitelně tu objasňuje princip toho, co je dnes nazýváno *stejnorodé markovské řetězce*, a dále zde dokazuje zákon velkých čísel pro posloupnost náhodných veličin spojených v řetězec za předpokladu existence konečných rozptylů u jednotlivých sčítanců.

Vedle vědecké práce věnoval Markov značnou pozornost přípravě svých přednášek do tisku (úvod do analýzy, sférická trigonometrie, diferenciální počet, teorie pravděpodobnosti aj.). V nich můžeme vysledovat zvláštní rysy Markovovy osobnosti, jeho chápání předmětu, představy o úrovni znalostí, jež by studenti měli mít. Co se týče stylu jeho prací je charakteristická srozumitelnost jazyka, pečlivé zpracování detailů (dovedení řešení k číslu a algoritmu), ale i nastínění možností praktického použití.

2.5.2 Hlásková statistika románu Evžen Oněgin

Prvotní impuls provést statistický výzkum výskytu souhlásek a samohlásek v ruských textech vzešel z Markovovy korespondence s A. A. Čuprovym. Statistické šetření prováděl Markov celkem na dvou textech, v románu A. S. Puškina *Evžen Oněgin* (*Jevgenij Onegin*) a v povídce Sergeje T. Aksakova (1791–1859) *Detskije gody Bagrova-vnuka*, a to v podstatě z důvodů didaktických. Hlavním cílem tohoto statistického rozboru bylo totiž přiblížit čtenářům teorii *prostých*

markovských řetězců na obecně srozumitelném příkladu. Zajímavé je to, že ačkoliv Markov rozvinul svou teorii markovských řetězců jako čistě matematickou teorii, aplikoval ji na literární text. S největší pravděpodobností se tedy jedná o první důslednou aplikaci matematiky v jazykovědě. Nejdříve prováděl Markov rozbor textu *Evžena Oněgina* (*Jevgenij Onegin*). Pojďme si jej tedy představit detailněji.

Konkrétně bylo zkoumáno 20 000 ruských hlásek (bez „*tvrdého jeru*" a „*měkkého jeru*") veršovaného románu A. S. Puškina *Evžen Oněgin*, které tvoří celou první kapitolu a 16 slok kapitoly druhé. Jak Markov postupoval?

Nechť těchto 20 000 hlásek tvoří posloupnost 20 000 za sebou následujících pokusů, z nichž každý může nabývat dvou hodnot, a to hodnoty samohláska nebo souhláska. Připusťme existenci konstantní pravděpodobnosti p, která vyjadřuje pravděpodobnost toho, že daná hláska je samohláskou. Tu odhadneme z výsledků pozorování pomocí počtu samohlásek v textu. Stejně odhadneme i pravděpodobnosti p_1, p_0, $p_{1,0}$, $p_{0,1}$, $p_{0,0}$, $p_{1,1}$, kde:

p_1 je pravděpodobnost, že samohláska následuje za samohláskou;
p_0 je pravděpodobnost, že samohláska následuje za souhláskou;
$p_{1,0}$ je pravděpodobnost, že samohláska následuje za souhláskou, které předchází samohláska;
$p_{0,1}$ je pravděpodobnost, že samohláska následuje za samohláskou, které předchází souhláska;
$p_{0,0}$ je pravděpodobnost, že samohláska následuje za dvěma souhláskami;
$p_{1,1}$ je pravděpodobnost, že samohláska následuje za dvěma samohláskami.

Analogicky odhadneme z textu pravděpodobnosti q, q_1, q_0, $q_{1,0}$, $q_{0,1}$, $q_{0,0}$, $q_{1,1}$, kde:

q je pravděpodobnost, že daná hláska je souhláskou;
q_1 je pravděpodobnost, že souhláska následuje za samohláskou;
q_0 je pravděpodobnost, že souhláska následuje za souhláskou;
$q_{1,0}$ je pravděpodobnost, že souhláska následuje za souhláskou, které předchází samohláska;
$q_{0,1}$ je pravděpodobnost, že souhláska následuje za samohláskou, které předchází souhláska;
$q_{0,0}$ je pravděpodobnost, že souhláska následuje za dvěma souhláskami;
$q_{1,1}$ je pravděpodobnost, že souhláska následuje za dvěma samohláskami.

Nyní porovnáme hodnoty získané z textu s předpokládanými teoretickými hodnotami, a to ve dvou případech:

1) výskyt hlásek odpovídá nezávislým pokusům,

2) hlásky jsou uspořádány v *řetězec*.

ad1) Předpokládejme, že výskyt samohlásek a souhlásek v textu jsou nezávislé pokusy. To odpovídá situaci, kdy z osudí, v kterém jsou bílé a černé koule (v našem případě samohlásky a souhlásky), opakovaně táhneme s tím, že vytaženou

kouli vždy vracíme zpět do osudí. 20 000 hlásek textu rozdělíme na 200 skupin po 100 hláskách (bez porušení pořadí) a spočítáme, kolik je v každé stovce samohlásek. Získáme tak následující tabulku:

Počet samohl. ve skupině	37	38	39	40	41	42	43	44	45	46	47	48	49
Počet skupin	3	1	6	18	12	31	43	29	25	17	12	2	1

Tabulka 2.5: Počet samohlásek v 200 skupinách po 100 hláskách (bez porušení pořadí hlásek)

Z tabulky snadno určíme aritmetický průměr, který je roven 43,19, a odtud pak odhadneme pravděpodobnost výskytu samohlásky, kde

$$p \cong 0{,}4319 \cong 0{,}432.$$

Nyní vypočteme součet druhých mocnin odchylek počtu samohlásek každé stovky od aritmetického průměru a dostaneme číslo,

$$1022{,}8,$$

odkud po vydělení číslem 200 získáme hodnotu rozptylu (pro skupinu sta hlásek), tj.

$$5{,}114.$$

Nezávislost stanovených veličin (tj. počet samohlásek v stovce hlásek za sebou následujících) potvrzuje i ten fakt, že jestliže je spojíme po dvou, čtyřech a pěti (vytvoříme tedy skupiny po dvou stech, čtyřech stech a pěti stech hláskách po sobě následujících) a vypočítáme pro těchto sto, padesát a čtyřicet posloupností sumu kvadrátů jejich odchýlení od 86,4; 172,8 a 216 (aritmetický průměr počtu samohlásek mezi dvěma sty, čtyřmi sty a pěti sty hláskami), získáme hodnoty 827,6; 975,2; 1004, které se mnoho neliší od výše získané hodnoty 1022,8.

Přejdeme-li ale od stovek pokusů k jednotlivým pokusům, pozorujeme, že číslo

$$\frac{5{,}114}{100} = 0{,}05114$$

(zjištěná hodnota rozptylu) se výrazně liší od hodnoty

$$p(p-1) = 0{,}432 \cdot 0{,}568 = 0{,}245376$$

(teoretická hodnota rozptylu za předpokladu, že pokusy jsou nezávislé). Koeficient disperze je v tomto případě roven

$$\frac{5114}{24537{,}6} \cong 0{,}208,$$

což naznačuje „*vázanost*“ našich pokusů (tato hodnota se výrazně liší od hodnoty 1, která odpovídá nezávislým pokusům). Vidíme tedy, že sumy vázaných

veličin (po sobě následující hlásky textu) mohou tvořit téměř nezávislé veličiny (skupiny po sto hláskách). Později se ukázalo, že tento příklad je značně přínosný pro teorii spojů.

Vytvoříme nová spojení po 100 hláskách. Nejprve si uspořádáme každou původní stovku hlásek do čtvercové matice takto:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
91	92	93	94	95	96	97	98	99	100

Z každé pětistovky hlásek vytvoříme pět nových stovek hlásek následujícím způsobem: první stovku sestavíme ze všech prvních a šestých sloupců, druhou z druhých a sedmých sloupců atd. Nyní se vedle sebe nevyskytují žádné dvě sousední hlásky z původních stovek. Dále spočteme, kolik je samohlásek v jednotlivých sloupcích. Tato čísla vždy po dvou sečteme takto: počet samohlásek v prvním a šestém sloupci, v druhém a sedmém sloupci, třetím a osmém sloupci, čtvrtém a devátém sloupci, pátém a desátém sloupci. Pro každou stovku písmen tak získáme pětici čísel, označenou symboly $(1,6)$, $(2,7)$, $(3,8)$, $(4,9)$, $(5,10)$, jejichž součet nám dá počet samohlásek dané stovky. Počet samohlásek v nových stovkách pak odpovídá sumám

$$\sum(1,6), \sum(2,7), \sum(3,8), \sum(4,9), \sum(5,10),$$

které jsou složeny z odpovídajících pěti sčítanců. (Všechny tyto výsledky uvádí Markov ve 40 přehledných tabulkách – každá tabulka odpovídá pětistovce hlásek. V prvním řádku tabulky je pět čísel $(1,6)$ a jejich suma, v druhém řádku pět čísel $(2,7)$ a jejich součet atd., poslední řádek uvádí počet samohlásek v první stovce, počet samohlásek v druhé stovce atd., až na posledním místě počet samohlásek ve všech pěti stovkách zmenšený o dvě stě.) Sestavíme tabulku četností samohlásek v jednotlivých stovkách podobnou té předešlé:

Počet samohl. ve skupině	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
Počet skupin	1	0	0	0	1	2	1	3	5	1	2	9	13	12	13	11
Počet samohl. ve skupině	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
Počet skupin	17	16	15	10	10	16	10	10	5	5	3	3	3	0	1	2

Tabulka 2.6: Počet samohlásek v 200 skupinách po 100 hláskách (s porušením pořadí hlásek)

Aritmetický průměr je stejný, a to 43,19. Součet kvadrátů jejich odchylek od 43,2 je značně větší než předešlý, a to

$$5788{,}8.$$

Odtud získáme hodnotu rozptylu

$$\frac{5788{,}8}{200} \cong 28{,}944$$

pro těchto 200 posloupností. Přejdeme-li nyní od stovek hlásek k jednotlivým hláskám, dostaneme hodnotu

$$0{,}28944,$$

která se mnoho neliší od teoretické hodnoty rozptylu pro nezávislé náhodné veličiny

$$0{,}432 \cdot 0{,}568 = 0{,}245376.$$

Koeficient disperze je pak roven

$$\frac{28944}{24537{,}6} \cong 1{,}18,$$

což se rovněž příliš neodlišuje od teoretické hodnoty koeficientu disperze, která je pro nezávislé náhodné veličiny rovna 1. Na závěr ještě tyto nové stovky hlásek rovněž pospojujeme po dvou, čtyřech a pěti. Jestliže pro sto, padesát a čtyřicet skupin vypočítáme postupně hodnotu sumy druhých mocnin jejich odchýlení od aritmetických průměrů počtu samohlásek v dané skupině, tj. od 86,4; 172,8; 216, získáme místo čísla 5788,8 hodnoty 3551,6; 3089,2; 1004, z nichž poslední je téměř šestkrát menší než číslo 5788,8, což poukazuje na původní vázanost v každé pětistovce hlásek.

ad 2) Uvažujme nyní závislost jednotlivých pokusů. To odpovídá situaci, kdy jsou dána dvě osudí, jedno bílé a druhé černé. V každém z těchto osudí jsou bílé a černé koule (samohlásky a souhlásky). Postupně konáme tahy následujícím způsobem. Vytáhneme-li při n-tém tahu kouli bílé barvy, konáme $(n+1)$-ní tah z bílého osudí, vytáhneme-li při n-tém tahu kouli černé barvy, konáme $(n+1)$-ní tah z černého osudí. Vytažené koule se vždy vloží zpět do osudí.

K vysvětlení závislosti nám může posloužit přibližné vyjádření pravděpodobností p_1, p_0. Mezi 20 000 hláskami zkoumaného textu se objeví posloupnost „*samohláska – samohláska*" celkem 1104krát. Po vydělení počtem všech samohlásek v textu získáme

$$p_1 = \frac{1104}{8638} \cong 0{,}128.$$

Podobně najdeme

$$p_0 = \frac{7534}{11361} \cong \frac{7534}{11362} \cong 0{,}663.$$

Všimněme si, že pravděpodobnost toho, že hláska je samohláskou, se výrazně liší podle toho, zda jí předchází souhláska ($p_0 = 0{,}663$) nebo samohláska ($p_1 = 0{,}128$). Už tento fakt naznačuje jistou závislost. Označme si dále

$$\delta = p_1 - p_0 = 0{,}128 - 0{,}663 = -0{,}535.$$

Připusťme, že naše posloupnost 20 000 hlásek tvoří prostý řetězec, pak je teoretický koeficient disperze pro $\delta = -0{,}535$ (v souladu s Markovovou prací *Issledovanije zamečatelnogo slučaja zavisimych ispytanij*) roven

$$\frac{1+\delta}{1-\delta} = \frac{465}{1535} \cong 0{,}3;$$

to se sice neshoduje úplně s dříve získaným číslem 0,208, ale přibližuje se mu podstatně více, než číslo 1, odpovídající případu nezávislých pokusů.

Nyní připusťme, že posloupnost hlásek tvoří „*složitý*" řetězec a použijeme výsledky práce *Ob odnom slučaje ispytanij svjazanych v složnuju cep*. Sečteme, kolikrát se v posloupnosti hlásek objevuje posloupnost „*samohláska – samohláska – samohláska*" a „*souhláska – souhláska – souhláska*". Odtud dostáváme:

$$p_{1,1} = \frac{115}{1104} \cong 0{,}104$$

$$q_{0,0} = \frac{505}{3827} \cong 0{,}132.$$

Abychom mohli aplikovat výsledky uvedené Markovovy práce, předpokládejme, že:

$$p \cong 0{,}432,$$

$$q \cong 0{,}568,$$

$$p_1 \cong 0{,}128,$$

$$q_1 \cong 0{,}872,$$

$$p_0 \cong 0{,}663,$$

$$q_0 \cong 0{,}337,$$

$$p_{1,1} \cong 0{,}104,$$

$$q_{0,0} \cong 0{,}132.$$

A na základě těchto hodnot najdeme

$$\delta = p_1 - p_0 \cong -0{,}535,$$

$$\varepsilon = \frac{p_{1,1} - p_1}{q_1} = \frac{-24}{872} \cong -0{,}027,$$

$$\eta = \frac{q_{0,0} - q_0}{p_0} = \frac{-205}{663} \cong -0{,}309.$$

Zjištěné hodnoty dosadíme do vzorce pro výpočet koeficientu disperze

$$\frac{\{q(1-3\varepsilon)(1-\eta)+p(1-3\eta)(1-\varepsilon)-2(1-\varepsilon)(1-\eta)\}(1-\delta)+2(1-\varepsilon\eta)}{(1-\delta)(1-\varepsilon)(1-\eta)} =$$

$$= \frac{1+\delta}{1-\delta}\left\{\frac{1+\varepsilon}{2(1-\varepsilon)}+\frac{1+\eta}{2(1-\eta)}\right\}+\frac{(q-p)(\eta-\varepsilon)}{(1-\varepsilon)(1-\eta)}$$

a získáme hodnotu

$$0{,}195,$$

která je již velmi blízká nalezené hodnotě

$$0{,}208.$$

Nemůžeme tvrdit, že příklad vyhovuje zcela teoretickým podmínkám, na druhou stranu je zřejmé, že tato výrazná shoda teoretické hodnoty koeficientu disperze s hodnotou získanou z výsledků měření není náhodná.

V tomtéž roce Markov rozšířil svá pozorování týkající se střídání souhlásek a samohlásek v ruské literatuře na 100 000 hlásek pověsti S. T. Aksakova *Detskije gody Bagrova-vnuka* a získal tyto hodnoty: $p_1 = 0{,}552$, $p_0 = 0{,}365$, $\delta = 0{,}187$. Markov byl se svými výsledky spokojen a usoudil, že jeho pozorování dostatečně dobře potvrdilo shodu skutečného pozorování hlásek s hypotézou existence prosté řetězové závislosti.

To, že se o problematiku využití matematických metod v lingvistice zajímal i později, svědčí například práce *Ob odnom primeneniju statističeskogo metoda* z roku 1916. Markov zde reaguje na stať N. A. Morozova *Lingvističeskije spektry*, která vyšla v 20. díle *Izvestij russkogo jazyka i slovesnosti*. Práce pojednává o využití „statistické metody" ke studiu jazyka různých spisovatelů. Morozov zkoumal útržky textu po tisíci slovech z děl L. N. Tolstého, A. S. Puškina, N. V. Gogola a I. S. Turgeněva. V těchto úryvcích počítal výskyt různých slov a na základě jejich počtu se snažil určit charakteristiky jazyka jednotlivých spisovatelů. Markov tuto myšlenku považuje za velmi zajímavou a konstatuje, že podobné práce mohou mít v budoucnu velký význam. Důležité je ovšem uvědomit si, že jako důkaz „stability" zjištěných výsledků a jejich obecné platnosti je odkaz na obecný zákon velkých čísel či shoda s jinými výsledky nedostatečná. Zároveň Markov konstatuje řadu základních chyb. Není tu žádný náznak toho, že získané výsledky jsou typické pro tyto ruské spisovatele a nevztahují se pouze na těchto několik málo útržků. Součty, které provádí Markov sám, jsou často v rozporu se závěry Morozova (např. Gogolův jazyk se podle Morozova vyznačuje značnou převahou výskytu předložky „*na*" oproti předložce „*v*", Markov ale nachází v první tisícovce slov knihy *Mrtvé duše* předložku „*na*" 12krát a „*v*" dokonce 37krát, což podle tabulky odpovídá jazyku Puškina). Markov konstatuje, že zjištěné výskyty slov neopravňují autora k závěrům o individuálním stylu autora, neboť záměna jedné tisícovky slov jinou může tyto závěry zcela převrátit. Jistý stupeň opodstatnění těchto závěrů by byl možný, sčítal-li by autor ne pět tisícovek slov (někdy i méně), ale sto tisíc

slov (pokud by se ovšem neukázala druhá poměrně pravděpodobná situace, že výsledky všech autorů se budou blížit k jednomu a témuž číslu, podléhajíce obecným zákonům jazyka). Na závěr Markov kritizuje pomocné prostředky, jež Morozov ve své práci využívá, jakými je např. převádění jedné tabulky na druhou tabulku pomocí zvláštních dělitelů ($\frac{1}{26}$, $\frac{1}{20}$ apod.) či náčrtky, které podle Markova podstatu díla nezmění.

V kapitole *Teorie informace a lingvistika* v [58] provedl P. Novák pokus s nápodobou českého textu podle teorie pravděpodobnosti a výsledky porovnal se srovnatelnými výsledky R. L. Dobrušina pro ruštinu, G. Shannona pro angličtinu a W. Meyer-Epplera pro němčinu (nápodoba textu daného jazyka za předpokladu, že všechna písmena mají stejnou frekvenci, s přihlédnutím k relativní frekvenci jednotlivých písmen, relativní frekvenci dvojic písmen a relativní frekvenci trojic písmen). Získal tyto výsledky (viz tab. 2.7).

1. Za předpokladu, že všechna písmena v textu mají stejnou frekvenci:
   čeština: ďj mrgučxýďyaýweaožá
   ruština: сухерробьбщ яыхвщиюайжтл
   angličtina: xfoml rxkhrjff juj zlpwcfwkkcyj
   němčina: aiobnin tarsfneoulpiitdregedcoads

2. S přihlédnutím k relativní frekvenci jednotlivých písmen:
   čeština: žia ep atndi zéuořmp
   ruština: еынт цияьа оерв однг
   angličtina: ocro hli rgwr nmielwis eu ll
   němčina: eragepterprteiningeit gerelen re

3. S přihlédnutím k relativní frekvenci dvojic písmen:
   čeština: lí di oneprá sguluvicechupsv
   ruština: умароно кач всванный рося ных
   angličtina: on ie antsoutinys are t inctore
   němčina: billunten zugen hin se sch wel

4. S přihlédnutím k relativní frekvenci trojic písmen:
   čeština: dves a vaše miléklár
   ruština: покак пот дчрноскака наконепно
   angličtina: in no ist lat whey cratict froure
   němčina: eist des nich in den plassen kann

Tabulka 2.7: Nápodoba českého, anglického a německého textu dle zákonů teorie pravděpodobnosti

Zde je vidět, že pokud uvažujeme u všech písmen stejnou frekvenci jejich výskytu, dostáváme text, který prakticky vůbec nepřipomíná text příslušného jazyka. Při rostoucím přirozeném čísle n ale dostáváme text, který se textu daného jazyka přibližuje. Pokusy v některých jazycích ukázaly, že již při $n = 32$ se takto vzniklý text prakticky rovná textu skutečného jazyka.

2.6 George Kingsley Zipf



O kvantitativní lingvistice se někdy mluví také jako o Zipfově lingvistice podle amerického lingvisty a psychologa německého původu George Kingsleyho Zipfa z Harvardovy univerzity. Zipf zkoumal ve 20. a 30. letech 20. stol. relativní frekvenci hlásek a došel k několika zajímavým závěrům – např. čím je hláska artikulačně obtížnější, tím menší je její frekvence; ve všech jazycích jsou neznělé hlásky přibližně dvakrát častější než znělé. Rovněž se zajímal o psychologické a fyziologické faktory ovlivňující produkci a percepci řeči[89]. Hlavní zásadou je podle něj tendence k co nejmenší námaze mluvčího. Tuto svou teorii nazývá *psychobiologií* a navrhuje vyčlenit v lingvistice zvláštní disciplínu, tzv. *biolingvistiku*[90]. Největší Zipfův přínos ale spatřujeme ve třech zákonech (nazýváme je *Zipfovy zákony*), které upozorňují na vztah mezi frekvencí slov a jejich pořadím, dále na vztah mezi frekvencí slova a počtem různých slov, která tuto frekvenci mají, a na vztah mezi frekvencí slova a počtem jeho významů.

První Zipfův zákon:

$$r \cdot f = k$$

(součin frekvence slova a jeho ranku je konstantní), kde
r ... rank (pořadí slova v seznamu podle klesající frekvence, kde jednomu údaji o frekvenci odpovídá pouze jeden rank),
f ... frekvence slova,
k ... konstanta.

[89] Viz např. *Relative Frequency as a Determinant of Phonetic Change*. 1929.

[90] Více viz *The Psychology of Language. An Introduction to Dynamic Philology*. Boston 1935; *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. 1949.

Druhý Zipfův zákon:

$$a \cdot b^2 = k$$

(počet slov o jisté frekvenci krát frekvence na druhou je konstantní), kde
a ... počet slov,
b ... frekvence,
k ... konstanta.

Třetí Zipfův zákon:

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

(slova s vysokou frekvencí mají zpravidla větší počet významů), kde
m ... počet významů daného slova,
f ... frekvence,
k ... konstanta.

Nyní se zastavíme u jednotlivých zákonů podrobněji.

2.6.1 První Zipfův zákon

Říká nám, že součin frekvence slova a jeho ranku je konstantní, tedy

$$r \cdot f = k.$$

Již v roce 1912 se ve 3. vydání svých *Gammes sténographiques*[91] pokusil francouzský stenograf J. B. Estoup formulovat zákonitost při uspořádání slov podle klesající frekvence. Zjistil, že v každém delším kontextu je součin frekvence slova f a jeho ranku r přibližně stejný, tedy

$$f \cdot r = k.$$

Zipf došel k podobnému závěru nezávisle na Estoupovi, proto se někdy v literatuře setkáváme s označením *zákon Estoupův-Zipfův*, častější je ale označení „*první Zipfův zákon*".

Ve většině dostupné literatury se uvádí jako datum uveřejnění tohoto výsledku G. K. Zipfem rok 1949, W. Plath ale ve svém článku [49] uvádí, že ve skutečnosti se jednalo již o rok 1929, když ve své dizertační práci Zipf uvedl vzorec

$$X \cdot Y = n$$

(později $r \cdot f = k$) a přiložil týž graf.

Zipf vychází z hypotézy, že v jazyce existují tendence udržet rovnováhu mezi frekvencí slova a počtem slov, které tuto frekvenci mají. Jde o dvě protichůdné síly. *Síla sjednocující* (*force of unification*) působí na to, aby slova měla co nejvyšší frekvenci a jejich počet byl co nejmenší – v ideálním případě snížit počet slov na jedno s frekvencí 100 %. Proti ní působí *síla rozlišující* (*force of diversification*), jejímž cílem je, aby v jazyce bylo co nejvíce slov různých s poměrně

[91] 4. vyd. 1916, 5. vyd. 1917.

nízkou frekvencí (ideálně 1). První z nich je způsobena tzv. *ekonomií mluvčího* (a *principem minimálního úsilí* při formulaci výpovědi, který pojmenoval analogicky podle principu nejmenší akce ve fyzice), druhá pak je způsobena tzv. *ekonomií posluchače* (*minimální úsilí* při slyšení a porozumění). Tato interpretace byla ale jen zřídka přijímána jako teorie vysvětlující rozložení vztahu pořadí a četnosti.

Podle Zipfa je tedy součin frekvence slova a jeho ranku konstantní (mezi frekvencí slova a jeho rankem je nepřímo úměrná závislost), přičemž tato konstanta závisí na délce textu. Zipf jej vyjadřuje rovnicí:

$$r \cdot f = k,$$

kde
r ... rank slova (pořadí slova při uspořádání slov podle klesající frekvence bez zřetele k počtu slov s touž frekvencí),
f ... frekvence slova,
k ... konstanta.

Zipf své teorie ověřoval na různých textech, například na románu Jamese Joyce *Odysseus* (*Ulysses*), který je tvořen 260 430 slovy, z nichž je 29 899 slov různých (průměrná slovní zásoba dospělého člověka se odhaduje na 3 000 až 10 000 slov). Při studiu tohoto textu Zipf zjistil, že součin ranků s příslušnými frekvencemi jsou přibližně stejná čísla: 26 530 (10. slovo v pořadí má frekvenci 2 653), 27 800 (50. slovo v pořadí má frekvenci 566), 26 500 (100. slovo v pořadí má frekvenci 265), 26 000 (1000. slovo má frekvenci 26) apod., tedy čísla blížící se konstantě. Zipf rovněž tvrdí, že se zákon osvědčil i při ověřování na textu publicistickém[92], při aplikaci na starou angličtinu i na texty cizojazyčné (čínština, gótština aj.).

Pravidelnost tohoto vztahu, ověřená Zipfem na textech řady jazyků, upoutala pozornost spousty badatelů, neboť se zdálo, že by mohla poskytovat klíč k nějakému obecnému principu jazykového chování. Zákon byl ale také podroben značné kritice, největší pozornost mu věnoval (zejména se zřetelem k jeho odvození) francouzský matematik B. Mandelbrot. Ukázal, „*že Zipfův vzorec sice udává „obecný spád" křivek, ale velmi špatně zobrazuje podrobnosti*"[93]. Podrobný rozbor vztahu mezi rankem a frekvencí slova se podle něj řídí tzv. *harmonickým* a tzv. *kanonickým* zákonem.

Harmonický zákon:

$$p_r = \frac{P}{r},$$

kde
r ... rank (pořadí),

[92] Na materiálu R. C. Eldridge, který analyzoval denní americký tisk o rozsahu 43 989 slov, z nichž bylo 6 002 slov různých.

[93] *Structure formelle des textes et communication*. Word 10, 1954, s. 1–27; český překlad *Komunikace a formální struktura textů*. In: Teorie informace a jazykověda, s. 130–150.

p_r ... příslušná četnost výskytu,
P ... konstanta pro každý text.
Jde v podstatě o tzv. první Zipfův zákon, kdy četnost slova je nepřímo úměrná jeho ranku.

Kanonický zákon:

$$p_r = P(r+\rho)^{-\beta}$$

nebo

$$\log p_r = \log P - \beta \log(r+\rho),$$

kde
β ... konstanta pro daný text,
ρ ... konstanta korigující frekvenci slov s nízkým pořadím,
P ... konstanta (charakteristika rozsahu výběru),
r ... rank (nabývá pro každou jednotku jiné hodnoty a je vždy nižší než počet všech různých slov, jejichž celkový počet se značí R),
p_r ... příslušná četnost výskytu.

Harmonický zákon je speciálním případem zákona kanonického (pro $\beta = 1$ a $\rho = 0$). Mandelbrot se pokusil také kanonický zákon vysvětlit pomocí předpokladů vzatých částečně z teorie informace.

Vysvětlit rozložení r a f, které by se opíralo o jiné hypotézy, než jaké definoval Zipf, se v pozdější době pokoušela řada autorů. Významné jsou zejména práce Mandelbrotovy, v nichž předkládá své výsledky v rámci tzv. *makrolingvistiky*. Tou rozumí nové odvětví lingvistiky, jež se má zabývat studiem jazykových jevů velkého rozsahu pomocí statistických metod. Vztah *makrolingvistika* versus *mikrolingvistika* (gramatika) je podle Mandelbrota podobný vztahu termodynamika versus mechanika jednotlivých molekul plynu: ačkoliv je popis na makroskopické úrovni slučitelný s mikroskopickým chováním popisovaným gramatikou, detailních rysů na nižší úrovni si nevšímá. I když je pouze neúplným popisem chování, může bez obtíží vést k početním výsledkům, které by prakticky nebylo možno obdržet z úvah o jednotlivých gramatických jevech. „*Mandelbrot naznačuje, že makrolingvistika může poskytnout podobný cenný nástroj pro popis hrubých rysů textového materiálu velikého rozsahu, jehož úplné, podrobné gramatické zpracování by bylo nezvládnutelně obsáhlé a složité.*“ ([49], s. 27) Navrhuje úpravu Zipfova vzorce tak, aby věrněji odpovídal pozorovaným faktům. Zavádí dva nové parametry ρ a B a s jejich pomocí získává tzv. *kanonický zákon*:

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

(je zde uveden v té podobě, v níž se objevuje v dřívějších Mandelbrotových článcích z let 1954 a 1955), kde r značí pořadí, p_r je relativní četnost slova pořadí r, P, ρ, B jsou konstanty pro daný text (ρ udává korekci pro slova nízkého pořadí).

Mandelbrotovi se podařilo odvodit kanonický zákon matematicky ze dvou různých teoretických modelů generování textů. Podle prvního jednoduššího modelu se slova textu vytvářejí písmeno za písmenem pomocí Markovova procesu

s konečným počtem stavů, v němž má každý symbol (i mezera) jistou stanovenou pravděpodobnost výskytu. Generování textu tímto způsobem vede k rozložení četností slov, které odpovídá kanonickému zákonu s B větším než 1. Druhý Mandelbrotův model byl vypracován analogicky k poměrům v termodynamice; matematicky určuje „*nejpravděpodobnější stav*" textu, který je podroben dvěma omezením: 1) všechna slova musí být během dekódovacího procesu oddělena mezerou, 2) musí být stanoven náklad optimálního dekódování (tj. systému dekódování, v němž nejkratší řada operací je přiřčena nejčastějšímu slovu atd.). Toto pojetí má za následek maximalizaci *entropie* (v duchu Shannonovy teorie informace) charakterizující rozložení pravděpodobností slov a opět vede k tomu, že rozložení četností slov je v souhlasu s kanonickým zákonem, tentokrát bez omezení pro hodnoty B. Mandelbrot dává zřejmě druhému modelu přednost a na jeho základě vyvozuje některé závěry – např. ten, že slova jsou základní jednotky textů, dále že teorie informace má pro lingvistiku zásadní význam.

Z dalších modelů zmiňme model Simonův (1955)[94], který považuje vytváření textu za stochastický „*proces generování*" a na základě tohoto předpokladu odvozuje distribuční funkci, která určuje vztah mezi počtem slovních typů o dané četnosti a jejich četností výskytu. Mandelbrot[95] vytýká Simonovi to, že odvozování se děje vlastně kruhem. Belevitch[96] a Somers[97] zase vyslovují názor, že k odvození prvního Zipfova zákona stačí pouze předpoklad, že logaritmy relativních četností slov mají normální rozložení. Aproximací useknutého normálního rozložení pomocí prvního nebo prvního a druhého členu Taylorovy řady odvozuje Belevitch postupně *Zipfův zákon* a *Mandelbrotův kanonický zákon*.

Mnoho úsilí věnoval Zipfovu zákonu i G. Herdan[98], který ale odmítal mluvit o zákonu, protože už P. Guiraud[99] uvedl, že se jedná o empirickou formuli. Touto problematikou se zabýval i P. Novák[100], který tvrdí, že pravdivost takových vzorců je třeba dále prověřovat. Jedině pokud je možno takovéto vztahy odvodit, tak se zapojí do soustavy dané vědní oblasti a nezůstávají pouhými empirickými pravidly.

Prvním Zipfovým zákonem se u nás podrobně zabývala i M. Těšitelová, která se jej pokoušela ověřit pro češtinu. Ukázalo se, že příslušné vztahy platí dobře ve střední části frekvenčního slovníku, ale ne u slov s frekvencí vysokou a nízkou. K ověřování použila těchto textů:

1. K. Čapek: *Život a dílo skladatele Foltýna* (21 963 slov)

2. M. Pujmanová: *Předtucha* (24 357 slov)

[94] Simon, H. A.: *On a Class of Skew Distribution Functions*. Biometrika, sv. 42, s. 425–440.
[95] Mandelbrot, B.: *A Note on a Class of Skew Distribution Functions: Analysis and Critique of a Paper by H. A. Simon*. Information and Control, sv. 2, 1959, s. 90–99.
[96] Belevitch, V.: *On the Statistical Laws of Linguistic Distributions*. Annales de la Société Scientifique de Bruxelles, sv. 7B, 1959, s. 310–326.
[97] Somers, H. H.: *Analyse Mathematique du Langage*. Louvain 1959.
[98] Herdan, G.: *Type-Token Mathematics*. The Hague 1960.
[99] Guiraud, P.: *Problèmes et méthodes de la statistique linguistique*. Paris 1960.
[100] Novák, P.: *Význam kvantitativních metod pro lingvistiku*. In: [58].

3. I. Olbracht: *Bratr Žak* (29 803 slov)

4. *Slovník šestiletých dětí* – Praha (51 981 slov)

5. *Slovník šestiletých dětí* – Zlín (109 061 slov).

Podle šetření M. Těšitelové „*nabývá poměr mezi rankem a frekvencí slova konstantnosti v tom úseku, kde se rank slova kryje s jeho pořadím, tj. kdy za sebou následují slova s různou frekvencí a kdy se rozdíl v klesající frekvenci blíží 1. S rostoucím počtem slov s touž frekvencí, tj. se zvětšující se disproporcí mezi rankem a pořadím slova, podmíněném klesající frekvencí, hodnota „konstanty" výrazně klesá a pozbývá své platnosti zhruba u pásma slov s nejnižší frekvencí, tj. 10–1. Vzhledem k podmínce, aby se rank slova kryl s pořadím, posunuje se platnost vztahu mezi rankem a frekvencí slova mezi slova s relativně vysokou frekvencí. V jistých frekvenčních intervalech, tj. při stejném ranku, především v pásmu slov se střední frekvencí, ale i v pásmu slov s frekvencí nižší a nejnižší, vytvářejí se frekvenční roviny R, vymezené stejným rankem a stejnou frekvencí. Ty se spojují v celé frekvenční bloky B, které se vyznačují relativní stabilností ranku a frekvence, srov.*

$$B = \sum_{r_i}^{r_n} R_1 + R_2 + \ldots + R_n,$$

kde r_i je rank, od něhož začíná první frekvenční rovina, r_n je rank odpovídající poslední frekvenční rovině v bloku R_n, R_1 je rovina daná rankem r_1 a frekvencí f_1 ([66], str. 46n.).“

2.6.2 Druhý Zipfův zákon

Vyjadřuje, že počet slov o jisté frekvenci krát frekvence na druhou je konstantní, a zapisujeme jej

$$a \cdot b^2 = k.$$

Poprvé byl uveřejněn v *The Psycho-Biology of Language* (Boston 1935). Zipf tomuto svému zákonu nepřipisuje obecnou platnost, omezuje jej pouze na slova s nepříliš velkou frekvencí, která představují značnou část slovníku. Slova s vysokou frekvencí z tohoto zákona vylučuje. Rovněž tvrdí, že tento zákon platí pro všechny jazyky.

Na českém materiálu (stejně jako u *prvního Zipfova zákona*) se pokusila M. Těšitelová ověřit jeho platnost. Ukázalo se, že tento zákon platí dobře rovněž u slov ve středním frekvenčním pásmu (a nevztahuje se ale na slova s vysokou frekvencí, ani na slova s frekvencí relativně malou). Druhý Zipfův zákon vyjadřuje v podstatě známou skutečnost, že počet slov o vysoké frekvenci je vždy malý, zatímco čím je frekvence nižší, tím větší počet slov má tuto frekvenci. B. Trnka[101] vytýká této formuli, že nepřihlíží k rozsahu textu, to znamená, že

[101] Trnka, B.: *G. K. Zipf: The Psycho-Biology of Language. G. K. Zipf: Human Behavior and the Principle of Least Effort.* ČMF 33, 1950, s. 3–4.

nemůže platit pro text libovolné délky, což ostatně M. Těšitelová na základě svých výzkumů rovněž potvrdila.

Zipf má zcela jistě pravdu v tom, že v distribuci slov existuje značná pravidelnost, která svědčí o tendencích v jazyce udržet rovnováhu mezi frekvencí slov a počtem slov různých, ovšem nelze to vyjádřit tak jednoduchou formulí.

2.6.3 Třetí Zipfův zákon

Říká nám, že slova s vysokou frekvencí mají zpravidla větší počet významů. Lze jej zapsat jako

$$\frac{m}{\sqrt{f}} = k.$$

Poprvé tento výsledek G. K. Zipf uvedl ve své práci *Human Behavior and the Principle of Least Effort*[102]. Matematicky se pokouší vyjádřit vztah mezi frekvencí slova a počtem jeho významů. Zipf se zde vlastně pokusil postihnout pomocí statistiky stránku sémantickou, která dosud stála stranou statistické analýzy.

I aplikace tohoto zákona na český materiál[103] ukázala, že počet významů daného slova není závislý na frekvenci slova. Např. i substantiva s frekvencí nižší a nejnižší mohou mít v podstatě týž počet významů jako substantiva s frekvencí vyšší a nejvyšší. Totéž platí u sloves. Naopak u slov formálních (např. u předložek) se ukazuje, že s klesající frekvencí ubývá předložek s relativně velkým počtem významů. Lze tedy konstatovat, že zákon platí rámcově jen pro slova formální, kterým i Zipf věnoval zvláštní pozornost.

2.6.4 Ostatní výsledky

Formálních slov se týká v podstatě další Zipfova empirická formule, ve které tvrdil, že frekvence slov je v nepřímém poměru k jejich délce, to znamená, že čím je slovo delší, tím má nižší frekvenci a naopak[104]. Je pravda, že slova nejvíce frekventovaná jsou poměrně krátká (srov. spojky, předložky, členy, zájmena aj.) a že je v jazyce tendence tato slova častěji opakovat. Viz například tab. 2.3 (Pořadí prvních 10 nejčastějších slov podle FSČ). Souvisí to s funkcí těchto slov, kterou lze nazvat jako funkci spojovací. Mezi délkou slova a jeho frekvencí nemůže ale existovat bezvýhradný vztah příčiny a následku, neboť v určitých komunikačních situacích volíme slova plnovýznamová, a to tak, abychom se jasně dorozuměli o určité věci.

Další myšlenka, kterou se zabýval G. K. Zipf, se týká výskytu slov v textech. Slova se v textu vyskytují v určitých *intervalech*. Jelikož se slova můžou opakovat se značnou volností, délku takového intervalu lze určit pouze přibližně. Nověji se touto problematikou zabýval i G. Herdan, který tvrdil, že to, jak často bude které slovo užito v kombinaci s jiným, je určeno kontextem. Přesto zůstaly tyto otázky jen naznačeny.

[102] Cambridge 1949.
[103] Těšitelová, M.: *On the Role of Nouns in Lexical Statistics*. PSML 2, 1967, s. 121–139.
[104] Zipf, G. K.: *Human Behavior and the Principle of Least Effort*. Cambridge 1949, s. 66n.

2.7 Textologie

Textologie (též *textová kritika*) je společenskovědní obor stojící na pomezí jazykovědy a literární vědy, který podrobuje text všestranné analýze.

Součástí textologie je problematika tzv. *atribuce textu*, jejímž zvláštním případem je *určování autorství*. Právě spory a polemiky kolem autorství některých známých děl daly podnět k důkladnějšímu propracování textologie a rovněž i matematických metod v ní používaných, tedy i k rozvoji matematické lingvistiky. Ze světové literatury si připomeňme například spor o autorství divadelních her Williama Shakespeara, jež byly někdy připisovány Francisi Baconovi či dokonce královně Alžbětě I.[105] U nás k těm známějším patří problematika autorství *Rukopisů* (viz kap. 2.12), dále například polemiky kolem díla K. H. Máchy, J. Nerudy či *Slezských písní* P. Bezruče. Rovněž jsou to spory o autorství 21 fejetonů, zvláštních svým ostrým útočným tónem, které vyšly v časopise *Lumír* v roce 1885 a které byly podepsány pseudonymy jako Anubis, Osiris, Trut atd. Kdo byl za pseudonymy ukryt, není známo[106]. Vůbec první náznaky použití matematických metod k atribuci textu nacházíme ovšem ještě dříve, a to během 19. století, v metodě nazývané *stylometrie*, která se začala používat pro řešení tzv. *platónské otázky*.

Dalším textologickým problémem, v kterém sehrála důležitou roli matematika (zejména teorie grafů a také teorie množin), je sestavování *stemmatu*, tj. schématu zachycujícího genealogickou návaznost textových pramenů.

V samostatných kapitolách si tyto oblasti textologie využívající s úspěchem matematiky (*stylometrie*, *určování autorství*, *stemma*) blíže představíme.

2.7.1 Stylometrie

Vůbec první užití matematiky (přesněji deskriptivní statistiky) pro atribuci textu spatřujeme v tzv. *stylometrii*. Tato metoda byla vypracována pro řešení *platónské otázky* a zajímaly ji dva okruhy problémů:

1) které *Dialogy* a *Listy* napsal skutečně Platón,

2) v jakém roce vznikly, popřípadě v jakém pořadí za sebou.

Poprvé byla metoda stylometrie formulována a užita v práci W. Dittenbergera[107] z roku 1880, ale náznaky stylometrie můžeme nalézt už u L. Campbella[108].

Metoda je založena na zjišťování frekvence výskytu formálních slov, krátkých úsloví, hiátů, dubletních synonymických slov apod., tj. takových slov,

[105] 10. 11. 2011 měl v ČR premiéru koprodukční film (Spolková republika Německo, Spojené státy americké, Velká Británie) režiséra Rolanda Emmericha *Anonymous* (u nás uváděno jako *Anonym*), který ukazuje jednu takovou možnou odpověď na otázku týkající se autorství Shakespearových her.

[106] Více viz Ivanov, M.: *Historie téměř detektivní*. Praha 1964.

[107] Dittenberger, W.: *Sprachliche Kriterien für die Chronologie der platonischen Dialoge*. Hermes 16, Berlin 1880, s. 321–345.

[108] Campbell, L.: *The Sophistes and Politicus of Plato*. Oxford 1867.

která nesouvisí s tematikou textu. Vycházelo se z Platónových textů obecně považovaných za pravé, respektive z textů časově zařazených, a postupovalo se následovně:

1) Pokud se našel jev, jehož číselný výskyt byl zhruba stejný, a sporný text této hodnoty nenabýval, považovalo se to za důkaz, že Platón autorem textu není;

2) k určení chronologie se hledal rovnoměrný vzestup či pokles užívání některých jevů.

Problém stylometrie spočíval v tom, že výsledky se měnily podle toho, který jev byl vybrán. Už např. E. Zeller[109] proto vyžaduje „obšírné závěry" o všech jevech, ne o jednotlivinách. Nedostatky stylometrie shrnul ve své práci rovněž F. Čáda[110], její první propagátor a uživatel u nás. Můžeme je popsat takto:

1) Stylometrie spočívala na rovině pouhé deskripce;

2) jazykové jevy je pro účely atribuce nutno hledat podle povahy problému, nelze je volit libovolně;

3) jevy nemohou být patrné na první pohled (a tedy napodobitelné).

I přes uvedené nedostatky se ovšem nejedná o metodu zastaralou, neboť jádrem stylometrie je pojem *individuálního stylu* a snaha po nalezení jeho kvalit z hlediska kvantitativního (stylometrie = „měření stylu"). Patrná je rovněž snaha po objektivním zkoumání problému bez subjektivních představ, čímž můžeme v stylometrii spatřovat kořeny metod atribuce, které využívají matematiky.

Z historických aplikací stylometrie se zmiňme o Ritterově práci z roku 1903[111], v níž se pokusil řešit autorství 35 recenzí z *Frankfurter Gelehrter Anzeigen* z let 1772–1773, které byly připisovány Goethovi. Soustředil se především na Goethovy jazykově stylistické dublety (např. *meistens – meist*, *nirgends – nirgend*, *vergeblich – vergebens* aj.).

Zmínku si zaslouží rovněž pozoruhodná práce N. A. Morozova[112]. Ten se zaměřil zejména na pomocná (formální) slova a určoval počet jejich výskytů v 1 000 slovech textu soudobých autorů. Údaje vynesl do grafu, který nazývá *lingvistické spektrum* (rozlišuje spektrum předložkové, spojkové, zájmenné apod.); shodnou podobu křivky považoval za důkaz autorství. Brzy se kolem

[109] Zeller, E.: *Über die Unterscheidung einer doppelten Gestalt der Ideenlehre in den platonischen Schriften*. Berlin 1887.

[110] Čáda, F.: *Datování Platónova Faidra*. Listy filologické 28, 1901, s. 173–193, 342–359, 401–439; Čáda, F. – Novotný, F.: *Příspěvek k řešení otázky o pravosti listů Platónových*. Listy filologické 33, 1906, s. 193–210, 336–337. Srov. i pozdější práce Novotného, např.: Novotný, F.: *Platonovy listy a Platon*. Brno 1926.

[111] Ritter, C.: *Die Sprachstatistika in Anwendung auf Platon und Goethe*. Neue Jahrbücher für das klassische Altertum, 1903, s. 241–261, 313–325; přetištěno jako samostatná kapitola v Ritterově knize *Neue Untersuchungen über Plato*. München 1910, s. 183–227.

[112] Morozov, N. A.: *Lingvističeskije spektry*. Izvestija ORJAS 20, kn. 4, 1915; srov. též jeho práci *Christos*. Moskva 1927.

spekter rozvinula značná diskuze, které se mimo jiné zúčastnil i A. A. Markov, zakladatel stochastických procesů a vlastně i výzkumů souvisejících s entropií jazyka. Podrobně Morozovovu práci rozebírá V. V. Vinogradov[113].

K dalším autorům, kteří se zabývali stylometrickou metodou, patří například W. Lutosawski[114], z novějších pak A. Q. Morton[115].

2.7.2 Určování autorství

Při určování autorství lze postupovat v zásadě dvěma způsoby podle toho, zda se dochoval či nedochoval originální text. Pokud je původní text k dispozici, můžeme jej dále zkoumat paleograficky, chemicky, rentgenologicky apod., jak tomu bylo v případě sporů o pravost *Rukopisů zelenohorského a královédvorského* (podrobně je této problematice věnována kap. 2.12). Pokud se ovšem původní text nezachoval[116] (existují např. pouze jeho opisy, novinové články atd.), je třeba postupovat jiným způsobem. A právě tento druhý typ sporných textů podnítil vznik speciálních matematických metod určených ke zkoumání literárních děl, které jsou založeny především na určování frekvence jazykových jevů a hledání různých kvantitativních charakteristik typických pro daného autora.

Metody řešící sporné autorství jsou rovněž dvojího druhu:

1) textový, resp. jazykový rozbor díla – zkoumáme jazyk, styl apod., a to i s využitím různých kvantitativních charakteristik a jejich porovnávání matematickými, resp. matematicko-statistickými metodami;

2) literárněhistorický rozbor – zkoumáme námět, časové umístění apod.

Zaměřme se na ty metody, které využívají poznatků matematické lingvistiky. „*Považujeme-li zkoumané dílo za určitý soubor (text) X, v kterém lze zjistit nejrůznější kvantitativní charakteristiky (ty si nakonec volí badatel podle svého záměru), a díváme-li se právě tak i na díla každého kandidáta autorství A_i, jde o to, kterému autorovi A_i přisoudíme soubor (text) X. Užijeme-li jako charakteristiky textu např. rozložení délky vět, rozhodujeme metodami matematické statistiky, který z autorů A_i má s textem X statisticky shodné rozložení.*“ (Viz [79], str. 365)

V prvním kroku je třeba vyřešit problém tzv. *porovnávací množiny* M autorů A_i. „*Musí se nalézt taková porovnávací množina M, jejímuž některému prvku (prvek = autor) chceme přisoudit X vzhledem k volbě porovnávacích charakteristik* ([79], str. 365).“ Počet kandidátů autorství (tj. počet prvků porovnávací množiny M) může být známý jednak z dodatečných informací, jednak

[113] Vinogradov, V. V.: *O jazyke chudožestvennoj literatury.* Moskva 1959.

[114] Lutosawski, W.: *The Origin and Growth of Plato's Logic.* London 1897.

[115] Morton, A. Q.: *The Authorship of Greek Prose (with Discussion).* Journal of Royal Statistical Society 1965, č. 128, část 2, s. 169–233; Morton, A. Q. – Levison, M.: *Some Indicators of Authorship of Greek Prose.* In: The Computer and Literary Style, Kent Studies in English 2, Kent 1967, s. 141–179.

[116] Tak tomu bylo v případě W. Shakespera, neboť přímo od něj se dochovalo jen několik podpisů a jeho záznam 147 řádek ze hry několika autorů.

zcela neznámý. Pokud počet kandidátů neznáme vůbec, měli bychom teoreticky za porovnávací množinu vzít populaci všech lidí daného časového období, ale vychází se obvykle z předpokladu, že autorem je někdo z literárně činných lidí daného období. Tento počet by se měl dále omezit na zpracovatelný počet osob a pak řešíme vlastně případ první, kdy je počet kandidátů autorství znám. Nejjednodušší situace nastává, jsou-li dva autoři A_1 a A_2, pak jde o alternativní rozhodnutí mezi nimi. Ale například Ellegård při řešení autorství tzv. Juniových dopisů pracoval se souborem 100 literárně činných Angličanů tohoto období. Problém nastane v případě, není-li skutečný autor prvkem množiny X, tj. jeho literární činnost se omezila jen na zkoumané dílo. Správně by se mělo totiž dojít k závěru, že dílo X není možno přisoudit žádnému z autorů. Důležité je rovněž to, aby se shodovaly závěry literárněhistorické s těmi statisticko-lingvistickými.

Aby bylo možno použít matematických, zejména kvantitativních metod, je nutné pomocí kvalitativní analýzy stanovit jev, který je v textech porovnávacího souboru pro předpokládaného autora slohově *invariantní*. Je třeba distancovat se od jevů izolovaných a nápadných (ty lze snadno napodobit) a soustředit se zejména na jevy časté, a tedy i kvantifikovatelné, a současně užívané bez vědomého záměru. Při používání matematických metod je nutné rovněž vycházet z jasně definovaných výchozích pojmů (bezesporných z hlediska vědy, ve které se matematický přístup uplatňuje) a výsledky interpretovat v rámci vědy, ve které se matematický přístup uplatňuje.

Jako pomůcka při řešení sporného autorství pak mohou badatelům posloužit rovněž různé frekvenční a konkordanční slovníky – např. slovníky směrů, období, jednotlivých autorů apod. (viz kap. 2.3.1). Velké možnosti skýtá v této oblasti využití počítačů, a to jednak při sestavování těchto slovníků či získání úplného soupisu textových variant určitého díla (využitelné při sestavování tzv. *stemmatu* – více viz kap. 2.7.3), jednak při vyhledávání a shromažďování bibliografických údajů.

U nás v hojné míře využíval matematických metod v textologii zejména P. Vašák (např. [78], [80]). To, že by mohla vzniknout jakási „matematická literární věda", však nikdy nepředpokládal, neboť vždy je součástí textologické práce kvalitativní analýza literárněvědná, popřípadě jazykovědná. A ovšem ani nejpřesnější statistické metody nemohou o autorovi rozhodnout jednoznačně.

Nyní se podívejme blíže na některé zásadní práce z oblasti určování autorství. Již roku 1887 se anglický badatel T. C. Mendenhall[117] pokoušel určit, zda autorem Shakespearových her není filozof Francis Bacon. Za charakteristiku autorského stylu si zvolil délku slov. Grafickou reprezentací distribuce délky slov je podle Mendenhalla „*characteristic curve of composition*", která dává možnost identifikovat autora, nebo alespoň některé kandidáty vyloučit. Výsledky ovšem Mendenhall hodnotil pouze grafickým porovnáním distribucí 400 000 slov z děl W. Shakespeara a 200 000 slov z Baconových spisů. Obě distribuce slov se graficky lišily – největší rozdíl se ukázal u slov délky 4 (Shakespeare je

[117] Mendenhall, T. C.: *The Characteristic Curve of Composition.* Science 9, 1887, s. 237–249; Mendenhall, T. C.: *A Mechanical Solution of a Literary Problem.* The Popular Science Monthly 60, 1901, s. 97–105.

užívá častěji než Bacon) a u slov o počtu 8 a více písmen (Shakespeare jich užívá daleko méně než Bacon). Délka slova se tedy podle Mendenhalla ukázala jako charakteristická a protože ji lze jen obtížně napodobit nebo potlačit, usoudil, že Bacon nemůže být autorem Shakespearova díla.

Vůbec první systematickou studií aplikující kvantitativní lingvistiku na problematiku sporného autorství je práce G. U. Yula[118]. K identifikaci autora knihy *De Imitatione Christi*, za nějž byl považován Tomáš Kempenský nebo Jean Gerson, zavádí Yule konstantu K (tzv. *Yulova konstanta*), která udává distribuci slov v textu. Pro Yula je tato konstanta víceméně *koncentrací slovníku* a lze ji vypočítat jako

$$K = \frac{\sum_{r=1}^{s} r^2 n_r}{N^2} - \frac{1}{N},$$

kde n_r je počet slov o frekvenci r a N je délka textu.
Yule postupoval tak, že porovnával konkrétní hodnoty této konstanty a její stejné hodnoty považoval za důkaz autorství. Yule experimentálně dokazuje, že pro homogenní materiál je tato charakteristika nezávislá na rozsahu výběru. Bohužel nedořešil, jak se tato veličina – respektive *slovníkové bohatství* – mění v rámci různých textů téhož autora (v některých případech může totiž variace vykazovat rozdíly stejné jako rozdíly mezi texty různých autorů). Stabilita je zaručena pouze pro skupiny děl pojednávající o velmi podobných tématech. Ale to podle Yula může rozhodnout až velké množství různých kontrolních výpočtů. I přes tyto nedostatky bylo ale možno na Yulovu práci navázat právě v rámci tzv. *slovníkového bohatství*[119]. Vedle toho se tento anglický matematik pokusil k atribuci využít délku věty jako vzdálenost mezi dvěma tečkami[120]. O nový způsob odvození konstanty K se pokusil G. Herdan[121] (bez Yulova předpokladu Poissonova rozložení).

[118] Yule, G. U.: *Statistical Study of Literary Vocabulary*. Cambridge 1944.

[119] Guiraud, P.: *Les caractères statistiques du vocabulaire*. Paris 1954; Kuraszkiewicz, W.: *Statystyczne badania sownictwa polskich tekstów XVI wieku*. Polskie Studie Sawistyczne. Warszawa 1958, s. 241–257; Woronczak, J.: *Metody obliczania wska»ników bogactwa sownikowego tekstów*. Poetyka i matematika, Warszawa 1965, s. 145–163; Těšitelová, M.: *On the so called Vocabulary Richness*. Prague Studies in Mathematical Linguistics 3, Praha 1972, s. 103–120; táž: *Otázky lexikální statistiky*. Praha 1974.

[120] Yule, G. U.: *On Sentence Length as a Statistical Characteristic of Style in Prose*. Biometrika 30, 1939, s. 363–390; pro formální chápání věty rovněž viz Morton, A. Q.: *The Authorship of Greek Prose (with Discussion)*. Journal of the Royal Statistical Society, 1965, No 128, part 2, s. 169–233; dále viz Shermann, L. A.: *Some Observation upon the Sentence Length*. University of Nebraska Studies 1, 1888, s. 119–130; Gerwig, G. A.: *On the Decrease of Predication and of Sentence Weight in English Prose*. University of Nebraska Studies 2, 1894, s. 17–86; Williams, C. B.: *A Note on the Statistical Analysis of Sentence Length*. Biometrika 31, 1939–40, s. 356–361 (polemika s Yulem); dále srov. práce G. A. Lesskise: *O razmerach predloženij v russkoj naučnoj i chudožestvennoj proze 60-ch godov 19 v.* Voprosy jazykoznanija, 1962, No 2, s. 78–95; týž:*O zavisimosti meždu razmerom predloženija i jego strukturoj v raznych vidach teksta*. Voprosy jazykoznanija, 1964, No 3, s. 99–123; týž: *O zavisimosti meždu razmerom predloženija i charakterom teksta*. Voprosy jazykoznanija, 1963, No 3, s. 92–112.

[121] Herdan, G.: *A New Derivation and Interpretation of Yules Characteristic K*. Journal of Applied Mathematics and Physics (ZAMP) 6, 1955, 332–334.

Trojice I. Neiescu, A. Stan a I. Stan[122] řešila autorství díla *Cintarea Rominiei* a rozhodovala se, zda je autorem Al. Russo nebo N. Balcescu. O tomto díle *Cintarea Rominiei* se v souvislosti s korelací mezi autory zmiňuje rovněž G. Herdan[123].

Podrobně se nyní zastavme u prací švédského lingvisty Alvara Ellegårda[124] z roku 1962, v nichž se pokoušel určit autora dopisů, jenž v letech 1769–1772 vycházely v londýnských novinách *Veřejný oznamovatel* (*Public Advetiser*) pod jménem Junius[125]. Hlavním cílem Ellegårda ovšem nebylo rozřešit autorství dopisů, ale vyvinout takovou statistickou metodu, která by byla obecně vhodná jako test autorství. Tyto dopisy tvoří celek poměrně jednolitého materiálu (asi 150 000 slov), proto mohou podle něj dobře posloužit. Jazyk a styl autora se Ellegård snaží postihnout pomocí kvantitativních údajů. Vychází z předpokladu, že jazyk a styl autora obsahuje rysy a jejich kombinace, které jsou konstantní nebo se mění předpověditelným způsobem. Dále předpokládá, že alespoň pro některé tyto rysy existují mezi různými autory rozdíly, které jsou větší než rozdíly mezi jednotlivými texty téhož autora. Chápe tedy styl jako „*konstantní rysy nebo kombinace rysů autorova způsobu psaní*". V jeho pojetí pak i „*špatný styl*" může dobře posloužit jako prostředek identifikace. Junius ve svých dopisech užíval tzv. antitetický styl, tedy styl založený na střetávání protikladů. Tento styl byl v Juniově době značně rozšířený, a není proto pro Junia nijak typický[126]. Protože jej lze snadno napodobit, nemůže být použit pro identifikaci autora. Ellegårdovou snahou je tedy odhalit nevědomé rysy autorova způsobu psaní. Nejprve na základě Yulových poznatků[127] použil jako podvědomé charakteristiky stylu rozložení délky vět. Variabilita délky vět v Juniových dopisech ovšem převyšovala variabilitu mezi texty různých autorů, proto musela být tato charakteristika zamítnuta. Rovněž zamítl i test zakládající se na tzv. *Yulově konstantě* K, která má podle něj malou rozlišovací sílu. Nakonec navrhl a použil tzv. *slovníkový test*. Ten vycházel ze zjištění, že některá slova, slovní obraty či slovní spojení používá daný autor častěji než autoři jiní – jsou tedy pro jeho způsob psaní jistým způsobem charakteristická, tzv. *kladná slova*

[122]Neiescu, I. – Stan, A. – Stan, I.: *Contributii statistice la studiul paternităţii Cintării Rominiei*. Cercetări de lingvistică 1963, s. 329–342. Titíž: *Noi contributii statistice la studiul paternităţii „Cintării Rominiei"*. Cercetări de lingvistică 1964, s. 311–315.

[123]Herdan, G.: *The Advanced Theory of Language as Choice and Chance*. Berlin 1966, s. 163.

[124]Ellegård, A.: *Who was Junius?* Stockholm 1962, 159 stran (Kniha je úvodem do juniovské problematiky – historické souvislosti, vznik dopisů, bibliografie juniovské literatury, nárys použité metody.); týž: *A Statistical Method for Determining Authorship. The Junius Letters, 1769–1772*. Stockholm 1962, 115 stran (Podrobně zde rozebírá statistickou metodu pro určování autorství.).

[125]První dopis vyšel 21. ledna 1769 a v následujících třech letech se objevovaly další. Autora dopisů hledala celá řada badatelů a Junius byl považován za největší záhadu světové žurnalistiky.

[126]Rovněž G. U. Yule nesouhlasí s tím, aby se na autora usuzovalo na základě charakteristických zvláštností jeho stylu, neboť ty mohou být imitovány. Žádá rozbor slovníku. Viz *Statistical Study of Literary Vocabulary*. Cambridge 1944, s. 2.

[127]Viz Yule, G. U.: *On Sentence Length as a Statistical Characteristic of Style in Prose*. Biometrika 30, 1939, s. 363–390; též Williams, C. B.: *A Note on the Statistical Analysis of Sentence Length*. Biometrika 31, 1939–40, s. 356–361.

(*plus word*). Druhou skupinu pak tvoří tzv. *slova záporná* (*minus word*). Obě skupiny se setkávají uprostřed, tzv. *neutrální slova*[128], tj. taková slova, která autor užívá zhruba tak často, jako jiní autoři. Autorova slova kladná a záporná určuje pomocí tzv. *distinktivního poměru* (*distinctiveness ratio*), který získá prostým dělením relativních frekvencí slov v textech zkoumaného autora a v literatuře téhož žánru v daném časovém období[129]. Protože se při sestavování seznamu Juniových kladných a záporných slov chtěl vyhnout totální excerpci textů, která je časově velmi náročná (ideální situace by byla, kdyby existovaly frekvenční slovníky pro jednotlivé autory, období apod.), postupoval při tvorbě seznamu jiným způsobem. Nejprve pozorně pročetl texty Juniovy a pro seznámení s jazykem doby i texty z porovnávací množiny autorů (asi 100 autorů, cca 1 milion slov), potom četl Juniovy texty znovu a zaznamenával si slova a slovní spojení, která podle něj použil Junius častěji než jeho současníci. Stejně postupoval při četbě textů z porovnávací množiny, kdy si zaznamenával slova a slovní spojení, jež byla užita častěji než u Junia a která si u něj vůbec nepamatoval. U těchto slov pak excerpcí zjistil jejich přesnou frekvenci výskytu, vypočítal distinktivní poměry a sestavil předběžný testový seznam. U některých výrazů se ukázalo, že ne všechny předpoklady získané čtením byly správné. Takto získaný testový seznam o rozsahu 458 výrazů se stal východiskem pro identifikaci. Podle Ellegårda je totiž nepravděpodobné, aby pro různé autory vznikly identické testové seznamy. Lze sice namítnout, že tento seznam nevystihuje přesně autorův slovník a že výběr slov je subjektivní, ale díky výpočtu distinktivního poměru se tato subjektivita do jisté míry anuluje, neboť nás zajímají frekvence výskytu. Aby Ellegård potvrdil hypotézu, že užívání slov zůstává během díla rozumně konstantní, rozděluje textový materiál na části o rozsahu 2 000 slov. Po zpracování na počítači a porovnání získaných frekvencí konstatuje, že konstrukci testového seznamu lze považovat za přípustnou. Potíže s výběrovými fluktuacemi[130] získaných frekvencí navrhuje vyřešit buď zvýšením rozsahu výběru, nebo určení autorství založit na skupinách slov, a ne na jednotlivých slovech. Ellegård dává přednost možnosti druhé a na základě stejných distinktivních poměrů slova seskupuje do dvou skupin, skupiny slov kladných a skupiny slov záporných. Pro obě skupiny odhaduje u Junia i ostatních autorů průměr a rozptyl, konstruuje interval spolehlivosti a porovnává, které texty ostatních autorů dosahují Juniových hodnot. Nejlepší výsledky získává u Philipa Francise, u něhož pět z jeho textů dosahuje Juniových hodnot v obou skupinách, každý text alespoň v jedné. Protože však podle Ellegårda tento test založený pouze na dvou skupinách slov (slovech kladných a záporných) není dostatečně citlivý, snaží se získat takový seznam, který by

[128] Stejné závěry i Těšitelová, M.: *K statistickému výzkumu slovní zásoby.* SaS 22, 1961, s. 171–182.

[129] Například přídavné jméno *uniform* má v Juniových textech frekvenci 0,000280, srovnávací výběr literatury z doby Juniovy má frekvenci 0,000065. Dělením těchto frekvencí, tj. 0,000280/0,000065, dostaneme hodnotu 4,31 (*distinktivní poměr*), který slovo *uniform* přiřadí mezi Juniova kladná slova.

[130] Má-li slovo frekvenci 0,0001, neznamená to ještě, že výběr o rozsahu 10 000 slov bude toto slovo obsahovat právě jednou. Teoreticky lze počet výskytů vypočítat na základě Poissonova rozdělení.

byl k individuálním Juniovým charakteristikám citlivější než k charakteristikám, které má Junius stejné s ostatními pamfletisty. Všiml si totiž, že Juniův slovník se přizpůsoboval jednak stylu *Veřejného oznamovatele*, jednak slovníku ostatních pamfletistů. Proto pořizuje nový výběr, který obsahoval 100 000 slov z politických dopisů otištěných ve *Veřejném oznamovateli* v Juniově době. Na základě tohoto výběru vytváří pomocný testový seznam. Nejprve z něj vyřazuje ty výrazy, v kterých se oba seznamy příliš lišily. Dále od tohoto seznamu oddělil tzv. *alternativy* (např. *on – upon*, *among – amongst*), tj. lingvisticky závislé jednotky. Je-li totiž slovo „*on*" klasifikováno jako kladné, slovo „*upon*" pak je s největší pravděpodobností klasifikováno jako záporné. Tím se snažil dosáhnout nezávislosti skupiny kladných a záporných slov. Konečný testový seznam tak obsahoval 220 jednotek a 122 alternativ. Po statistickém zpracování na základě stejných distinktivních poměrů dochází opět k závěru, že autorem dopisů je Philip Francis, úředník ministerstva, společensky tehdy poměrně vysoce postavená osoba (navíc z obsahu dopisů vyplývá, že šlo o osobu s dobrým přehledem o politické situaci a s informacemi z „první ruky").

C. Brinegar se ve své práci z roku 1963[131] pokoušel určit autorství 10 dopisů z novin *New Orleans Daily Crescent* z roku 1861 podepsaných pseudonymem Quintus Curtius Snodgrass. Tyto dopisy byly připisovány Marku Twainovi, čímž by dokumentovaly jeho účast v občanské válce. Jako charakteristiku autorského stylu zvolil Brinegar délku slova a všechna slova klasifikoval podle počtu písmen. Inspirací mu byly starší Mendenhallovy práce[132]. Nejprve zjistil rozložení délky slov ze zaručeně Twainových dopisů z let 1861, které porovnal s rozložením délky slov jeho dopisů z let 1872 a 1897. Zjistil, že rozložení délky slov se u Twaina během téměř 40 let nezměnilo. Konečně provedl srovnání délky slov v dopisech Curtise Snodgrasse a v dopisech M. Twaina z roku 1861. Na rozdíl od Mendenhalla se ovšem nespokojuje pouze s grafickým porovnáním distribucí, ale tyto dvě skupiny dopisů porovnává χ^2 testem a t-testem rozložení délky slov. Přínosné na této práci je zjištění, že délka slov se během let výrazně nemění a lze ji tedy považovat za charakteristiku autorského stylu (alespoň pro angličtinu). Jestliže se distribuce délek slov dvou dostatečně rozsáhlých textů liší, lze to brát jako důkaz toho, že se jedná o texty různých autorů; naproti tomu shoda distribucí podporuje hypotézu o stejném autorovi, ale nedokazuje ji. Brinegar došel k závěru, že je velmi nepravděpodobné, aby Mark Twain byl autorem 10 dopisů z roku 1861 podepsaných Q. C. Snodgrassem.

V témže roce se badatelé F. Mosteller a D. L. Wallace[133] pokusili určit autorství dvanácti tzv. *Federalistických článků* z let 1787–1788, které byly připisovány A. Hamiltonovi, J. Madisonovi či J. Jayovi. Srovnávali velice podobný

[131] Brinegar, C. S.: *Mark Twain and the Curtius Snodgrass Letters: A Statistical Test of Authorship*. Journal of American Statistical Association 58, 1963, s. 85–96.

[132] Mendenhall, T. C.: *The Characteristic Curve of Composition*. Science 9, 1887, s. 237–249; týž: *A Mechanical Solution of a Literary Problem*. The Popular Science Monthly 60, 1901, s. 97–105; viz též Williams, C. B.: *Studies in the History of Probability and Statistics*. Biometrika 49, 1956, s. 248–256.

[133] Mosteller, F. – Wallace, D. L.: *Inference in an Authorship Problem*; *A Comparative Study of Discrimination Method Applied to the Authorship of the Disputed Federalist Papers*. Journal of American Statistical Association 58, 1963, s. 275–309.

styl psaní u textů A. Hamiltona a J. Madisona. Délka věty jako charakteristika autorského stylu opět selhala, neboť se ukázala shoda mezi oběma autory (průměrná délka jejich vět byla 34,55 a 34,59 se směrodatnými odchylkami 19,2 a 20,3). Otázkou je, jak by v tomto případě dopadl Brinegarův test zaměřený na délku slov. Mosteller a Wallace k určení autorství používají *diskriminační funkci* a *Bayesův teorém*. Jako proměnné jim posloužila slova pomocná[134] (např. *upon*, *also*, *an*, *of* atd.), podobně jako tomu bylo ve stylometrii. Tato slova mají totiž vysokou frekvenci a jejich výběr nezávisí na tématu. Statistickým porovnáním usoudili, že autorem článků je J. Madison.

O slova pomocná a o délku věty se při řešení otázky sporného autorství tzv. *epištol svatého Pavla* opíral A. Q. Morton[135], který došel k závěru, že apoštolu Pavlovi je možné připsat pouze 4 ze 14 epištol.

Z Mostellerovy a Wallaceho metody vychází i Nirasawa[136] při řešení autorství spisu *Yura Monogatari*, který byl napsán v polovině 18. stol. v Japonsku.

Vidíme, že všichni autoři (Yule, Brinegar, Ellegård i Mosteller a Wallace, Morton, Nirasawa) užívají při řešení sporného autorství slovníkových testů. Yule vychází především ze substantiv (konstruuje charakteristiku K), Ellegård užívá plnovýznamových slov (vytváří testový seznam), Brinegar slov plnovýznamových i neplnovýznamových (určuje rozložení délky slov), Mosteller a Wallace a z nich vycházející Nirasawa či Morton neplnovýznamových slov.

Určitě by bylo zajímavé aplikovat některý z těchto přístupů na český materiál a rozhodnout tak otázku, co je možno z hlediska českého materiálu považovat za charakteristiku individuálního autorského stylu. Zatím se ovšem zdá, že při řešení problému sporného autorství musí badatel vypracovat do určité míry metodu novou, která je přizpůsobena povaze daného problému.

Opomineme-li práci A. Seydlera který se pomocí počtu pravděpodobnosti pokoušel dokázat nepravost *Rukopisů*, nacházíme další zmínku o využití počtu pravděpodobnosti k řešení jazykovědné povahy u Františka Wolfa[137]. Tento

[134] Jako zdroj tabelovaných anglických pomocných slov užívají práci Miller, G. A. – Newman, E. B. – Friedman, E. A.: *Length-frequency statistics for written English.* Information and Control 1, 1958, s. 370–389.

[135] *The Authorship of Greek Prose (with Discussion).* Journal of Royal Statistical Society 1965, č. 128, část 2, s. 169–233; Morton, A. Q. – Levison, M.: *Some Indicators of Authorship of Greek Prose.* In: The Computer and Literary Style, Kent Studies in English 2, Kent 1967, s. 141–179.

[136] Nirasawa, T.: *Inference in the Authorship of „Yura Monogatari".* Mathematical Linguistics 33, 1965, s. 21–27, 45–46.

[137] Wolf, F.: *Použití počtu pravděpodobnosti k identifikaci textu.* Inaugurace rektorů v Brně 1928/29, 1929/30, s. 99–105. (František Wolf se narodil 30. listopadu 1904 v Prostějově, kde vystudoval reálku. Studoval na brněnské univerzitě matematiku a fyziku (mimo jiné u profesorů B. Hostinského a E. Čecha). Poté působil v Cambridgi, v roce 1938 se habilitoval u prof. Jarníka na Přírodovědecké fakultě Karlovy univerzity. 1. prosince 1938 odjel na stipendium do Švédska (zde pracoval v Mittag-Leflerově ústavu u profesora Carlemanna). Na jaře roku 1941 se přes Finsko, Sovětský svaz a Japonsko dostal do Ameriky, kde od roku 1942 působil na University of California v Berkeley. První jeho práce se zabývaly problematikou konvergence obecných trigonometrických řad, od roku 1952 se specializoval na funkcionální analýzu (teorie perturbací) a od roku 1959 na teorii singulárních okrajových úloh v parciálních diferenciálních rovnicích. Zemřel 12. srpna 1989.)

absolvent brněnské univerzity byl prvním doktorandem prof. Čecha a promoval v roce 1928 (sub auspiciis). Jeho promoční přednáškou byla úvaha z počtu pravděpodobnosti *O identifikaci textů*.

Dále se u nás problematikou sporného autorství zabýval v 50. letech 20. století O. Králík[138]. Ten se pomocí literárněhistorických a textových výzkumů pokoušel určit autorství povídky *Kříž pod Petřínem* otištěné v almanachu *Máj* z roku 1858 a několika básní, za jejichž autora byl považován Josef Barák. Impuls k celému problému Neruda – Barák dala nepřímo francouzská badatelka Marie Scherrerová[139] článkem, v němž poukazovala na pozoruhodné shody a podobnosti mezi zapadlými verši Josefa Baráka z konce 50. let 19. století a Nerudovými *Prostými motivy*. O. Králík pak na základě svých výzkumů za pravděpodobného autora označil Jana Nerudu. Ovšem z textových paralel nelze ještě usuzovat na autorskou totožnost Baráka a Nerudy. V roce 1972 se pokoušel určit autorství povídky *Kříž pod Petřínem* znovu P. Vašák[140], a to již pomocí kvantitativních metod. Délka věty (vymezená formálně jako úsek od tečky k tečce) jako charakteristika autora se v případě Jana Nerudy stala opět neprůkazná, neboť se ukázalo, že i přes její velké rozpětí od 11,09 do 17,00 slov nezahrnuje případ *Kříže pod Petřínem* s průměrnou délkou věty 10,46 slov. Vzal tedy za jednotku zkoumání soubory vět rozlišené podle úlohy věty v textové stavbě díla a podle uplatnění primárního a sekundárního vypravěče. Ukázalo se, že hodnoty průměrné délky věty charakteristické pro povídku *Kříž pod Petřínem* jsou buď na okraji hodnot typických pro Nerudovy povídky, či leží mimo ně. Rovněž aplikace Yulovy teze, že pro autorský sloh je konstantní i rozptyl délky věty okolo stanoveného průměru, ukázala, že to platí pouze pro uvozovací věty Nerudových povídek (a i v tomto případě ležely hodnoty *Kříže pod Petřínem* mimo hodnoty typické pro Nerudu). I další kvantitativní zkoumání věty (pozice posledního podstatného jména ve větě, délka slov) neprokázala autorství Jana Nerudy, dokonce u délky slov dospěl Vašák k závěru, že získané hodnoty jsou příznačné pro jiné Barákovy texty, proto mohl říci, že autorem povídky je opravdu podepsaný J. Barák.

Statistickými metodami se v Ústavu pro jazyk český ČSAV analyzovaly například *Rukopisy královédvorský a zelenohorský*. Podrobněji bude o nich pojednáno v samostatné kapitole 2.12.

[138]Králík, O.: *Neruda a Barák*. sb. VŠP Olomouc, Jazyk a literatura 3, 1956, s. 71–93; *Svědectví Anny Holinové*. Host do domu, 1956, s. 107–109; *Neruda nebo Barák?* Literární noviny 6, 1957, č. 4, s. 6; *Z doby Májů*. Olomouc 1958; *Almanach Máj a jeho redaktor*. Slezský sborník 56, 1958, s. 355–359; *Křižovatky Nerudovy poezie*. Praha 1965; *Problém Barákova autorství*. Česká literatura 20, 1972, s. 179–206. S Králíkovými závěry polemizovali Vodička, F.: J*ještě jednou Neruda nebo Barák?* Literární noviny 7, 1958, č. 4, s. 6; Vašák, P.: *Metody ustanovlenija spornogo avtorstva (problema Barák – Neruda)*. Prague Studies in Mathematical Linguistics 3, Praha 1972, s. 143–162; *Vztah mezi délkou a funkcí věty v Nerudových Arabeskách*. Česká literatura 19, 1971, s. 272–299; *Barákovo autorství jako problém*. Česká literatura 22, 1974, s. 145–155; Macek, E.: *Biografická realita Josefa Baráka a jeho básnické texty*. Česká literatura 22, 1974, s. 156–170; *Králík kontra Barák*. sb. Literární archiv VIII – IX. Praha 1978, s. 495–582.

[139]Scherrerová, M.: *Neruda a Barák*. Slovesná věda 2, 1949, s. 112–115.

[140]Vašák, P.: *Metody ustanovlenija spornogo avtorstva (Problema Barak – Neruda)*. PSML 3, 1972, 143–162.

Podle P. Vašáka [79] je volba náležité statisticko-lingvistické charakteristiky základní otázkou při řešení sporného autorství. Druhou otázkou je pak hodnocení výsledků. Podle něj by bylo vhodné zjistit větší množství charakteristik a ty pro jednotlivé texty či autory porovnávat vícerozměrnými statistickými metodami. Některé z těchto charakteristik se mohou ukázat jako naprosto necharakteristické a texty mezi sebou nerozlišující. Vhodné se rovněž ukazuje užití např. *faktorové analýzy*, která každý údaj hodnotí vzhledem k celému souboru ostatních údajů (údaje nepodstatné „odstraňuje", údaje podstatné, tj. charakteristické, při její aplikaci vystupují „na povrch"). Rovněž se zdá, že pro sporné autorství by byly vhodné *testy neparametrické*, které nevycházejí z předpokladu jistého typu rozložení, neboť hledání statistických rozložení lingvistických jednotek je velmi obtížné a často problematické[141]. Potom se totiž stává, že statisticky signifikantní výsledek není signifikantní lingvisticky. Jako užitečné se ukazují tyto kvantitativní údaje: a) četnost (frekvence) lingvistických jednotek; b) statistické charakteristiky (průměr, rozptyl apod.); c) indexy a koeficienty (např. Yulova konstanta aj.); d) informačně-teoretické charakteristiky (viz kap. 2.8).

Rozsáhlou bibliografii tohoto oboru lze nalézt v práci [78]. Jedná se o knižní publikace, sborníky i jednotlivé články, které jsou jednak obecného charakteru, jednak se zabývající problematikou atribuce. Vynechány jsou populární články a rozsáhlá literatura týkající se podvrhů. Důraz je kladen na metodologicky přínosné práce a práce důležité pro vývoj oboru a problému. Tato bibliografie vznikla na základě poznámek z dlouholetého studia této problematiky, excerpcí pramenů z literatury.

2.7.3 Stemma

Jak bylo řečeno na začátku kapitoly, je cílem textologie rekonstrukce textového procesu. Mezi úkoly textologa pak patří vedle shromáždění množiny textových pramenů, atribuce díla, realizace textu díla i sestavení tzv. *stemmatu*. Stemmatem rozumíme schéma, které zachycuje genealogickou návaznost textových pramenů (dochovaných, nedochovaných, hypotetických) určitého textu díla. Stemma tedy zachycuje průběh textového procesu (jeho logiku) a je třeba si uvědomit, že z časového seřazení textových pramenů nevyplývá nutně paralelní textová návaznost.

Textový proces je založen na přenosu informace z pramene do pramene. Mohou nastat čtyři případy (a samozřejmě jejich libovolné kombinace):

1) formulace A je změněna na formulaci B, tj. A je přepsáno na B;

2) formulace A je vyškrtnuta, tj. A je přepsáno na 0;

3) doplnění formulace A, tj. 0 je přepsáno na A;

[141] Výsledkem takového pokusu je kniha Wimmer, Gejza – Altmann, Gabriel: *Thesaurus of Univariate Discrete Probability Distributions*. Stamm-Verlag Essen, 1999, která má 838 (+ xxviii) stran. Připomeňme, že G. Wimmer je v současné době pracovníkem ÚMS PřF MU v Brně.

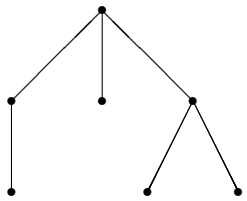
4) přesun formulace A na jiné místo, tj. Ax je přepsáno na xA.

Každý textový pramen tedy obsahuje jisté množství informace o průběhu textového procesu. Porovnáváme-li dva textové prameny mezi sebou, lze konstatovat buď informační shodu (ta nastává při dokonalé reprodukci – věrný opis, kopie, fotografická reprodukce apod.), nebo rozdíl. Množství informace zjišťujeme pomocí textových změn a jejich počtu. Pro n textových pramenů může být v konkrétním textovém místě až n různých znění, ale rovněž znění jediné.

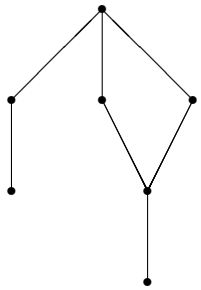
Sestavování stemmatu se skládá ze dvou fází:

1) *zřetězení pramenů* (zjišťujeme, které prameny spolu bezprostředně souvisely);
2) jejich *orientace* (zjišťujeme směr jejich textové závislosti, tj. zda z pramene A vyplývá pramen B nebo naopak).

Pokud z jednoho textového pramene vznikne 2 a více opisů, má stemma podobu rozvětvujícího se stromu:



Pokud textový pramen vznikne kontaminací dvou a více pramenů, pak se na stemmatu vytvoří uzavřené obrazce (mnohoúhelníky), např.:



Zdůrazněme ale, že kontaminaci pramenů žádná z dále uvedených metod sestavování stemmatu v čisté modelové podobě neuvažuje.

Když textolog shromáždil textové prameny k uvažovanému textu díla, má před sebou zpravidla velmi nesourodý materiál – autografy, opisy, tisky, pracovní texty, korektury apod., a to původu autorského i neautorského (často bez možnosti rozlišení), různého rozsahu, z různých časových období i fází tvorby. Jeho úkolem je dát této nesourodé množině jakousi logiku, řád – provést její typologii z hlediska textové návaznosti, tzn. zachytit strukturu vztahů množiny textových pramenů, sestrojit stemma. Protože textolog pracuje s materiálem velmi rozmanitým, je třeba vyvarovat se při sestavování stemmatu toho, že

bude sestrojeno podle určité apriorní hypotézy. Vždy je potřeba použít takovou metodu sestavování stemmatu, která způsobem nezávislým na badateli zformuluje hypotézu o stemmatu, jež je potom prověřena v rámci textové reality a historie textu.

My si zde uvedeme tři metody sestavování stemmatu využívající matematiky. Tyto metody jsou založeny na porovnávání pramenů mezi sebou a registrování odlišných míst (*metoda společných chyb*, *metoda rozkladu množiny* a *metoda taxonomická*). Poslední dvě metody budeme demonstrovat na modelovém příkladu podle P. Vašáka ([78], [80]).

Metoda společných chyb

Byla zformulována ve 2. čtvrtině 19. století Němcem K. Lachmannem. I když podle J. Frogera[142] ji již předtím užíval G. Gröber a G. Paris. Princip metody je následující: prameny se shodnými chybami mají shodný počátek, resp. textový vzor, z něhož vyšly. Problémem této metody je už tzv. *chyba*, která se určovala pomocí *konjekturální kritiky* (*konjektura* je dohad správného znění). Jednak je obtížné samotné rozpoznání chyby, jednak k tomu přistupují problémy další: co s chybami společnými všem pramenům (pak by to byl společný počátek), co s chybami vyskytujícími se pouze v jednom případě. Předpokládalo se, že společné chyby určují větev stemmatu, tj. základní linii, chyby samostatné větve rozlišují. Zcela jistě je tato metoda založena na racionálním jádru, neboť shoda v jisté informaci vypovídá o vzájemné závislosti. Metoda byla často kritizována a diskuse stále pokračují. Už v roce 1913 ji kritizoval J. Bédier a v roce 1928 konstatoval, že většina stemmat sestavených touto metodou se dělí pouze do dvou linií. Podobné námitky uvádí rovněž D. S. Lichačev[143]. Co této metodě ale nemůžeme upřít, je snaha vnést do nesourodého textového materiálu logiku, a to nezávisle na badateli, tedy formálním způsobem. A výsledné stemma je třeba vždy brát pouze jako hypotézu o stemmatu, kterou dále prověříme na konkrétním textovém materiálu.

Podívejme se ale blíže na kritizovaný fakt, že většina stemmat sestavených touto metodou byla bifidní („*two branch stemma*"), tj. dělila se do dvou linií. Protiklad „*chyba – správné znění*" aplikovaný na každý textový pramen vcelku vede nevyhnutelně k bifidnímu stemmatu. Každý textový pramen se ale skládá z množiny takových míst „*ano-ne*". V metodě společných chyb tedy chybí informace o tom, kolik takových „*ano-ne*" míst je a v jaké kombinatorice jsou k ostatním pramenům. Pokud se tedy opozice „*chyba – správné znění*" aplikuje na jednotlivá textová místa, ještě z toho nevyplývá nutnost bifidního stemmatu.

V roce 1920 metodu společných chyb modifikoval Dom Quentin, když pojem *chyba* nahradil metodologicky správnějším pojmem *různé znění*, resp. *varianta*. Jeho postup však nakonec stejně vedl k modifikaci *metody společných chyb*. Jeho největším přínosem ale je rozlišování dvou fází stemmatu, a to *zřetězení pramenů* a *orientace pramenů*. Tyto fáze bezprostředně vyplývají z matema-

[142] Froger, J.: *La critique des textes et son automatisation*. Paris 1968.
[143] Lichačev, D. S.: *Tekstologija*. 2. vydání, Leningrad 1983.

tické teorie grafů, kde se rozlišují grafy neorientované (spojení prvků) a orientované (směr závislosti).

Následující dvě metody sestavování stemmatu, které si představíme, pracují nejen s textovými změnami, ale i jejich počtem a s kombinacemi textových změn mezi prameny. Nejsou to samozřejmě metody jediné – při sestavování stemmatu se využívají například postupy založené na složitějších matematických technikách, jako je faktorová analýza (jejíž variantou je ostatně vratislavská taxonomie), multivariační analýza aj.

Taxonomická metoda

Vyšla z myšlenek polského antropologa Jana Czekanowského, který pro typologické srovnání lebek různých ras navrhl metodu dnes známou jako *vratislavská taxonomie*. Každý prvek množiny, jejíž typologii hledáme, se popíše soustavou nějakých charakteristik (kvantitativních, resp. kvantifikovatelných), na jejichž základě se sestaví tabulka udávající vzdálenost každého prvku od všech ostatních. Ke každému prvku nalezneme na základě vzdálenosti prvek nejbližší, tedy „nejpříbuznější". Tento fakt lze znázornit do podoby grafu – v terminologii taxonomie se mluví o *dendritu*. Převedeme-li to do řeči textologie, pak textový pramen, který vznikl jako opis nějakého vzoru, je tomuto vzoru bližší než pramen vzniklý jako opis opisu atd. Lze totiž předpokládat, že každým dalším opisováním vzrůstá počet změn vůči prvotnímu vzoru. Více si ukážeme na modelovém příkladu.

Metoda rozkladu množiny (resp. metoda skupinová)

Byla vypracována francouzským textologem J. Frogerem (v jeho terminologii je to metoda „*par les groupes*") přímo pro potřeby textologie. Je založena na poznatku, že každá textová změna vytváří v množině dochovaných textových pramenů určité typologické skupiny podle toho, jaká konkrétní textová znění prameny mají. Čím více je míst textových změn, tím více máme dokladů o rozkladu množiny textových pramenů. Máme-li například pět textových pramenů A, B, C, D, E, z nichž mají na určitém textovém místě shodné znění a prameny A a B, znění b pramen C a znění c prameny D a E, pak tato tři různá znění a, b, c nám vypovídají o typologických skupinách AB, C a DE, tzn. o jakési možnosti společné textové historie pramenů AB, jiné historie u pramene C a jiné u DE. Proto P. Vašák oproti původnímu Frogerovu názvu „*par les groupes*" zavádí označení *rozklad množiny* textových pramenů, který je podle něj výstižnější.

Při porovnávání textů může být takových rozkladů větší počet. Postupujeme tedy následujícím způsobem. Průběžně očíslujeme pořadovými čísly 1, 2, 3... všechna textová místa, v kterých se textové prameny odlišují. Potom si u každého z těchto textových míst zaznamenáváme skupiny textů, které mají jisté shodné znění a, b, c atd. Teoreticky se může v textovém místě u n pramenů vyskytovat n různých znění.

P. Vašák sestavoval stemma na základě textových pramenů dochovaných k Máchovým *Cikánům*. Celkem se dochovalo 7 různých textových pramenů, z toho pro první kapitolu pouze 5 (L – tisk v časopise *Lumír* z roku 1851, $O6$ – Schulzův opis z roku 1851, C – knižní vydání u Jeřábkové 1857, K – tisk u Kobra z roku 1861, $R21$ – Máchův autograf obsahující první dva odstavce první kapitoly). Při porovnávání těchto pěti pramenů bylo nalezeno celkem 57 textových míst, v nichž se prameny odlišovaly. Zjištěné údaje jsou uvedeny v tab. 2.8.

znění a	znění b	znění c	počet míst
$R21$, L, C, K	$O6$	–	10
$R21$	$O6$, L, C, K	–	30
$R21$, $O6$	L, C, K	–	11
$R21$, $O6$, L, C	K	–	1
$R21$	$O6$	L, C, K	5

Tabulka 2.8: Rozklad textových pramenů k první kapitole Máchových *Cikánů* do typologických skupin

Z tabulky je vidět, že na 52 místech nacházíme 2 různá znění, na 5 místech nacházíme 3 různá znění. Rovněž je vidět, že na základě dochované textové reality lze provést pět různých rozkladů množiny textových pramenů. Těmito rozklady (tj. typologickými skupinami) jsou:

1) $R21$, L, C, K proti $O6$ (10 případů),

2) $R21$ proti $O6$, L, C, K (30 případů),

3) $R21$, $O6$ proti L, C, K (11 případů),

4) $R21$, $O6$, L, C proti K (1 případ),

5) $R21$ proti $O6$ proti L, C, K (5 případů).

Modelový příklad

Na příkladu, který uvádí P. Vašák ([78], [80]) si nyní ukážeme postup při sestavování stemmatu pomocí posledních dvou metod založených na počtu textových změn a jejich kombinatorice (*metoda rozkladu množiny*, *metoda taxonomická*). Budeme vycházet z originálního textu N (pět úvodních vět *Babičky* Boženy Němcové), který byl přeměněn, jako by byl historicky opisován (reprodukován): opisy A, B, C, D, E, F, G, H (viz tab. 2.9).

	1	2		3	4	5	6	7
N	Babička	měla	syna a dvě	dcery.	Nejstarší	žila	mnoho	let
A	Babička	měla	syna a dvě	dcery.	Nejstarší	žila	mnoho	let
B	Babička	měla	syna a dvě	dcery.	První	žila	mnoho	let
C	Moje babička	měla	syna a dvě	dcery.	Nejstarší	žila	mnoho	roků
D	Babička	měla	syna a dvě	dcery.	Nejstarší	žila	několik	let
E	Babička	měla	syna a dvě	dcery.	Nejstarší	žila	mnoho	let
F	Moje babička	měla	jednoho syna a dvě	dcerky.	Nejstarší	žila	mnoho	roků
G	Babička	měla	syna a dvě	dcery.	Nejstarší	byla	několik	let

	8	9	10	11	12	13	14
N	ve Vídni u	přátel, od	nichž	se	vdala.	Druhá	dcera
A	ve Vídni u	známých, od	nichž	se	vdala.	Druhá	dcera
B	ve Vídni u	přátel, od	kterých	se později	vdala.	Druhá	dcera
C	ve Vídni u	přátel, od	nichž	se	vdala.	Druhá	
D	ve Vídni u	známých, od	nichž	se	vdala.	Druhá	dcera
E	ve Vídni u	známých, od	nichž	se	odstěhovala.	Mladší	dcera
F	v Polné u	přátel, od	nichž	se	vdala.	Druhá	
G	ve Vídni u	známých, od	nichž	se	vdala.	Druhá	dcera

	15	16	17	18	19	20	21	22
N	šla	pak	na	její	místo. Syn,	řemeslník,	též byl	samostatným
A	odešla	pak	na	její	místo. Syn,	truhlář,	též byl	samostatným
B	šla	pak	na	její	místo. Syn,	řemeslník,	též byl	samostatným
C	šla	pak	na	její	místo. Chlapec,	řemeslník,	též byl	samostatným
D	odešla	potom	na	její	místo. Syn,	truhlář,	též byl	samostatným
E	odešla	pak	na	její	místo. Syn,	truhlář,	byl též	samostatným
F	šla	pak	převzít	její	místo. Chlapec,	řemeslník,	též byl	samostatným
G	odešla	potom	na	toto	místo. Syn,	truhlář,	též byl	samostatný

	23	24	25	26	27	28
N	a přiženil se do	městského	domku.	Babička	bydlela v	pohorské
A	a přiženil se do	městského	domku.	Babička	bydlela v	pohorské
B	a přiženil se do	městského	domku.	V té době babička	bydlela v	pohorské
C	a oženil se do	městského	domku.	Babička	zůstala v	pohorské
D	a přiženil se do	městského	domu.	Babička	bydlela v	pohorské
E	a přiženil se do	městského	domku.	Babička	bydlela v	pohorské
F	a oženil se do	městského	domku.	Babička	zůstala v	pohorské
G	a přiženil se do	malého	domu.	Babička	bydlela v	malé

	29	30	31		32	33		34
N	vesničce	na	slezských	hranicích;		žila	spokojeně v malé	chaloupce
A	vesnici	na	slezských	hranicích;		žila	spokojeně v malé	chaloupce
B	vesničce	na	slezských	hranicích;		žila	spokojeně v malé	chaloupce
C	vesničce	na	slezských	hranicích;	vím že	žila	spokojeně v malé	chaloupce
D	vesnici	na		hranicích;		žila	spokojeně v malé	chaloupce
E	vesnici	na	slezských	hranicích;		žila	spokojeně v malé	chaloupce
F	vesničce	při	slezských	hranicích;	vím že	žila	spokojeně v malé	chalupě
G	vesnici	na		hranicích;		žila	klidně v malé	chaloupce

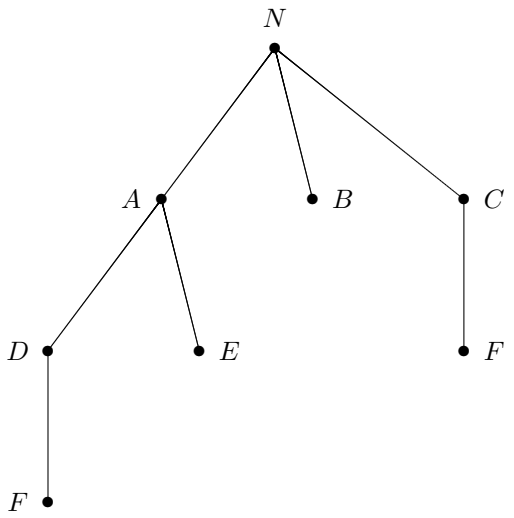
	35		36	37		38		39
N	se	starou	Bětkou,	která	byla	její vrstevnice	a	již u rodičů sloužila.
A	se	starou	Bětkou,	která	byla	její vrstevnice	a	již u rodičů sloužila.
B	se	starou	Bětkou.	Ta	byla	její vrstevnice	a	již u rodičů sloužila.
C	se	starou	Bětkou,	která	byla	její vrstevnice	a	již u rodičů sloužila.
D	se	starou	Bětkou,	která	byla	její vrstevnice	a	sloužila již u rodičů.
E	se	službou	Bětkou,	která	byla	stejně stará	a	již u rodičů sloužila.
F	se	starou	Bětkou,	která	byla	její vrstevnice	a	již u rodičů sloužila.
G	se	starou	Bětou,	která	byla	její vrstevnice	a	sloužila již u rodičů.

Tabulka 2.9: Modelové texty pro porovnávání textových pramenů – pět úvodních vět *Babičky* B. Němcové (podle P. Vašáka)

Budeme vycházet z jistých modelových předpokladů:

1) neuvažujeme kontaminaci pramenů;
2) reproduktor textu zachovává změny svých předchůdců a přidává změny nové;
3) různí reproduktoři mění text na různých místech.

Obrazem tohoto textového procesu reprodukce je následující stemma:



a) Metoda rozkladu množiny

Zvolíme si jeden z textových pramenů jako *referenční* – ostatní prameny s ním budeme porovnávat a zaznamenávat textové změny. My zvolíme jako referenční pramen A. Tento referenční pramen lze volit zcela libovolně, vhodné je však zvolit takový pramen, který je textově nejúplnější, neboť vůči němu se lépe zaznamenávají textové změny nebo mezery pramenů ostatních. Po porovnání ostatních pramenů s referenčním pramenem A dostáváme následující tabulku (viz tab. 2.10):

Číslo místa změny	Prameny mající shodně jiné znění než referenční text A	Číslo místa změny	Prameny mající shodně jiné znění než referenční text A
1	$C\ F$	21	E
2	F	22	G
3	F	23	$C\ F$
4	B	24	G
5	G	25	$D\ G$
6	$D\ G$	26	B
7	$C\ F$	27	$C\ F$
8	F	28	G
9	$N\ B\ C\ F$	29	$N\ B\ C\ F$
10	B	30	F
11	B	31	$D\ G$
12	E	32	$C\ F$
13	E	33	G
14	$C\ F$	34	F
15	$N\ B\ C\ F$	35	E
16	$D\ G$	36	G
17	F	37	B
18	G	38	E
19	$C\ F$	39	$D\ G$
20	$N\ B\ C\ F$		

Tabulka 2.10: Porovnání textových pramenů vůči referenčnímu prameni A

Údaje shrneme do tabulky 2.11, která zachycuje příslušné rozklady množiny textů. V každém řádku jsou vždy uvedeny texty, které mají shodně odlišné znění, než má referenční text A, s čísly míst změn a jejich počtem. Zbylé texty se pak s referenčním textem A shodují a tabulka je nezachycuje.

Rozklad množiny	Čísla míst změn	Počet míst (frekvence)
N B C F	9, 15, 20, 24	4
C F	1, 7, 14, 19, 23, 27, 32	7
D G	6, 25, 31, 39, 16	5
F	2, 3, 8, 17, 30, 34	6
B	4, 10, 11, 26, 37	5
G	5, 18, 22, 24, 28, 33, 36	7
E	12, 13, 21, 35, 38	5

Tabulka 2.11: Rozklad množiny textových pramenů pěti úvodních vět *Babičky* B. Němcové (podle P. Vašáka)

Našli jsme tedy sedm rozkladů množiny všech textových pramenů – jeden je čtyřprvkový ($NBCF$), dva jsou dvouprvkové (CF a DG) a čtyři jednoprvkové (F, B, G, E). Nyní je naším úkolem sestavit stemma, a to ve dvou fázích: jednak provést zřetězení pramenů, jednak jejich orientaci. Při hledání zřetězení pramenů není důležité, od kterého pramene začneme. My zvolíme za výchozí pramen F, neboť je obsažen v nejvíce rozkladech původní množiny všech pramenů ($NBCF$, CF, F). Pramen F obsahuje tedy změny společné s NBC, dále s C a změny samostatné. Proces přenosu změn do textu F tak probíhá v řetězci $NABCDEFG - NBCF - CF - F$. Vzhledem k výchozím předpokladům, že reproduktor zachovává změny svých předchůdců a přidává změny nové, můžeme textový pramen F v každé vyšší skupině eliminovat (graficky škrtnout, viz dále), protože takto probíhá přenos změn do pramene F. Graficky znázorněn vypadá přenos takto (podle Frogera):

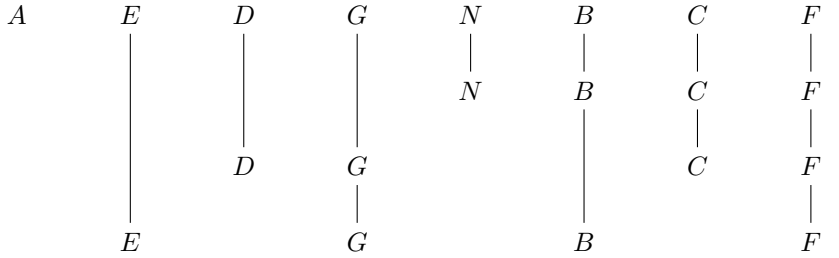
N	*A*	*B*	*C*	*D*	*E*	~~*F*~~	*G*
N		*B*	*C*			~~*F*~~	
			C			~~*F*~~	
						F	

Pro přehlednost můžeme upravit pořadí textů ve skupině:

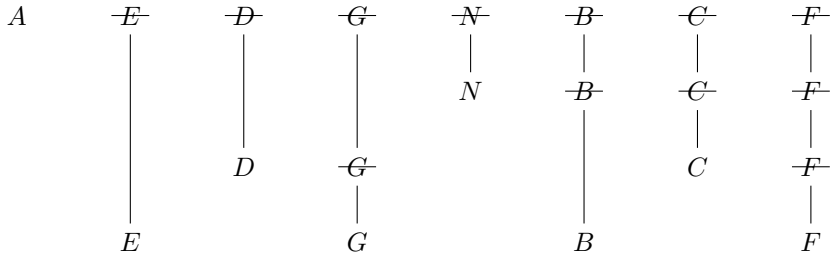
A	*D*	*E*	*G*	*N*	*B*	*C*	~~*F*~~
				N	*B*	*C*	~~*F*~~
						C	~~*F*~~
							F

Takto postupujeme při zhodnocení celého materiálu. Nalezené rozklady množiny (zde konkrétně $NBCF$, CF, DG, F, B, G, E) napíšeme pod sebe (od celé množiny pramenů přes množinu o jeden prvek menší atd., až po množiny jednoprvkové) a postupně zaškrtáváme prameny ve větších skupinách v souladu s přenosem změn z pramene do pramene. V tomto případě by přenos změn

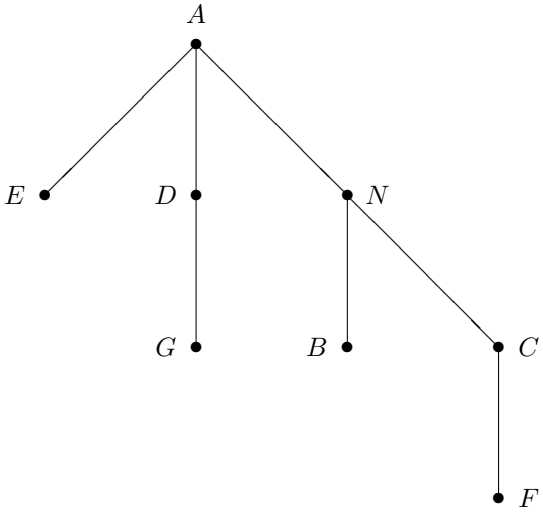
z pramene do pramene probíhal takto (spojnice mezi textovými prameny naznačují přenos změn):



Po eliminaci textových pramenů dostane schéma podobu:



Ponecháme-li pouze nezaškrtnuté texty jako konečný prvek přenosu změn od skupiny ke skupině, dostaneme následující neorientované stemma:



Toto stemma odpovídá stemmatu, které mělo vyjít a které jsme neznali. Souhlasí zřetězení pramenů, nesouhlasí pouze směr vztahu. Tímto způsobem totiž nejsme schopni určit výchozí prvek celého procesu přenosu. Získané neorientované schéma bereme pouze jako hypotézu o stemmatu, kterou je třeba dále prověřit. Vracíme se tedy do původního textového materiálu a posuzujeme jazykové, stylistické, motivické aj. proměny textového procesu podle navrženého

stemmatu, s využitím informací literárních, historických, politických, estetických apod. Hledáním logiky textového procesu se hledá výchozí textový pramen a neorientované schéma se mění na orientované, provádí se zřetězení pramenů.

V konkrétní textové praxi se můžeme poměrně často setkat s nesplněním některých modelových předpokladů (reproduktor nezachová již provedenou změnu, různí reproduktoři mění text na stejných místech apod.). Pak nalezneme například rozklad složený z pramenů AB a současně i BC. Pramen B tedy figuruje ve dvou rozkladech. V těchto situacích je vhodné držet se výchozích představ, tj. abstraktního systému, a na závěr v rámci navrhované hypotézy o stemmatu hledat racionální kritéria k odstranění těchto anomálií. Například při analýze první kapitoly pramenů k Máchovým *Cikánům* byl nalezen rozklad KC (63 textových míst) a současně i rozklad CL (4 textová místa). Porovnáním frekvencí P. Vašák dospěl k závěru, že rozklad KC je pravděpodobnější, proto rozklad CL v první fázi eliminoval. Po vytvoření hypotézy o stemmatu je ovšem nutné vysvětlit, proč k takové textové situaci došlo.

Další možnou anomálií je eliminace skupiny. Pak je jasné, že v tomto místě stemmatu je logické předpokládat existenci neznámého textového pramene. Při prověřování hypotézy o stemmatu musí být samozřejmě zdůvodněn i tento hypotetický pramen. Ztracené prameny se s jistou mírou pravděpodobnosti dají předpokládat v těch místech, kde se linie stemmatu rozvětvují. Jinak na jedné linii (spojnici) dvou pramenů lze teoreticky umístit libovolné množství hypotetických pramenů.

Nutnou a postačující podmínkou platnosti modelu je existence pouze dvou různých znění v daném textovém místě. Pokud je znění více, pak předpoklady nebyly splněny. Jako nejvýhodnější se ukázal postup, který taková místa prozatím neuvažuje, a stemma se sestaví na základě míst o dvou zněních. V další fázi se opět musí doložit, proč místa o více zněních vznikla.

b) Taxonomická metoda

Tato metoda je založena na pojmu *vzdálenost*. Jisté narušení modelových předpokladů o přenosu textu z pramene do pramene není při této metodě tolik důležité, protože nemění strukturu vztahů mezi prvky. Pojem *vzdálenosti* umožňuje pravděpodobnostní pohled: co má vyšší frekvenci výskytu, je reálnější.

Nechť se uvažovaná množina M textových pramenů skládá z prvků M_1, M_2,...,M_k. Vzdálenost textového pramene M_m od textového pramene M_n, tj.

$$d(M_m, M_n),$$

je dána počtem textových změn, které mezi nimi existují. Je zřejmé, že vzdálenost textových pramenů je symetrická (tzn. že vzdálenost M_m od M_n je stejná jako vzdálenost M_n od M_m). Dále platí, že vzdálenost pramene od sebe samého je nulová.

Postupným porovnáváním textových pramenů získáme pro každý pramen M_i, $i = 1, \ldots, k$, množinu vzdáleností od všech pramenů ostatních, z nichž vždy nalezneme vzdálenost nejmenší, která odpovídá pramenu textově nejbližšímu. Máme-li např. prameny M_a, M_b, M_c, přičemž M_b je reprodukcí M_a, M_c

je reprodukcí M_b, potom je vzhledem k předpokladům o procesu reprodukce zřejmé, že

$$d(M_a, M_b) \leq d(M_a, M_c),$$

tj. opis opisu má (vzhledem k principu pravděpodobnosti) více textových změn než původní opis.

Na základě minimálních vzdáleností sestrojíme strom (stemma), který vyjadřuje strukturu textové příbuznosti pramenů. Ukažme si postup na našem hypotetickém příkladu s osmi prameny z *Babičky* Boženy Němcové. Porovnáme-li každý pramen s každým, získáme počet textových změn mezi libovolnými dvěma prameny, jak ukazuje následující tabulka 2.12:

	N	A	B	C	D	E	F	G	součet
N	0	4	5	7	9	9	13	16	63
A	4	0	9	11	5	5	17	12	63
B	5	9	0	12	14	14	18	21	93
C	7	11	12	0	16	16	6	23	91
D	9	5	14	16	0	10	22	7	83
E	9	5	14	16	10	0	22	17	93
F	13	17	18	6	22	22	0	29	127
G	16	12	21	33	7	17	29	0	135

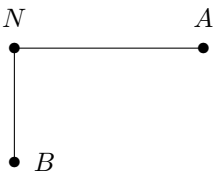
Tabulka 2.12: Vzdálenosti textových pramenů prvních pěti vět *Babičky* Boženy Němcové

Nyní sestrojíme stemma. Pro každý textový pramen nalezneme v řádku (resp. sloupci) tabulky nejmenší vzdálenost, tj. pramen nejbližší, a tyto prameny v grafickém znázornění spojíme čarou. Tak postupujeme u všech pramenů, tedy napojujeme je na prameny již spojené.

Konkrétně v našem příkladu je pramenu N nejbližší pramen A (vzdálenost 4), z druhého řádku plyne, že pramenu A je nejbližší pramen N (vzdálenost 4), což znázorníme jako



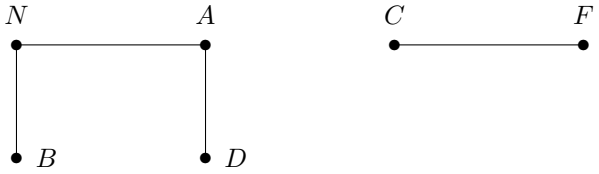
Ve třetím řádku zjišťujeme, že textovému pramenu B je nejbližší pramen N (vzdálenost 5), ke kterému byl již nejbližší pramen nalezen (pramen A), tj. graficky



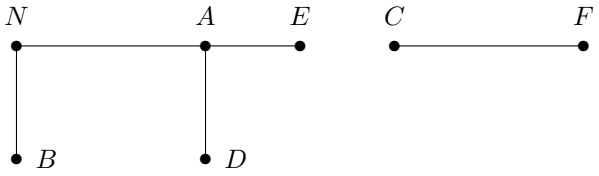
Pramenu C je nejbližší pramen F (vzdálenost 6), tj.



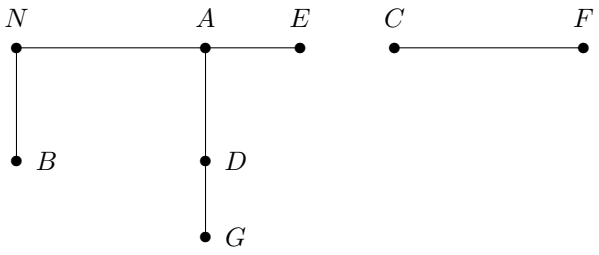
Pramenu D je nejbližší pramen A (vzdálenost 5), kterému nejbližší byl již nalezen, tj.



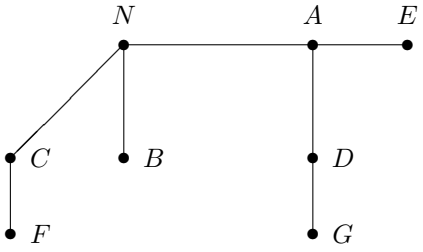
Pramenu E je nejbližší pramen A (byl již nalezen) se vzdáleností 5, tj.



Pramenu F nejbližší pramen C se vzdáleností 6 byl již nalezen a pramenu G je nejbližší D (vzdálenost 7), tj.



Nyní musíme tyto dva vzniklé podstromy spojit, tj. nalézt nejkratší spojnici mezi skupinou textů $NBAEDG$ a CF. Z tabulky 2.12 zjistíme, že nejmenší vzdálenost je pro dvojici C a N, tj. dostáváme výsledné stemma:



Tento výsledek souhlasí s kontrolním stemmatem. Podobně jako u předchozí metody jsme získali neorientované stemma, u kterého je třeba provést zřetězení pramenů. Taxonomická metoda má ale tu výhodu, že zde existuje možnost, jak omezit počet pramenů, v nichž lze nalézt počátek textového procesu. Vzhledem k předpokladu o narůstání počtu změn je zjevné, že počátkem textového procesu je ten pramen, jehož součet vzdáleností od všech ostatních je nejmenší. Není však pramenem jediným, neboť teoreticky existují dva, kdy druhým je pramen, který má od skutečného počátku nejmenší vzdálenost. V tomto případě existují dva možné počátky textového procesu, a to prameny N a A se vzdáleností 63. Takto získané stemma je opět chápáno jako jakási hypotéza o stemmatu, kterou je třeba podložit konkrétní následnou analýzou.

O vzdálenosti d mezi dvěma prameny můžeme říct, že:

1) čím je vzdálenost d relativně menší, tím menší byl stupeň textové změny vůči reprodukovanému vzoru;

2) je-li vzdálenost d relativně větší, je možné dvojí vysvětlení:
   a) reprodukovaný vzor byl výrazně měněn buď aktivním motivovaným zásahem, nebo výraznou nepozorností,
   b) mezi uvažovanými dvěma textovými prameny historicky existovaly prameny, které se nedochovaly.

Pro množinu M skládající se z k textových pramenů má vytvořené stemma celkem $(k-1)$ spojovacích linií. Odstraníme-li spojnici nejdelší, rozpadne se stemma na dvě části (dva podstromy). V této části je stemma nejvolnější a zde má textolog právo logicky předpokládat existenci hypotetického pramene. Jeho existence však musí být textově možná a musí být dokázána další analýzou.

Každá seriózní metoda musí vycházet z určitých modelových předpokladů (stejně jako dvě předložené metody). Tyto předpoklady mohou mít na konkrétním materiálu pravděpodobnostní průběh, tj. stačí, aby platily ve většině případů. Rovněž, s oporou v teorii systémů, se z praktických důvodů doporučuje sestavovat stemma pouze na základě míst o dvou různých zněních. Textová místa o třech a více zněních slouží k ověření takto vytvořené hypotézy o stemmatu.

2.8 Teorie informace

Je matematická disciplína zabývající se přenosem, kódováním a měřením *informace*. Vznikla v souvislosti s rozvojem kybernetiky a její počátky klademe na přelom čtyřicátých a padesátých let 20. století. Za zakladatele tohoto oboru jsou považováni anglický matematik a inženýr Claude Elwood Shannon a americký matematik a fyzik Warren Weawer, kteří vyložili základy *teorie informace* v roce 1949 ve své práci *Matematická teorie komunikace*[144]. Protože byla tato

[144]Shannon, C. E. – Weawer, W.: *The Mathematical Theory of Communication.* Urbana 1949. Viz též Shannon, C. E.: *A mathematical theory of communication.* Bell System Technical Journal, vol. 27, 1948, s. 379–423, 623–656; Shannon, C. E.: *Prediction and Entropy of Printed English.* Bell Systém Technical Journal 30, 1951, s. 50–64 (čes. překlad ve sborníku *Teorie informace a jazykověda*, Praha 1964, s. 75–88.

práce určená především matematikům a vědcům z ostatních oborů byla těžko srozumitelná, velký význam měla její rozsáhlá recenze od Ch. F. Hocketta[145], který *teorii informace* (a rovněž jí blízkou *teorii komunikace*[146] přiblížil lingvistům. Podnítil tak jejich zájem o spolupráci při řešení otázek týkajících se přirozených jazyků.

Od 50. let 20. století se začala teorie informace významnou měrou uplatňovat i v nově vznikající kvantitativní lingvistice[147]. V centru pozornosti kvantitativní lingvistiky stály zejména její dva pojmy – *entropie* a *redundance*[148], které si blíže představíme společně s pojmy *bit* a *šum*.

Jedním ze zásadních výsledků teorie informace je zjištění, že množství informace se dá měřit. K jejímu měření byl převzat z fyziky (přesněji z termodynamiky) termín *entropie*, který můžeme definovat jako průměrné množství informace obsažené v jednom výsledku příslušného pokusu. Rovněž lze entropii definovat jako míru neurčitosti pokusu. Shannon pro výpočet entropie H[149] zavedl vzorec

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i,$$

kde N ... počet prvků v množině, p_i ... pravděpodobnost výskytu i-tého prvku pro $i = 1, 2, \ldots, N$. Současně musí platit, že

$$\sum_{i=1}^{N} p_i = 1,$$

kde $p_i \geq 0$. Entropie má tyto vlastnosti:

1. Entropie je maximální, jestliže jsou všechny prvky stejně pravděpodobné.

2. Entropie je nulová, jestliže je pravděpodobnost jednoho z prvků 1 (a tedy ostatních prvků 0).

[145] Hockett, Ch. F.: *Review of C. E. Shannon and W. Weaver The Mathematical Theory of Communication*. Language 29, s. 69–93.)

[146] Teorie komunikace se zabývá formální stránkou přenosu informace. Schéma komunikačního procesu je následující: *informace zakódovaná* podle pravidel daného *kódu* přechází ve formě *signálů kanálem* od *zdroje* směrem k *příjemci*, kde je *dekódována*. Přirozené jazyky jsou jen jedním, i když nejdůležitějším, z mnoha komunikačních systémů. Naproti tomu teorie informace se zabývá samotnou *informací*, má proto pro jazykovědu větší význam.

[147] Aktuálnost řešení lingvistických otázek pomocí teorie informace dokumentuje i ten fakt, že v roce 1957 byl v USA založen nový vědecký časopis *Information and Control*, v jehož redakci vedle zakladatele teorie informace C. E. Shannona a zakladatele kybernetiky Norberta Wienera, zasedl známý ruský lingvista Roman Jakobson, který se mimo jiné podílel na založení Pražského ligvistického kroužku a jistý čas působil i na univerzitě v Brně.

[148] Viz Novák, P.: *Teorie informace a lingvistika*. In: [58], s. 115–125; též Řeháček, L.: *Populární výklad základů moderní matematické a strojové lingvistiky*. Slovo a slovesnost 27, 1966, s. 147–151; srov. též Herdan, G.: *The Advanced Theory of Language as Choice and Chance*. Berlin – Heidelberg – New York 1966, s. 259n.

[149] Shannon mluví o tzv. *selektivní informaci*. Více viz [49].

3. Entropie je aditivní, tj. má-li nějaké konečné schéma[150] A entropii H^A a konečné schéma B entropii H^B, pak entropie složeného systému AB (při nezávislosti obou) je rovna

$$H^A + H^B.$$

Zjednodušeně řečeno je entropie tím větší, čím je výsledek pokusu (jev) méně předvídatelný. Převedeme-li vše na lingvistickou problematiku, pak *předvídatelnost* (*predictability*) označuje míru pravděpodobnosti, s jakou je posluchač schopen na základě dosud poznané části výpovědi předem odhadnout její další část (viz též kap. 2.5). Množství informace je největší tehdy, když nejsme schopni vůbec předvídat další část výpovědi, což nastane v případě, že jsou všechny prvky stejně pravděpodobné. Nulové množství informace získáme tehdy, když následující prvek uhodneme s jistotou (takový prvek nazýváme *redundantní*, tj. nadbytečný, neboť nám neposkytuje žádnou informaci). Uvědomme si ale, že míra informace užívaná v teorii informace nemá nic společného se sémantickým obsahem přenášených sdělení a zabývá se výhradně statistickou strukturou formálního zobrazení. To vylučuje možnost aplikací teorie informace na studium sémantických problémů.

Příklad 2.1: Mějme dva prvky A_1 a A_2, které mají stejnou pravděpodobnost výskytu, tj.

$$\begin{pmatrix} A_1 & A_2 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Dále mějme prvky B_1 a B_2, jejichž pravděpodobnosti výskytu odpovídají schématu

$$\begin{pmatrix} B_1 & B_2 \\ \frac{1}{100} & \frac{99}{100} \end{pmatrix}.$$

Vidíme, že v druhém případě lze předpovědět výsledek pokusu snadno, neboť z každých 100 pokusů nastane jev B_1 pouze jedenkrát a jev B_2 ve všech ostatních případech. V případě prvním mohou v následujícím pokusu nastat oba jevy A_1 a A_2 se stejnou pravděpodobností, raději se proto o výsledku následujícího pokusu nevyslovíme. Entropie (neurčitost) prvního schématu je tedy zcela jistě větší než entropie schématu druhého. Entropie je v prvním případě rovna

[150] Konečným schématem A rozumíme množinu vzájemně neslučitelných jevů A_i s pravděpodobnostmi výskytu těchto jevů $p(A_i)$, kde $i = 1, 2, \ldots, N$, z nichž při každém provedení pokusu nastane právě jeden jev. Schéma lze znázornit takto:

$$A = \begin{pmatrix} A_1 & A_2 & \ldots & A_N \\ p(A_1) & p(A_2) & \ldots & p(A_N) \end{pmatrix}.$$

$$H = -\sum_{i=1}^{2} p_i \log_2 p_i = -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} = 1,$$

v druhém případě je entropie rovna

$$H = -\sum_{i=1}^{2} p_i \log_2 p_i = -\left(\frac{1}{100} \log_2 \frac{1}{100} + \frac{99}{100} \log_2 \frac{99}{100} \right) = 0{,}15.$$

Příklad 2.2: V tabulce 2.7 je uveden příklad s nápodobou českého, ruského, anglického a německého textu podle teorie pravděpodobnosti za předpokladu, že všechna písmena mají stejnou frekvenci, dále s přihlédnutím k relativní frekvenci jednotlivých písmen, relativní frekvenci dvojic písmen a relativní frekvenci trojic písmen. Na příkladu češtiny si ukážeme, jak můžeme při nápodobě textu postupovat. Symbolem H_0 označíme entropii v prvním případě, kdy se vycházelo pouze z počtu písmen v české abecedě. Entropii zjištěnou při druhém experimentu, kdy se vycházelo z pravděpodobného výskytu českých písmen, označíme H_1. Symbol H_2 bude označovat entropii s přihlédnutím k relativní frekvenci dvojic písmen, H_3 s přihlédnutím k relativní frekvenci trojic písmen.

Předpokládejme nejprve, že všechna písmena českého textu mají stejnou frekvenci výskytu. V českých textech se vyskytuje 42 písmen (nerozlišujeme písmena „*ů*“ a „*ú*“, uvažujeme i písmeno „*mezera*“). Nyní si vezmeme urnu U_0 a vložíme do ní 42 lístků (na každém lístku je právě jedno písmeno této abecedy různé od ostatních). Vytáhneme jeden lístek, zaznamenáme si písmeno, které je na lístku napsáno, lístek vrátíme zpět do urny a pokus opakujeme. Konečné schéma této situace by vypadalo takto:

$$0 = \begin{pmatrix} \text{-} & \text{a} & \text{á} & \text{b} & \text{c} & \text{č} & \text{d} & \dots \\ \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & \frac{1}{42} & \dots \end{pmatrix}.$$

Tímto postupem může vzniknout například takový „český“ text: *ďj mrgučxýďyaýweaožá*. Obdobné výsledky vytvořené pro ruštinu (podle Dobrušina), pro angličtinu (podle Shannona) a pro němčinu (podle Meyer-Epplera) můžeme srovnat v tabulce 2.7. Je zřejmé, že nápodoba „českého“ textu se příliš nezdařila, proto nyní přihlédneme k relativní frekvenci jednotlivých písmen.

Zjistíme si pravděpodobnosti, s jakými se jednotlivá písmena vyskytují. Text můžeme vytvářet například takto: do urny U_1 vložíme 1 000 lístků, z nichž 163 bude prázdných („*mezera*“), na 73 lístcích bude písmeno „*e*“, na 68 písmeno „*o*“ atd. podle relativní četnosti výskytu českých písmen (můžeme ztotožnit s pravděpodobnostmi jejich výskytu). Konečné schéma by vypadalo následovně:

$$1 = \begin{pmatrix} \text{-} & \text{a} & \text{á} & \text{b} & \text{c} & \text{č} & \dots \\ 0{,}163 & 0{,}054 & 0{,}021 & 0{,}014 & 0{,}010 & 0{,}008 & \dots \end{pmatrix}.$$

Provedeme-li s urnou U_1 stejný pokus jako s urnou U_0, dostaneme například takový výsledek: *žia ep atndi zéuořmp*. Pro první schéma je hodnota entropie $H_0 = 5{,}39$, pro druhé je $H_1 = 4{,}67$.

Nyní budeme pokračovat s nápodobou českého textu a zjistíme relativní četnosti dvojic českých písmen. Tím zjistíme rovněž i tzv. *podmíněnou pravděpodobnost* výskytu v závislosti na písmenu bezprostředně předcházejícím (srov. kap. 2.5). Situace s urnami by mohla vypadat následovně. Vezmeme si 42 uren, které označíme 42 písmeny české abecedy. Do každé urny vložíme lístky s dvojicemi písmen (první písmeno této dvojice bude shodné s písmenem umístěným na urně) v počtu, který odpovídá podmíněné pravděpodobnosti druhých písmen. Pokusy budeme provádět takto. Nejprve vezmeme urnu označenou symbolem „*mezera*" a vytáhneme lístek, z něhož opíšeme písmeno (např. písmeno „*l*"). Vezmeme urnu označenou písmenem „*l*" a z ní vytáhneme lístek, z něhož rovněž opíšeme druhé písmeno (např. „*í*"). Pokračujeme-li stejným způsobem i dále, můžeme dostat například takový text: *lí di oneprá sguluvicechupsv*. Této situaci pak odpovídá 42 konečných schémat

$$- = \begin{pmatrix} \text{-} & \text{a} & \text{á} & \text{b} & \text{c} & \text{č} & \dots \\ \text{p(-|-)} & \text{p(a|-)} & \text{p(á|-)} & \text{p(b|-)} & \text{p(c|-)} & \text{p(č|-)} & \dots \end{pmatrix}$$

$$A = \begin{pmatrix} \text{-} & \text{a} & \text{á} & \text{b} & \text{c} & \text{č} & \dots \\ \text{p(-|a)} & \text{p(a|a)} & \text{p(á|a)} & \text{p(b|a)} & \text{p(c|a)} & \text{p(č|a)} & \dots \end{pmatrix}$$

atd., kde například symbol $p(-|a)$ označuje pravděpodobnost, s jakou se po písmenu „*a*" vyskytuje „*mezera*". Entropii každého z těchto 42 konečných schémat umíme vypočítat podle vzorce pro entropii. V obecném tvaru by vzorec vypadal takto:

$$H(B|A_i) = -\sum_j p(B_j|A_i) \log_2 p(B_j|A_i).$$

Abychom získali na základě těchto jednotlivých entropií (tzv. *podmíněné entropie*) celkovou entropii, musíme vypočítat jejich střední hodnotu. Úlohu vah mají pravděpodobnosti výskytu jednotlivých písmen. Potom

$$E\{H(B|A_i)\} = \sum_i p(A_i) H(B|A_i),$$

což po dosazení vzorce pro podmíněnou entropii upravíme na vzorec

$$-\sum_i \sum_j p(A_i) p(B_j|A_i) \log_2 p(B_j|A_i),$$

z něhož dosazením vzorce

$$p(A)p(B|A) = p(A \cap B)$$

odvodíme vzorec pro entropii H_2, tj. pro entropii podle pravděpodobnosti výskytu dvojic písmen, který je roven

$$-\sum_i \sum_j p(A_i \cap B_j) \log_2 p(B_j|A_i).$$

Kdybychom pokračovali s nápodobou českého textu i dále a zohledňovali pravděpodobnost výskytu trojic písmen, mohli bychom získat text tohoto znění: *dves a vaše miléklár*, který už připomíná text českého jazyka. A konečně obecný vzorec pro výpočet entropie by vypadal takto:

$$H_n = -\sum_i \sum_j p(A_i(n-1) \cap B_j) \log_2 p(B_j|A_i(n-1)),$$

kde $A_i(n-1)$ je pravděpodobnost $(n-1)$-tice písmen, která bezprostředně předchází.

Je vidět, že při rostoucím čísle n dostáváme text, který se přibližuje textu daného jazyka. Protože je ale zjišťování relativní četnosti čtveřic, pětic atd. písmen velice náročné, objevily se různé experimentální metody (např. Shannonova, Kolmogorova) založené na tom, že pokusná osoba postupně hádá písmena, jimiž je tvořen nějaký text. Tímto způsobem sice nemůžeme napodobovat text, ale můžeme zjišťovat entropie vyšších řádů. Zde se ukázalo, že se H_{32} liší od H_{100} již tak nepatrně, že H_{32} lze považovat za dobrý odhad entropie $H \to \infty$. Objevily se samozřejmě také pochyby, zda pomocí těchto experimentů dostáváme stejné hodnoty, které bychom dostali na základě zjišťování relativní četnosti dlouhých posloupností písmen.

Jazyk	H_0	H_1	H_2	H_3	H_∞
Čeština	5,39	4,67	3,87		
Ruština	5	4,35	3,52	3,01	0,87 – 1,37
Angličtina	4,76	4,03	3,32	3,10	1,40
Němčina	4,76	4,10			1,60

Tabulka 2.13: Hodnoty entropie různých řádů pro některé jazyky

Podívejme se na entropii ještě jinak. Každé dostatečně dlouhé sdělení s entropií H lze zakódovat abecedou o m znacích tak, že se průměrný počet znaků překódovaného sdělení připadající na jeden znak sdělení původního prakticky rovná $H/\log_2 m$. Zvolíme-li $m = 2$, pak $H/\log_2 2 = H$. Lze tedy říci, že nám entropie udává průměrný počet znaků binárního kódu připadající na jeden znak původního sdělení, jehož entropii H známe. Rovněž lze pomocí entropie odhadnout počet všech posloupností o n znacích (při využití abecedy s entropií H) a tento počet má hodnotu 2^{nH}. Máme-li tedy knihu o 100 000 znacích s entropií 1,5, ze stejné abecedy by bylo možno sestavit ještě zhruba $2^{100000 \cdot 1,5}$ různých textů o 100 000 znacích s entropií 1,5.

Protože hodnota entropie na úrovni písmen $H^{p\acute{\imath}s}$ závisí na počtu písmen abecedy užívané v daném jazyce, pro srovnání se pracuje s tzv. relativní entropií h, kde

$$h = \frac{H_n}{H_0}$$

(n je řád entropie).

Podobně jako jsme uvažovali o entropii na úrovni písmen, lze mluvit i o entropii na úrovni fonémů, morfémů či slov. Například H_0^{sl} (entropii na úrovni slov) bychom dostali z počtu slovních tvarů jazyka, H_1^{sl} z frekvenčního slovníku na základě výskytu jednotlivých slov a H_∞^{sl} můžeme odhadnout takto: Předpokládejme, že v dostatečně dlouhém textu musí být stejné množství informace na úrovni grafematické či slovní. Známe-li $H_\infty^{pís}$, počet znaků textu a počet slovních tvarů v textu, můžeme sestavit rovnici:

$$H_\infty^{sl} \cdot \textit{počet slovních tvarů v textu} = H_\infty^{pís} \cdot \textit{počet písmen v textu}.$$

Po úpravě pak dostáváme

$$H_\infty^{sl} = H_\infty^{pís} \cdot (\textit{počet písmen v textu / počet slovních tvarů v textu}).$$

Ovšem zlomek „*počet písmen v textu / počet slovních tvarů v textu*“ není nic jiného, než *průměrná délka slova s* (v písmenech). A protože jsme při výpočtu $H_\infty^{pís}$ uvažovali i mezeru, musíme k průměrné délce slova s přičíst ještě 1 (totiž mezeru spojenou s každým slovem). Vzorec pro výpočet H_∞^{sl} je pak následující:

$$H_\infty^{sl} = H_\infty^{pís} \cdot (s + 1).$$

Předešlé úvahy by nás mohly přivést na myšlenku, že podobné pokusy, jaké jsme dělali na úrovni písmen, a to například na úrovni morfémů či slov, by nás mohly jednodušeji a rychleji přivést k nápodobě českého textu. Toto však není dobře možné, neboť přirozený jazyk není jazyk s konečným počtem stavů. Proto je uplatnění pojmu entropie v lingvistice omezeno jen na některé problémy. Teorie informace se v lingvistice uplatnila vlastně jen v oblasti grafematické a fonematické, velmi omezeně morfematické. Německý matematik W. Fuchs použil Shannonovy entropie jako statistické charakteristiky individuálního literárního stylu, když zjišťoval hodnoty entropie pro rozložení slovních délek vyjádřených počtem slabik. Uvádí vzorec pro vztah mezi průměrným počtem slabik $\bar{i}$ (ve slovech daného jazyka) a procentem p_i slov o i slabikách:

$$p_i = \frac{e^{-(\bar{i}-1)}(\bar{i}-1)^{i-1}}{(i-1)!}.$$

Sami představitelé kvantitativní lingvistiky a teorie informace (např. C. E. Shannon) si ovšem uvědomovali omezené možnosti své metody. A je třeba zdůraznit, že vedle kvantitativního hlediska je při studia jazyka nutno uplatňovat vždy i hledisko kvalitativní.

Dalším důležitým pojmem teorie informace užívaným v lingvistice je *redundance*. Používá se místo tzv. relativní entropie h dané vzorcem

$$h = \frac{H_n}{H_0}$$

(kde n je řád entropie). Redundance je určena vzorcem

$$R_n = 1 - \frac{H_n}{H_0},$$

kde index n u R značí, že jde o redundanci příslušnou k entropii řádu n. Redundance udává procento nadbytečných jednotek, znaků sdělení o entropii H_n. Redundance je číslo, které nabývá hodnot od 0 do 1. Hlavní význam redundance spočívá v tom, že zabezpečuje spolehlivost sdělení. Její význam lze rovněž vidět pro srovnávací studium jazyků.

Jazyk	R_0	R_1	R_2	R_3	R_∞
Čeština	0	0,13	0,28		
Ruština	0	0,13	0,30	0,40	0,72 – 0,82
Angličtina	0	0,16	0,30	0,35	0,71
Němčina	0	0,14			0,66

Tabulka 2.14: Hodnoty redundance příslušné k hodnotám entropie z tab. 2.13

Jednotka množství informace se nazývá *bit* (zkratka z angl. *binary digit*, tj. binární jednotka). Je to jednotka daná abecedou o jednom prvku a dvou stavech, tedy

$$bit = \log_2 2^1.$$

Je to jednotka založená na binárním (dvoustranném) protikladu *ano – ne*. Užívá se jí k měření informace z praktického důvodu, neboť většina technických systémů k přechovávání a přenosu informací je na tomto binárním principu založena. Například při přenosu informace Morseovou abecedou se využívají místa, v nichž buď je nebo není elektrický impuls, pokud tam je, pak je buď krátký („*tečka*“) nebo dlouhý („*čárka*“) apod. Z dvojčlenného charakteru binární jednotky vyplývá, že máme-li např. skupinu osmi prvků, pak si vystačíme s třemi rozhodnutími typu *ano – ne*, abychom určili jakýkoliv z těchto prvků (k určení každého z 16 prvků nám stačí 4 kroky, k určení 32 prvků 5 kroků atd.).

Příklad 2.3: Mějme sdělení S, které je zaznamenáno abecedou $A = \{a_i\}$ (kde $i = 1, 2, \ldots, m$[151]). Toto sdělení má N znaků a entropii H_n. Entropii lze interpretovat jako průměrný počet znaků připadající na jedno písmeno sdělení S, zakódujeme-li je nejekonomičtějším binárním kódem. Mějme například sdělení S dané posloupností znaků „*ACAABDABBAAADCBA*“ (tj. $N = 16$). Zjistíme relativní četnosti výskytu jednotlivých písmen. Pak tomuto sdělení odpovídá konečné schéma

$$A = \begin{pmatrix} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}.$$

[151] Pro zjednodušení uvažujeme, že je m mocnina 2.

Spočteme entropii H_1 (tj. entropii písmen s ohledem na relativní četnosti písmen) a dostaneme

$$-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{4}\log_2\frac{1}{4}+\frac{1}{8}\log_2\frac{1}{8}+\frac{1}{8}\log_2\frac{1}{8}\right)=$$

$$-\left(\frac{1}{2}(-1)+\frac{1}{4}(-2)+\frac{1}{8}(-3)+\frac{1}{8}(-3)\right)=\frac{7}{4}=1{,}75.$$

Nyní můžeme snadno zvolit příslušný nejekonomičtější binární kód: $A \rightarrow 0$, $B \rightarrow 10$, $C \rightarrow 110$, $D \rightarrow 111$. Překódováním dostaneme sdělení S' = „*01100010111010100001111110100*“. Toto sdělení S' má délku $N' = NH_1$, což je $16 \cdot 7/4 = 28$ binárních znaků. Zakódujeme nyní sdělení S' původní abecedou A, a to tak, že každé H_0-tici (další možná interpretace veličiny H_0), tj. v našem případě každé dvojici (abeceda A má 4 prvky, $H_0 = \log_2 4 = 2$) sdělení S' přiřadíme právě jedno a_i, například takto: $00 \rightarrow A$, $01 \rightarrow B$, $10 \rightarrow C$, $11 \rightarrow D$, dostaneme sdělení S'' = „*BCACDCCCABDDBA*“ s délkou $N'' = 14$. Protože S' je zakódována nejekonomičtějším binárním kódem, v němž se oba binární znaky vyskytují se stejnými relativními četnostmi a nezávisle na sobě, vyskytovaly by se i v dostatečně dlouhém sdělení typu S'' (v našem konkrétním případě ne, neboť posloupnosti znaků jsou příliš krátké) se stejnými relativními četnostmi všechny H_0-tice, a tedy i všechna písmena abecedy A. Protože S' mělo NH_1 binárních znaků a my jsme kódovali po skupinách o H_0 znacích, je S'' dlouhé $NH_1/H_0 = 28/2 = 14$. Nyní zjistíme, jaký je rozdíl mezi délkou S' a S''. Odečteme délku sdělení S'' od délky sdělení S a výsledek dělíme délkou sdělení S:

$$\frac{N - N\frac{H_1}{H_0}}{N}.$$

Po krácení proměnnou N dostaneme výraz

$$1 - \frac{H_1}{H_0} = R_1.$$

Můžeme tedy R_1 (a obecně R_n) interpretovat jako procento redundantních, nadbytečných znaků sdělení o entropii H_1 (obecně H_n).

Uvědomme si, že zprávu S'' jsme dostali postupem značně složitějším než seškrtáním $R_1 \cdot 100$ % znaků zprávy S. Dále si uvědomme, že o nadbytečných znacích můžeme mluvit pouze v ideálních podmínkách sdělovacího procesu (neexistence poruch, nemožnost zkreslení apod.). Souhrnně všechny takové poruchy nazýváme termínem *šum*, který byl do jazykovědy a do teorie informace převzat od spojovacích techniků a slouží k označení jakékoliv poruchy, k níž při přenosu informace dojde. Podmínky bez existence šumů ale nikdy splněny nejsou. Nelze tedy redundanci chápat jako něco nadbytečného, zbytečného. Význam redundance spočívá v zajištění spolehlivosti sdělovacího procesu.

Kdybychom využívali naši českou abecedu co nejekonomičtěji, všechna písmena by se vyskytovala se stejnou pravděpodobností. Každá variace s opakováním sestavená z písmen naší abecedy by byla českým slovem. Pak by ovšem

jakákoliv chyba při psaní či tisku mohla zcela změnit význam sdělení. Skutečnost je ale zcela jiná. Má-li čeština 42 písmen, z toho 14 „*samohláskových*“ a 28 „*souhláskových*“, pak by při nejekonomičtějším využití české abecedy existovalo 392 slov tvaru „*samohláska + souhláska*“ (tj. $14 \cdot 28 = 392$) a 42 slov tvořených pouze jedním písmenem. Ve skutečnosti jen malá část z nich jsou česká slova – dvoupísmenná například *ač*, *ach*, *ať*, *au*, *as*, *až*, *ob*, *oč*, *od*, *och*, jednopísmenná *a*, *i*, *o*, *u*, *k*, *s*, *v*, *z* (jednopísmenných je jen zhruba $\frac{1}{4}$). A pouze díky redundanci pak můžeme správně porozumět i zkomolenému textu:

PRIJEDU ZIBRA VEKER.

Hodnoty příslušné redundance zjištěné z hodnot entropie pro různé jazyky kolísají v rozmezí $0{,}70 \pm 0{,}10$. Nelze říci, jestli je toto kolísání dáno skutečnými odchylkami redundance v různých jazycích nebo jestli je to následek nejednotnosti metody v zjišťování redundance. V každém případě se tu naskýtá rozsáhlá oblast jazykovědných výzkumů. Pokud se opravdu hodnoty redundance v různých jazycích různí, pak je třeba zkoumat, jakými vlastnostmi konkrétního jazyka je hodnota redundance podmíněna. Pokud se ukáže, že hodnoty redundance jsou pro všechny jazyky zhruba stejné, pak nás napadá řada dalších otázek:

1) Proč má redundance právě tuto zjištěnou hodnotu? Zde pravděpodobně může být nápomocna psychologie, fyziologie apod., neboť tato hodnota bude nejspíše ovlivněna podmínkami mluvení, vnímání řeči atd.

2) Jaké je místo oněch prostředků, postupů, kterými se v přirozených jazycích dosahuje zjištěné hodnoty redundance, mezi všemi takovými postupy? Zde by mohlo být nápomocno srovnávací studium přirozených jazyků a umělých kódů studovaných a konstruovaných v teorii kódování.

3) Proč se v přirozených jazycích uplatňují právě tyto postupy?

Tyto výzkumy jsou velmi cenné, neboť začleňují tyto ukazatele do jiných poznatků o jazyce. A rovněž svůj praktický význam měly některé jednoduché aplikace teorie informace na studium přirozených jazyků – na základě jejich poznatků byly například sestavovány telegrafní kódy novoindických jazyků.

2.9 Glottochronologie

Zajímavou aplikací matematiky v jazykovědě je tzv. *glottochronologie* (též *lexikostatistika*). Je to lexikologická metoda, která pomocí statistiky zjišťuje dobu vzniku jazyka, respektive různých jazyků. Tato metoda vznikla v 50. letech 20. století a za jejího zakladatele je považován americký antropolog Morris Swadesh[152]. Ve stejné době uvádí podobné výsledky i Američan Robert B. Lees[153].

[152] Swadesh, M.: *Lexico-Statistic Dating of Prehistoric Ethnics Contacts*. Proceedings of the American Philosophical Society 96, 1952, 452–463.

[153] Lees, R. B.: *The Basis of Glottochronology*. Language 29, 1953, 113–127.



Obrázek 2.2: Morris Swadesh

Cílem glottochronologie je pomocí kvantitativních metod zjistit, v které době došlo k rozrůznění určitého jazyka nebo prajazyka na dva nebo více jazyků moderních. Příbuzenské vztahy jazyků se měří na základě změn v slovní zásobě, lexiku (odtud název *lexikostatistika*).

Glottochronologie byla inspirována tzv. *radiokarbonovou metodou*, u jejíhož vzniku stál americký chemik Willard Frank Libby a pomocí níž lze stanovit stáří organických látek. Po zániku organických látek dochází k rozpadu radioaktivních jader uhlíku a tento rozpad se řídí exponenciální funkcí, to znamená, že po uplynutí určitého časového intervalu klesne počet radioaktivních jader na polovinu, po uplynutí téhož času znovu klesne na polovinu atd. Tento časový interval, během něhož klesne počet radioaktivních jader uhlíku na polovinu, se nazývá *poločas rozpadu*. Pro izotop uhlíku stanovil Libby poločas rozpadu na 5 568 let. V roce 1960 byla Libbymu udělena za tuto metodu Nobelova cena za chemii. Později se přišlo na to, že poločas rozpadu izotopu uhlíku byl stanoven nepřesně a pomocí tzv. *dendrochronologické kalibrace* byla jeho hodnota upřesněna na 5 730 let.

A právě tato radiokarbonová metoda inspirovala na přelomu 40. a 50. let 20. století M. Swadeshe k její aplikaci na vývoj jazyků. V první verzi glottochronologie z roku 1952 určil jakési *jádro slovní zásoby*, nejprve o 215, později o 200 lexikálních jednotkách, u nichž předpokládal, že jsou dostatečně univerzální a stabilní, nezávislé na kulturních či zeměpisných podmínkách (např. *matka*, *otec*, *pták*, *ryba*, *pes*, *muž*, *žena*, *dlouhý*, *velký*, *malý*, *znát*, *spát* apod.). Předpokládal, že v tomto jádru slovní zásoby se výměna slov řídí obdobným exponenciálním zákonem jako rozpad radioaktivních jader izotopu uhlíku, tj. ke změnám dochází poměrně pomalu a mizení slov ze slovníkového jádra je konstantní. Ve spolupráci se specialisty pak zjišťoval, jakou rychlostí dochází ke změnám v tomto jádru, a to u jazyků s dostatečně dlouhou literární tradicí. Průměrná hodnota změn v testovacím souboru se pohybovala okolo 19,5 % za časový interval tisíc let. Po kritikách glottochronologie publikoval Swadesh v roce 1955 novou studii, v níž slovníkové jádro redukoval na 100 jednotek, některé pojmy slovníkového jádra nahradil pojmy univerzálnějšími a přepočítal i *rozpadovou konstantu* λ, jejíž hodnotu stanovil na 14 %, to znamená, že za tisíc let zůstane zachováno 86 % slov ze slovníkového jádra.

Nejčastěji se při zjišťování *časové hloubky* (tzn. období, kdy došlo k rozrůznění zkoumaných jazyků) používají tyto dva vzorce:

$$i_{(t)} = \frac{\ln C}{2 \ln r} \tag{2.1}$$

$$i_{(t)} = \frac{\ln C}{-2\lambda} \tag{2.2}$$

kde
$i_{(t)}$ ... časová hloubka (uběhlý čas);
C ... procento společně zděděných párů slov z počtu všech párů slov v slovníkovém jádru obou zkoumaných jazyků;
r ... index rychlosti mizení slov z jádra (procento dvojic uchovaných za určitou jednotku času, zpravidla tisíc let), index r je komplementární k rozpadové konstantě λ (tzn. $r = 1 - \lambda$), dle Swadeshových výpočtů klademe $r = 0{,}86$;
λ ... rozpadová konstanta (konstanta komplementární k r, tzn. $\lambda = 1 - r$), $\lambda = 0,14$.

Při zjišťování *časové hloubky* postupujeme tak, že stanovíme zpravidla 100 výrazů základního jádra slovní zásoby v těchto jazycích a jejich porovnáním zjistíme procento shodných a různých dvojic. Dosazením této hodnoty a indexu rychlosti mizení slov (popř. rozpadové konstanty) do příslušného vzorce získáme hledanou časovou hloubku.

P. Novotná ve své rigorózní práci [42] uvádí postup při hledání doby divergence němčiny a francouzštiny. V základním slovníkovém jádru obou jazyků nacházíme 33 párů společně zděděných slov, lze tedy položit $C = 0,33$. Dosazením do rovnice (2.1) dostaneme dobu divergence v tisíciletích:

$$i_{(t)} = \frac{\ln 0{,}33}{2 \ln 0{,}86} = \frac{-1{,}10886}{-0{,}30164} = 3{,}675.$$

Jiná situace nastává v případě, že nemáme synchronní dvojici příbuzných jazyků, tzn. každý z dvojice studovaných jazyků je zachycen v jiné době (od rozpadu jejich společného prajazyka do doby jejich zachycení uběhla různě dlouhá časová období $i_{(t_1)}$ a $i_{(t_2)}$). V tomto případě Swadesh vychází ze vztahu (2.1), který lze přepsat jako

$$2i_{(t)} = \frac{\ln C}{\ln r}, \tag{2.3}$$

a ze vztahu

$$i_{(t_1)} + i_{(t_2)} = \frac{\ln C}{\ln r}. \tag{2.4}$$

Zpravidla neznáme ani jeden z časových úseků $i_{(t_1)}$ a $i_{(t_2)}$, jen jejich rozdíl $\triangle i_{(t_{12})}$. Nahradíme proto součet $i_{(t_1)} + i_{(t_2)}$ výrazem $i_{(t_1)} + i_{(t_1)} + \triangle i_{(t_{12})} = 2i_{(t_1)} + \triangle i_{(t_{12})}$, kde $i_{(t_1)}$ je kratší z obou časových úseků $i_{(t_1)}$, $i_{(t_2)}$. Odtud dostáváme pro dva asynchronně zaznamenané jazyky vztah

$$i_{(t_1)} = \frac{\ln C}{2 \ln r} - \frac{\triangle i_{(t_{12})}}{2}, \tag{2.5}$$

kde

$$i_{(t_1)} = min(i_{(t_1)}, i_{(t_2)}).$$

Glottochronologie měla pozitivní ohlas u lingvistů zabývajících se jazyky bez dlouhé literární historie, naopak kritického přijetí se jí dostalo zejména od indoevropeistů, neboť u těchto jazyků s delší literární tradicí se výsledky glottochronologických výpočtů výrazně lišily od již známých výsledků. K nejvýraznějším kritikům patřili např. E. Coseriu a W. W. Arnold. Nejčastěji bylo glottochronologické metodě vytýkáno to, že výběr slov jádra a jejich počet je značně subjektivní, což je v rozporu s přesným aparátem zpracování. Dále bylo upozorňováno na to, že rychlost rozpadu slovníkového jádra není vždy konstantní, neboť na ni mají vliv i vnější vlivy.

Zajímavé modifikace lexikostatistické metody předložili zejména Kanaďanka Sheila Embletonová a Rus Sergej Starostin. Oba autoři se shodují v tom, že Swadeshova glottochronologie nerozlišuje záměnu slova za inovaci od náhrady výpůjčkou (více viz [42]). Starostin navrhuje proto nejprve očistit základní slovníkové jádro od jakýchkoliv výpůjček. Výsledkem je menší hodnota rozpadové konstanty λ. Ve svých úvahách jde ale mnohem dále a odvozuje nové vzorce pro výpočet časové hloubky (tzv. *rekalibrovaná glottochronologie*).

Metody glottochronologie využívala ke studiu různých jazyků celá řada autorů. Ke studiu indiánských a eskymáckých jazyků ji jako první použil sám M. Swadesh. V 60. letech ji aplikoval na studium příbuznosti jazyků kavkazských G. A. Klimov. U nás se problematikou diferenciace slovanských jazyků pomocí lexikostatistické metody zabývali brněnští jazykovědci M. Čejka a A. Lamprecht[154]. S její pomocí se pokusili vymezit dobu rozpadu praslovanské jednoty na větev západní, východní a jižní. Nedostatky metody se snažili vyrovnat rozborem fonologickým a morfologickým. Rozpad praslovanštiny stanovili na konec 10. století (podle glottochronologie 8. – 11. stol.). V současnosti se glottochronologické metodě věnují na FF Masarykovy univerzity v Brně P. Novotná a V. Blažek. Bibliografii vztahující se ke glottochronologii, ale i k problematice využívání kvantitativních metod v historické a srovnávací lingvistice či dialektologii, lze nalézt rovněž u M. Těšitelové (viz [66]).

2.10 Jan Amos Komenský

Ve svém článku *J. A. Komenský a racionalizace osvojování si slovní zásoby*[155] uvádí Marie Těšitelová, že Jan Amos Komenský byl jedním z prvních autorů, který využíval frekvence slov k ekonomickému rozvíjení slovní zásoby žáků, a to ve svém díle *Janua linguarum reserata* (*Brána jazyků otevřená*). Autorka vychází z článku D. Harkina[156], ve kterém autor čtenáře seznamuje s historií měření slovní zásoby k nejrůznějším účelům – například tiskařská

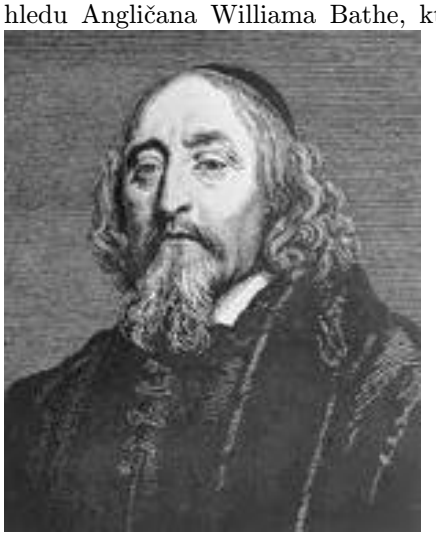
[154] M. Čejka – A. Lamprecht: *K otázce vzniku a diferenciace slovanských jazyků.* Sb. fil. fak. brněnské university 1963, A 11, s. 1–20; M. Čejka: *Lexicostatistic Dating and Slavonic Languages.* Sb. fil. fak. brněnské university 1972, A 20, s. 39–52.

[155] Slovo a slovesnost 31, 1969, s. 187–191.

[156] *The History of Word Counts.* Babel (Revue Internationale de la traduction) 2, 1957, č. 3, s. 113–124.

praxe, stenografie, technika, didaktika a jazykové vyučování. V tomto článku Harkin rovněž zmiňuje, že J. A. Komenský byl mezi prvními, kteří ocenili význam znalosti výskytu slov pro efektivnost vyučování, zvláště cizím jazykům, a to právě v díle *Janua linguarum reserata*. Dále Harkin naznačuje inspirační zdroje Komenského pro napsání tohoto díla a seznamuje nás se stavbou tohoto díla. Dopouští se ale řady nepřesností.

Jako přímého předchůdce J. A. Komenského uvádí Harkin ve svém přehledu Angličana Williama Bathe, který roku 1611 vydal *Janua linguarum*.



Komenský pak podle jeho vzoru vydal údajně roku 1665 v Amsterodamu svoji *Janua linguarum reserata*. Z komeniologické literatury je ale známo, že 1. vydání Komenského *Janua linguarum reserata* vyšlo již roku 1631 v Lešně. Ještě roku 1631 Komenskému vydává v Londýně J. Anchoranus tento původně pouze latinský text s doplněným paralelním překladem anglickým a francouzským, roku 1632 vychází v Gdaňsku německý překlad J. Mochingera a roku 1633 rovněž v Gdaňsku trojjazyčná paralela latinská, německá a polská, kterou pořídil A. Wengerscius. První český překlad s názvem *Dvéře jazyků odevřené, tj. Krátký a snadný způsob kteréhokoli jazyka, spolu i s začátky všech umění svobodných, pochopení, v kterémž pod stem titulů v tisíci průpovědech všecka slova celého jazyka obsažena jsou nejprv v jazyku latinském a nyní v českém zhotovené od J. A. C.* (zkráceně *Dvéře jazyků odevřené*) vychází roku 1633, a to „*pro nedostatek papíru a některé jiné překážky*“ (viz [26]) bez latinské paralely. V následujících letech vyšla *Janua linguarum reserata* v mnoha vydáních a v dalších jazycích, např. řečtině, italštině, švédštině, španělštině, ale i v jazycích orientálních, zpravidla ale s velkými zásahy upravovatelů – např. latinský text značně rozšířil Z. Schneider, překladatel do řečtiny. Mezi tato další vydání patří i vydání amsterodamské, jež měl patrně na mysli Harkin ve svém článku. Toto vydání pořídil J. Seidel v tzv. edici latinsko-vlámské (latino-belgicae) roku 1666, a ne roku 1665 s paralelami latinskými, řeckými a francouzskými, jak uvádí Harkin.

V předbělohorském období plánoval Komenský sepsat v mateřském jazyce základní naučná díla, která by pomohla vytvořit českou národní vědu. Čeština by se tak vymanila ze závislosti na latině a byla by oporou pro další literární tvorbu v rodné řeči. Tyto své plány vyjádřil v latinské předmluvě *Ad eruditos gentis meae* (*K vzdělancům mého národa*) ke spisu *Theatrum universitatis rerum* (*Divadlo světa*). Na této encyklopedicky zaměřené práci začal pracovat roku 1614 a z původně plánovaných 16 knih (oddílů) dokončil pouze malou část. Příčinou mohla být obtížnost tohoto úkolu, nepříznivá situace po bitvě na Bílé hoře či zničení již hotových částí při požáru v Lešně roku 1656. Pro

své didaktické plány Komenský zamýšlel projekt jakési encyklopedie[157], kterou měly tvořit dvě části: teoretická (vlastní didaktika) a část praktická (učebnice a informatoria pro učitele a vychovatele). Celek se měl jmenovat *Paradius ecclesiae renascentis*. Nám se z těchto plánů zachoval v rukopise jenom úvod, *Didactica* a *Informatorium školy mateřské*. Pravděpodobně i tištěné *Dvéře jazyků odevřené* byly součástí tohoto programu *Paradius ecclesiae renascentis*. Komenský je dokončil před ostatními didaktickými spisy, protože se mu hodily pro jeho učitelské působení na lešenském gymnasiu[158]. Ostatní části zamýšleného souboru Komenský nejspíše již nenapsal.

Sám Komenský v předmluvě k prvnímu latinskému vydání i k vydání českému kritizuje dosavadní způsob výuky cizích jazyků, kdy jsou studenti zdržováni gramatickými poučkami, učením se nazpamět dlouhých vět bez spojování jednotlivých slov s jejich významy apod. Přínosné by podle něj bylo vytvoření jakéhosi kompendia celého jazyka, ve kterém by byla uspořádána všechna slova a jejich spojení tak, aby pomocí něj mohl člověk snadno získat přístup k jakémukoliv autorovi „*jako dvéře odevřené měl*", aniž by musel svou slovní zásobu roztroušeným čtením, poslouchání či mluvením získávat. Uvádí, že pokud jde o myšlenku napsat pomůcku pro efektivní studium slovní zásoby, inspirací mu byla práce *Janua linguarum* v jazyce latinském a španělském, která před několika lety vyšla „*pod titulem Collegium hybernského v Salamantyce, hispánském městě*", „*kdež ve dvanácti stech sentencí obsáhl všecko přednější latinského jazyka slova, a to pořádně, tak, že žádné slovo (krom slovíček pojících jest, kterýž, od, ale etc.) neklade se než jednou v celé knížce*" ([26], str. 10).

[157]Tyto příručky inspirované obdobnými díly, které se v té době objevovaly, nepochybně souvisejí i s přípravou slovníku *Linguae Bohemicae Thesaurus* (*Poklad českého jazyka*). Tento velký slovník Komenský již připravoval pro tisk a pracoval na něm neuvěřitelných 44 let od roku 1612. Měl nahradit zastaralé slovníky, zejména slovníky Tomáše Rešela a Daniela Adama z Veleslavína. Nakonec shořel při požáru Lešna, když ze sebraného materiálu stihl vyjít tiskem jen jakýsi zárodek díla, *Kořenové jazyka českého*. Tento spisek je dnes nezvěstný, ale znal jej ještě Václav Rosa a těžil z něj pro svůj rukopisný slovník *Thesaurus linguae Bohemicae*.

[158]Vedle toho je J. A. Komenský rovněž autorem latinských spisů o matematice *Geometria theoretica* a *Secunda pars Geometriae Geodesia dicta*, dvoudílné učebnice pro lešenské gymnázium, která se dochovala v rukopise.


J. A. Comenii
JANUA LINGUARUM
reserata.
ANNO M. DC. LXVI.
J. A. Comenii
Ontsloote Deure der Talen,

Komenskému se metoda uspořádání slovní zásoby *Januy* zalíbila, ale k její realizaci měl některé výhrady:
1) Z názvu „*dvéře*“ Komenský cítí, že tato práce by měla být jakýmsi vstupem do poznání základů daného jazyka. Její součástí by měla být podle něj tedy slova každodenní potřeby, což v případě tohoto díla rozhodně neplatí. Je zde řada slov řídkých a zastaralých (Komenský používá termíny „*neobyčejných*“ a „*zvetšelých*“), která jsou obtížná a zbytečná pro počáteční studium jazyka. M. Rhenius poznamenává, že lépe by odpovídal tomuto pojetí místo označení „*dveře*“ název „*žebřík*“.
2) Každé slovo se uvádí pouze jednou, a to i tehdy, má-li více významů.
3) Slova se uvádějí většinou ve významech přenesených, protože „*autor ozdob sentencí květovaným mluvením všudy hledá*“ ([26], str. 10).

Na popud M. I. Rhenia se tedy Komenský ujal úkolu sestavit podobný slovník pro studium latinské slovní zásoby. Vyšel přitom z těchto zásad:

1) Je třeba, „*aby rozum a jazyk vždycky vedle sebe šli*“ ([26], str. 11). Sestavil tedy „*sto nejobyčejnějších všech věcí titulů*“, dnes můžeme říci 100 nejběžnějších sémantických okruhů (původ světa, živly, rostliny, zvířata, voda, země, obloha, člověk, tělo, nemoci, mysl, lidské činnosti a vlastnosti apod.).

2) K těmto tematickým okruhům vyhledal ze slovníků asi 8 000 slov a použil je v 1 000 větách, aby popsal všechny základní skutečnosti.

3) Snažil se dodržet, aby se každé slovo vyskytlo pouze jedenkrát: jednalo-li se ale o slova víceznačná (tj. s více významy), nebo o slova mající významový vztah k slovu jinému, potom je Komenský opakoval. U slov víceznačných se snažil dokonce dodržovat i pořadí významů, aby na prvním místě nestál význam vedlejší.

4) V hranatých závorkách uvádí synonyma.

V „*pozdravení*“ k prvnímu českému vydání z roku 1633 Komenský rovněž zmiňuje něco, co známe z moderního jazykového vyučování, a to rozlišování tzv. *aktivní* a *pasivní slovní zásoby*. Ačkoliv toto vydání postrádá latinskou paralelu, lze podle něj užívat tuto knihu i bez ní, samostatně. Komenský se snažil o výběr všech slov „*potřebnějších*“. Nejsou to sice všechna slova české

slovní zásoby, ale jsou to nejen slova, kterých se aktivně užívá (*aktivní slovní zásoba*), ale i slova, kterým stačí rozumět (*pasivní slovní zásoba*). Za zmínku stojí rovněž jeho lingvistická připomínka toho, že latina na rozdíl od češtiny nerozlišuje vid, „*protož nenajdeš-li přicházeti, hledej přijíti, nenajdeš-li odvozovati, hledej odvézti...*“ ([26], str. 12). Na závěr dodává, že měl ještě v úmyslu přidat rejstřík českých slov, ale nedostatek místa a čas potřebný k vypracování mu to neumožnili. Na volné místo ale přidal 4 překlady latinských žalmů z roku 1626.

Srovnáním latinského a českého znění můžeme vidět snahu o co nejvěrnější shodu. Podíváme-li se ale na amsterdamské vydání z roku 1665, které zmiňuje ve svém článku Harkin, jsou patrné výrazné rozdíly. Při stejném počtu tematických okruhů a počtu vět, výrazně vzrostl počet užitých slov. V textu jsou užita i slova málo běžná. Právě s tím souvisí nejspíše Harkinova poznámka, že se jedná o slovník speciální, ne obecný. Česká verze *Januy* nedosáhla takové obliby jako podoba latinská. Na vině patrně byla nepříznivá vnitřní situace v Čechách, možnost jejího využití a to, že nevyšla společně s latinským textem. *Dvéří jazyků* si všimli až jezuité, kteří vedeni cílem účinněji prosazovat protireformaci vydali latinsko-německo-českou verzi *Januy* roku 1669 (i když tisk byl započat již roku 1667). O účasti Komenského na tomto vydání nemáme žádné důkazy.

Podle Těšitelové lze Komenského právem „*zařadit mezi ty, kteří se snažili vnést systém do osvojování si slovní zásoby a hierarchizovat tu slova podle objektivních měřítek.*“ Komenský rozlišuje slova podle jejich běžnosti a potřebnosti, pracuje ale také (řečeno dnešní terminologií) s aktivní a pasivní slovní zásobou a snaží se stanovit jejich počet. Klade důraz na spojování formy slova s obsahem, zaměřuje se na synonyma i stylové zabarvení slov. Můžeme tu vysledovat moderní názory, které ale byly bohužel dalšími překladateli a upravovateli setřeny. Tyto myšlenky Komenského posouvají až do dvacátých let 20. století, kdy se zejména díky odkazu francouzského lingvisty Ferdinanda de Saussura začínají formovat hlavní zásady strukturální lingvistiky, která zcela změnila tehdejší pohled na jazykovědu. A až nové směry ve vyučování cizím jazykům se znovu vrací k těmto moderním myšlenkám.

Nelze tedy zpochybnit, že v lingvistickém díle J. A. Komenského (podobně jako v jeho díle pedagogickém či didaktickém) nacházíme řadu v jeho době průkopnických a moderních myšlenek. Vesměs se jedná o paralelu se *strukturalismem*, tak jak se formuloval ve 20. letech 20. století. Nikde v pracích J. A. Komenského nenacházíme ale žádné konkrétní odkazy na to, jak přesně slovní zásobu počítal, jaká objektivní měřítka používal (výběr slov do učebnice *Janua linguarum reserata* závisel dle mého mínění zcela na jeho subjektivním zhodnocení). Uvádí pouze počty slov zahrnutých do některých svých děl (např. první latinské vydání *Janua linguarum reserata* obsahuje osm tisíc slov uspořádaných do tisíce vět v deseti tematických okruzích). A ačkoliv strukturální jazykověda sehrála důležitou roli při rozvoji matematické lingvistiky, nelze ještě díky těmto názorům, ač jistě pokrokovým, považovat J. A. Komenského za předchůdce kvantitativní (respektive matematické) lingvistiky.

2.11 Časopis Krok a počátky využívání matematiky v lingvistice v českých zemích

Časopis *Krok* je prvním českým vědeckým časopisem. Vycházel v letech 1821 až 1840 především za redakce významného obrozenského vědce a propagátora češtiny Jana Svatopluka Presla. V tomto časopise byly v roce 1831 otištěny dva zajímavé články, které svou tematikou můžeme zařadit do matematické lingvistiky a které vzhledem k době vzniku můžeme považovat za jedno z prvních využití matematiky v lingvistice vůbec, nejen co se týče české jazykovědy.

Ačkoliv bývá jako jeden z prvních vědců, který se zabýval kvantitativními jevy v jazyce, uváděn americký lingvista W. D. Whitney (1827–1894) – viz např. [6], lze nalézt v prehistorii kvantitativní lingvistiky práce mnohem starší[159].

V této souvislosti si proto zcela jistě zaslouží pozornost neznámá hlásková statistika, kterou P. Grzybek zmiňuje ve svém článku [18] a která vyšla roku 1831 v prvním českém vědeckém časopise *Krok*, redigovaném Janem Svatoplukem Preslem (1791–1849), tedy celých patnáct let před nejstarší ze zmíněných prací a více než čtyřicet let před prací W. D. Whitneyho. Lze se tedy domnívat, že se jedná o jedno z nejstarších užití matematických metod v jazykovědě (konkrétně při zjišťování frekvencí hlásek) vůbec.

2.11.1 Jan Svatopluk Presl a časopis Krok



Obrázek 2.3: Jan Svatopluk Presl

[159]Současný německý badatel Peter Grzybek uvádí jako jedny z prvních prací v oblasti kvantitativní lingvistiky práce německých vědců Ernsta Förstemanna (1822–1906) a Augusta Schleichera (1821–1868). Förstemann se ve svých studiích z let 1846 a 1852 zabýval frekvencí hlásek a písmen v němčině, řečtině, latině a gótštině. August Schleicher v roce 1852 publikoval svou práci *Formenlehre der kirchenslawischen Sprache*, kterou zamýšlel jako doplněk k pracem Ernsta Förstemanna a ve které se zabýval frekvencí písmen a hlásek ve staré církevní slovanštině.

"kniha˙final" — 2012/7/27 — 22:04 — page 110 — #116

Jan Svatopluk Presl se narodil 4. září 1791 v Praze. Po studiu medicíny na Karlovo-Ferdinandově univerzitě v Praze (ukončil jej v roce 1816 titulem doktor medicínských věd) pracoval dva roky jako asistent zoologie a mineralogie, potom se stal profesorem těchto oborů v Olomouci a roku 1820 i v Praze. Patřil mezi nejvýznamnější české přírodovědce 19. století. V období národního obrození, kdy oficiálním jazykem byla němčina, byl aktivním propagátorem českého jazyka a kultury. Udržoval kontakt se všemi čelními představiteli vědy a kultury, kteří podporovali české národní obrození (Josef Jungmann, Josef Dobrovský, Václav Hanka, Jan Evangelista Purkyně, František Palacký, Pavel Josef Šafařík aj.). Zasloužil se o rozvoj českého odborného názvosloví celé řady vědních oborů – botaniky, zoologie, ale i chemie, mineralogie, technologie či paleontologie. Rovněž zavedl v chemii dvouslovné pojmenovávání anorganických sloučenin. Presl je například autorem slov *tuleň*, *vorvaň*, *kyslík*, *vápník*, *draslík*, *cínovec*, *kazivec* a *živec*. Do roku 1849 byla v češtině vytvořena odborná terminologie pro všechny tehdejší vědecké obory. Na jejím vytváření se vedle J. Svatopluka Presla a jeho bratra Karla Bořivoje Presla (1794–1852) podíleli například lékař Antonín Jungmann (antropologicko-lékařská terminologie), František Palacký (estetika a historie), Josef Jungmann (jazykověda) či profesor gymnázia a filozofie v Plzni Josef Vojtěch Sedláček (1785–1836), který zaváděl fyzikální a matematické (přesněji geometrické) názvosloví. Mezi nejvýznamnější vědecké práce J. S. Presla patří *Flora Čechica*, kterou vydal roku 1819 společně se svým bratrem Karlem Bořivojem Preslem a která je psána ještě latinsky. Od roku 1820 vydávají oba bratři práci *O přirozenosti rostlin aneb Rostlinář*, ve které jsou uvedena i pravidla tvoření českých rodových jmen rostlin a terminologie organické chemie.

J. S. Presl byl propagátorem myšlenky *České společnosti*, která vyústila později právě ve vydávání prvního českého vědeckého časopisu *Krok* (podrobněji se o něm zmíníme ještě dále), jehož byl J. S. Presl v letech 1821–1837 redaktorem. Stal se jedním ze zakladatelů *Vlasteneckého muzea* v Čechách v dubnu roku 1818 a definitivně byla myšlenka *České společnosti* naplněna v roce 1830, když Společnost muzejní zřídila *Sbor k vědeckému vzdělávání řeči a literatury české*, jehož členy se stali vedle Presla i Palacký a Jungmann. Tento sbor pomohl vytvořit pokladnici české literatury – tzv. *Matici českou*, která měla zajišťovat vydávání českých knih, především úplného českého slovníku a slovníku encyklopedického. Tím byly splněny základní podmínky pro pěstování národní vědy. Aktivně se účastnil politického života, v roce 1848 byl zvolen do *České národní rady* a byl rovněž jedním z 341 delegátů *Všeslovanského sjezdu* v Praze. Zemřel 6. dubna 1849 v Praze.

Jak bylo uvedeno výše, byl J. S. Presl jedním z propagátorů myšlenky *České společnosti*, jejímž autorem byl František Martin Pelcl a hlavním inspirátorem Josef Jungmann. Tato společnost měla pečovat o vydávání českých knih a o jejím zřízení se jednalo již v roce 1810. V souvislosti s těmito plány se Josef Jungmann a jeho přátelé a studenti zabývali od roku 1815 i myšlenkou tzv. *Malé encyklopedie*, vlastně české encyklopedie, která by vyložila celou oblast vědy v českém jazyce. Důraz měl být kladen na otázky povahy terminologické (jazykové otázky), méně pak na vědeckou koncepci. Společně s Jungmannem se

na přípravě *Malé encyklopedie* podílel například Antonín Marek (logika), Josef Kouble (přírodopis), Josef Kramář (zeměpis), Jan Jodl (historie), Antonín Jungmann (antropologie). Plán se neuskutečnil, ale stačily vyjít některé spisy, které s tímto počinem souvisely – v roce 1820 vydal Antonín Marek *Logiku neb Umnici* a Josef Jungmann svou *Slovesnost*. J. Jungman, J. S. Presl a jiní se pokoušeli přimět například nově vzniklé *Vlastenecké muzeum*, aby byl ústavem národním i svou řečí, ale protože jejich *České společnosti* se stále nedařilo získat úřední povolení, tyto snahy vyšly zpočátku naprázdno. Aby myšlenky po zvelebení českého jazyka a literatury došly alespoň nějakého naplnění, začala Jungmannova skupina vydávat samostatný vědecký časopis, který vlastně suploval plány spojené s *Malou encyklopedií*. Nový časopis dostal název *Krok* a stal se tak prvním českým vědeckým časopisem. Název „*Krok*" je v češtině homonymem a pravděpodobně byl pečlivě vybrán – jednak může označovat vojvodu Kroka, bájnou postavu české historie ze 7. století, jednak se může jednak o jednotku chůze. Časopis Krok byl totiž důležitým krokem pro rozvoj českého jazyka a literatury.

Přesně časopis nesl titul *Krok, veřejný spis všenaučný pro vzdělance národu československého* a byl vydáván „*přispěním mnoha učených vlastenců od Jana Svatopluka Presla*". Krok vycházel s přestávkami v letech 1821 až 1840. První díl vyšel ve čtyřech částkách v letech 1821 až 1823, druhý díl v letech 1824 až 1831, třetí díl v letech 1833 až 1836. Duchovním otcem časopisu byl Josef Jungmann, zásadní předmluvu napsal Jan Evangelista Purkyně. J. S. Presl redigoval časopis od jeho vzniku v roce 1821 až do roku 1837. Časopis *Krok* byl adresován především inteligenci a možná proto měl jen malý počet odběratelů – asi 150. *Krok* sdružoval především vědecké pracovníky a jeho úkolem bylo, jak jsme již řekli, vypracování základů české vědecké terminologie. Snahou vlastenců bylo totiž získat vyšší vzdělání v národním jazyce. J. E. Purkyně ve své předmluvě k časopisu *Krok* z roku 1821 píše, „*že v našem národu víc než v kterémkoli jiném sprostota a vzdělanost se dělí a taková mezera mezi oběma se nachází, že komu osud nepopřál od mládi své prostředkem cizího jazyka vyššího vzdělání sobě dobýti, ten (. . .) k samotlivé a skrovné prostotě odsouzen býti musí*".

Časopis uveřejnil mnoho příspěvků z oblasti různých věd společenských i přírodních. V otázkách jazykových, zejména lexikálních, zastával velmi průbojné stanovisko, neologizující a často až puristické. Byly zde uveřejňovány jednak původní české příspěvky (kterých bylo ale poměrně málo), jednak překlady (často Preslovy) článků z němčiny, francouzštiny a angličtiny. Co se týče věd společenských, mezi nejvýznamnější přispěvatele patřili Josef Jungmann[160], Jan Kollár[161], Pavel Josef Šafařík, František Palacký[162], Václav Hanka. Ve vědách přírodních byl nejvýznamnějším přispěvatelem sám Jan Svatopluk Presl, jeho bratr Karel Bořivoj Presl, dále Antonín Jungmann[163]. U kolébky *Kroka*

[160] Jinotajná pověst *Krok* (1821); dále články týkající se problematiky prozódie.

[161] Kritika *Česko-německo-latinského slovníku* od Jiřího Palkoviče (1823); *Myšlénky o libozvučnosti řeči vůbec, obzvláště československé* (1823), kde dokazoval, že je možné z důvodů libozvučnosti volit tvary nebo hlásková seskupení libozvučnější než ta, která předpisuje spisovná kodifikace.

[162] *Přehled dějin krásovědy a její literatury* (1823), kde řešil estetické otázky.

[163] Profesor lékařství, bratr Josefa Jungmanna.

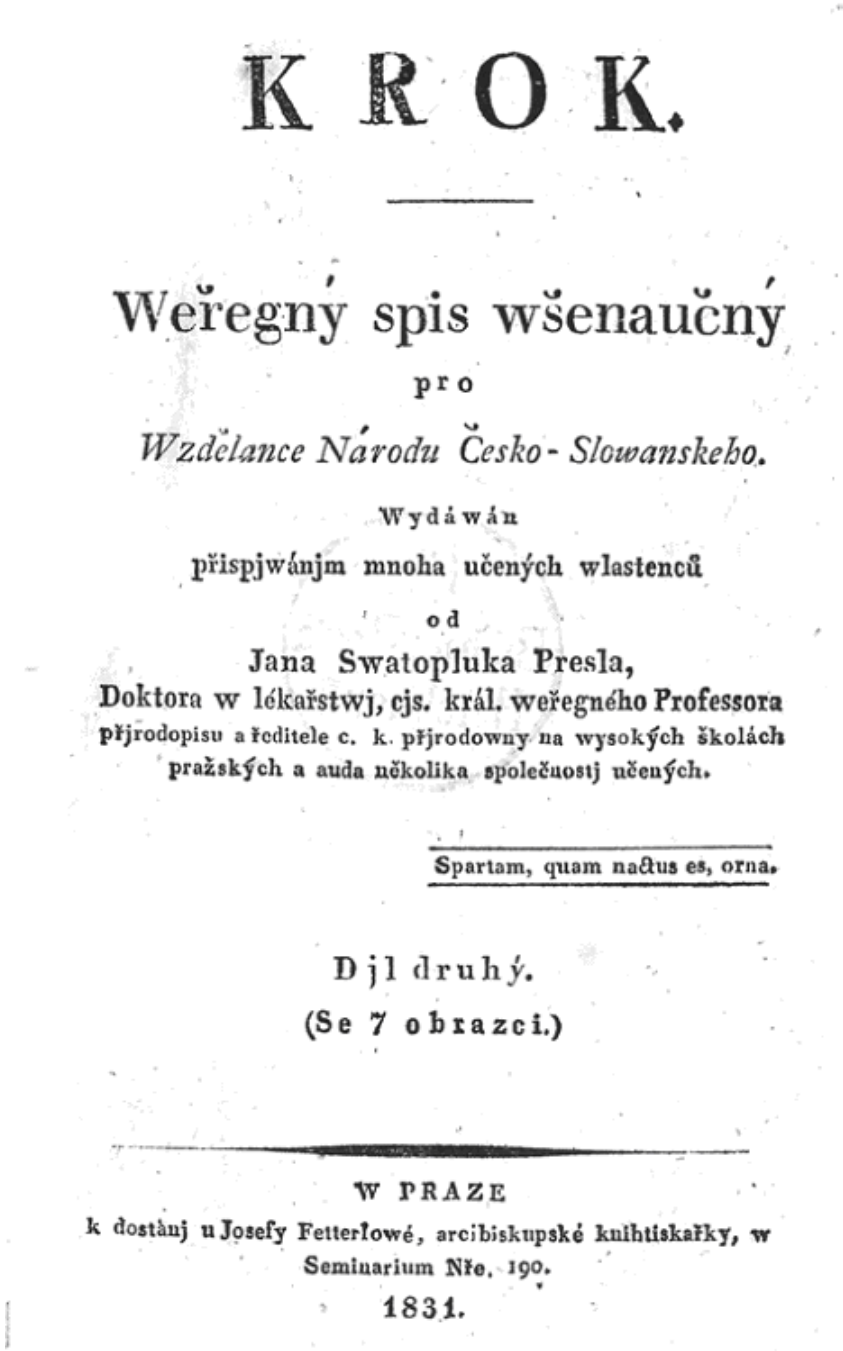
stál i Jan Evangelista Purkyně (1787–1869), vynikající fyziolog se širokými zájmy literárními i filosofickými, který byl ale v roce 1823 povolán na univerzitu do Vratislavi. Purkyně je pokládán za původce názvu *Krok*. Z dalších přispěvatelů jmenujme Josefa Chmelu[164], Josefa Jirečka[165], Josefa Myslimíra Ludvíka[166], Antonína Marka[167], Františka Sušila[168], Václava Aloise Svobodu[169], Františka Šíra[170], Karla Boleslava Štorcha[171], Bohuslava Tablice[172], Jana Slavomíra Tomíčka[173], Františka Vetešníka[174], Vincence Pavla Žáka[175].

V převážné míře časopis *Krok* redigoval Jan Svatopluk Presl (v letech 1821–1837), dále v roce 1833 rovněž Franta Josef Šumavský spolu s Janem Slavomírem Tomíčkem.

V roce 1827 vznikl konkurenční český *Časopis Společnosti vlastenského museum v Čechách* (od roku 1831 *Časopis Českého muzeum*, pod titulem *Časopis Národního muzea* vychází tento časopis dodnes). Tento časopis, který redigoval František Palacký, doprovázely zpočátku pochybnosti o jeho účelnosti, ale brzy se ukázalo, že v něm našla česká společnost reprezentativní orgán vědecký i literární. Vycházely zde například historiografické články Františka Josefa Studničky (1836–1903). Jeho upevněním pozbýval *Krok* na svém významu a roku 1840 přestal zcela existovat.

Při přípravě literární edice *Matice česká* (1831) byl J. S. Presl nucen vést rozsáhlé diskuse s vědci o tom, zda je čeština schopna fungovat dostatečně jako jazyk vědy. Pravděpodobně s těmito diskusemi souvisí také dva články otištěné v časopise *Krok* z roku 1831. Protože autor článku není uveden, můžeme se domnívat, že se jedná o samotného Jana Svatopluka Presla, redaktora časopisu *Krok*.

[164] Profesor latiny a řečtiny na akademickém gymnáziu v Praze.
[165] Bratr Hermenegilda Jirečka, úředník v ministerstvu kultu a vyučování ve Vídni, kde dosáhl vysokých hodností.
[166] Farář.
[167] Kněz – v *Kroku* vydal např. *Lovec aneb Oprava prohřešků proti duchu mluvy československé*.
[168] Církevní hodnostář, sběratel lidových písní.
[169] Profesor na gymnáziu v Písku, Jindřichově Hradci a na Malé Straně v Praze.
[170] Profesor a ředitel na gymnáziu v Jičíně.
[171] Účtárenský úředník.
[172] Farář, účastnil se založení *Ústavu řeči a literatury československé* v Bratislavě roku 1803.
[173] Novinář, soukromý učitel.
[174] Farář, děkan, jeden ze zakladatelů *Matice české*.
[175] Též Žiak, farář, profesor teologického semináře, rovněž kaplan a učitel češtiny v Brně, zpovědník italských politických vězňů na Špilberku.



KROK.

Weřegný spis wšenaučný

pro

Wzdělance Národu Česko-Slowanskeho.

Wydáwán

přispjwánjm mnoha učených wlastenců

od

Jana Swatopluka Presla,

Doktora w lékařstwj, cjs. král. weřegného Professora přjrodopisu a ředitele c. k. přjrodowny na wysokých školách pražských a auda několika společnostj učených.

Spartam, quam nactus es, orna.

Djl druhý.

(Se 7 obrazci.)

W PRAZE

k dostánj u Josefy Fetterlowé, arcibiskupské knihtiskařky, w Seminarium Nře. 190.

1831.

Obrázek 2.4: Titulní strana časopisu *Krok* z roku 1831

2.11.2 Hlásková statistika z roku 1831

Jak již bylo řečeno v úvodu, byla roku 1831 ve třetím díle časopisu *Krok* na stranách 624 a 625 publikována hlásková statistika *Srovnání spoluhlásek a samohlásek v češtině, vlaštině a němčině*. Ačkoliv se jedná o jednoduchou hláskovou statistiku, zaujímá v historii matematické lingvistiky vyjímečné postavení, které je dáno zejména dobou vzniku této statistiky, ale i zvolenou problematikou.

Statistika si klade za cíl vyvrátit tvrzení, že čeština je jazyk velmi tvrdý a těžký na výslovnost, neboť obsahuje velké množství souhlásek, jako je tomu u jazyka německého. Rovněž můžeme tuto statistiku chápat jako další příspěvek k problematice libozvučnosti (eufonii) jazyků, která byla v tomto období poměrně hojně řešena.

Autor v textu srovnává výskyt jednotlivých souhlásek a samohlásek na 1 000 hlásek v češtině, italštině a němčině (počítá tedy promile jednotlivých souhlásek a samohlásek). Je zřejmé, že statistika je doprovázena jistými metodologickými trhlinami. Nic se nedozvídáme o analyzovaném textovém materiálu (rozsah, druh textu apod.). Podíváme-li se na četnosti výskytu jednotlivých hlásek, dostáváme správný součet 1 000 pouze pro češtinu, součet pro němčinu se liší o jedničku (pravděpodobně chyba způsobená zaokrouhlováním). Špatný součet dostáváme i v případě italštiny, což může vést k dohadům, zda je statistika pro italštinu korektní.

Ještě jednou zdůrazněme, že statistika zjišťuje výskyt hlásek (tedy zvuků), ne jednotlivých písmen. To můžeme vidět na obr. 2.5, kde například pro němčinu zvuk [a] může být v textu zapsán skupinami písmen *aa*, *ah*, *au*, *ei*, *aei*, *ei*, *eu*, *ey*. Podobně pro němčinu například hláska [e] může být zapsána jako *ä*, *ee*, *oe*, *eh*. Stejně jsou řešeny zdvojené souhlásky, kdy písmennému zápisu *ll*, *mm*, *nn* odpovídá zvuk [l],[m],[n]. Rozlišuje se realizace písmene *z* jako [z] a [c], písmene *w* jako [v] a [f].

Na závěr je vyčíslen poměr samohlásek k souhláskám (tzv. *samohláskovitost jazyka*) a také poměr souhlásek k samohláskám v těchto třech jazycích. V češtině tedy na jednu samohlásku připadá 1,083 souhlásek, v italštině to je 0,953 souhlásek (v původním textu se uvádí hodnota 0,952) a v němčině jedné samohlásce odpovídá dokonce 2,067 souhlásek. Opačně pak jedné souhlásce v češtině odpovídá 0,923 samohlásek, v italštině je to 1,049 samohlásek a v němčině 0,484 (místo v textu uvedených 0,483). Na základě těchto hodnot vyplývá, že podobné poměry získáme pro češtinu a italštinu, naproti tomu němčina se od nich výrazně odlišuje. Čeština tedy neobsahuje zdaleka takové množství souhlásek jako němčina, a proto námitky, že je čeština jazyk tvrdý a těžký na výslovnost podobně jako němčina, jsou zcela liché.

Je nutno zdůraznit, že zvuková stránka jazyka nebyla běžně zpracovávána, v popředí zájmu odborníků stál vždy jazyk psaný. Přitom právě například v otázce libozvučnosti jazyka hrají statistiky založené na analýze hlásek důležitou roli. Různé hláskové statistiky vznikaly sice již zhruba od 50. let 19. století, ale ty byly sestavovány většinou k jiným účelům (např. těsnopis, šifrování, telegrafování) a ve skutečnosti vycházely z řeči psané, ne mluvené. Fonologické

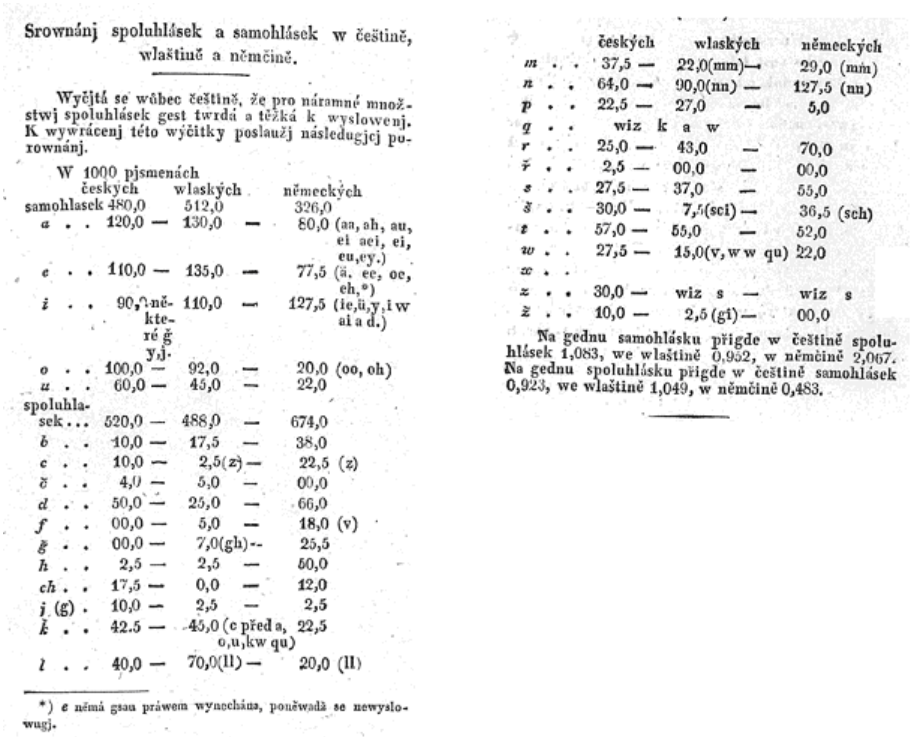
statistiky z počátku 20. století zase neanalyzovaly řeč mluvenou, ale izolovaná slova ze slovníku. Hláskových statistik sestavených z hlediska akustických vlastností jazyka u nás bylo vytvořeno málo, navíc vždy v nepatrném rozsahu či z jediného textu, takže tyto výsledky jsou pro celou soustavu českého jazyka nespolehlivé. První hláskovou statistiku češtiny na dostatečně rozsáhlém a reprezentativním materiálu[176] sestavila až v roce 1946 Věra Mazlová[177], která se vedle cíle zhodnotit libozvučnost češtiny[178] snažila o porovnání češtiny s ostatními jazyky (více viz [40]). Získané výsledky rovněž srovnávala s dostupnými hláskovými statistikami – např. hlásková statistika uvedená v časopise *Krok* z roku 1831 (její výsledky vychází více v neprospěch samohlásek – 40,7 %, *Krok* udává 48 % samohlásek), statistika určená k těsnopiseckým účelům J. Sedláčka na materiálu 88 422 hlásek, taktéž pro potřeby těsnopisu vytvořená statistika Fr. Hradilíka o rozsahu 10 000 slov aj.

I přes výše uvedené výhrady (absence informací o analyzovaném materiálu) k uvedené statistice z roku 1831 je však nutno vyzdvihnout ten fakt, že se jedná o jednu z prvních hláskových statistik češtiny, ale zejména o jednu z prvních hláskových statistik v rámci slovanské jazykovědy a možná i jazykovědy celosvětové.

[176] Jednalo se o 360 textů o 1 000 hláskách, tedy celkem 360 000 českých hlásek. Autorka texty roztřídila podle autorů a dále je rozdělila na poezii (150 000) a prózu (210 000).

[177] Poprvé již roku 1941 Mazlová na pokyn prof. Hály analyzovala 12 úryvků z literatury naučné, umělecké i poezie, z nichž každý obsahoval 1 000 hlásek (celkem tedy 12 000 českých hlásek). Protože výsledky ukázaly, že výskyt jednotlivých hlásek je v hrubých rysech stabilní a menší výkyvy jsou způsobeny druhem textu či individualitou autora, rozhodla se autorka pokračovat na rozsáhlejším materiálu, na jehož základě by bylo možno formulovat obecnější závěry o zvukové stránce češtiny.

[178] V článku [40] Mazlová cituje Jana Kollára (Kollár, J.: *Myšlenky o libozvučnosti řeči vůbec, obzvláště československé.* Krok 1823, I, 3, s. 33.), který se otázkou libozvučnosti intenzivně zabýval: „*Hmota řeči jsou spoluzvučky a samohlásky. Z obou těchto povahy, rozměru a rozmanitosti libozvučnost povstává. Spoluzvučky v řečech jsou tma, která světlem samohlásek osvícena býti může. Řeč samohlasnatá (mnoho vokálů mající) je libozvučná.*“ Pro zajímavost uveďme, že ve svých vzpomínkách (Kollár, J.: *Spisy Jana Kollára (Cestopis druhý a Paměti z mladších let života).* Díl čtvrtý, Praha 1863.) Kollár zmiňuje, že myšlenky týkající se libozvučnosti literárních děl ho zaměstnávaly již v době, kdy čekal na kandidátské zkoušky v Banské Bystrici, aby mohl započít studium na německé univerzitě v Jeně (zkoušky vykonal v únoru 1817, ke studiu v Jeně nastoupil na podzim roku 1817). Do rukou se mu tehdy od přítele Samuela Ferjenčíka dostalo vydání Komenského *Labyrintu světa a ráje srdce.* Z 1. a 2. kapitoly této knihy vypracoval statistiku souhlásek a samohlásek. Ve svých pamětech uvádí i konkrétní číselné údaje, kdy v 1. kapitole našel celkem 344 samohlásek (*a* 102; *e*, *é* 91; *ě* 34; *i*, *í*, *y*, *ý* 124; *j* (*g*) 15; *o* 76; *u*, *ů* 22; *ou* 4), souhlásek (zde používá termín „*spoluhlásek*“) nalezl celkem 568 (*b* 20; *c* 14; *č* 8; *d* 34; *h* 9; *ch* 10; *k* 33; *l* 52; *m* 58; *n* 58; *p* 32; *r* 20; *ř* 12; *s* 65; *š* 3; *t* 54; *v* 50; *z* 19; *ž* 9). Různým kombinováním, střídáním vokálů se snažil nalézt jakási pravidla vedoucí k libozvučnosti, ale nakonec konstatoval, že pouze taková mechanická cesta k libozvučnosti nevede, neboť důležitou roli zde hraje i estetický cit a sluch. I když se zdá, že se nejedná o statistiku hláskovou, ale o statistiku grafémů (tedy psaného jazyka), časově se dostáváme do období o zhruba 14 let dříve než je uvedená statistika z časopisu *Krok.* Jistě by bylo zajímavé zjistit více informací o této Kollárově statistice, popř. o dalších jeho statistických, pokud nějaké existovaly.



Srownánj spoluhlásek a samohlásek w češtině, wlaštiuě a němčině.

Wyčjtá se wůbec češtině, že pro náramné množstwj spoluhlásek gest twrdá a těžká k wyslowenj. K wywrácenj této wýčitky poslauží následugicj porownánj.

W 1000 pjsmenách	českých	wlaských	německých
samohlasek	480,0	512,0	326,0
a . .	120,0 —	130,0 —	80,0 (aa, ah, au, ei aei, ei, eu,ey.)
e . .	110,0 —	135,0 —	77,5 (ä. ee, oe, eh,*)
i . .	90, ně-které ğ y,j.	110,0 —	127,5 (ie,ü,y,i w ai a d.)
o . .	100,0 —	92,0 —	20,0 (oo, oh)
u . .	60,0 —	45,0 —	22,0
spoluhlasek ...	520,0 —	488,0 —	674,0
b . .	10,0 —	17,5 —	38,0
c . .	10,0 —	2,5(z) —	22,5 (z)
č . .	4,0 —	5,0 —	00,0
d . .	50,0 —	25,0 —	66,0
f . .	00,0 —	5,0 —	18,0 (v)
ğ . .	00,0 —	7,0(gh) —	25,5
h . .	2,5 —	2,5 —	50,0
ch . .	17,5 —	0,0 —	12,0
j (g) .	10,0 —	2,5 —	2,5
k . .	42.5 —	45,0 (c před a, o,u,kw qu)	22,5
l . .	40,0 —	70,0(ll) —	20,0 (ll)

*) e němá gsau právem wynechána, poněwadž se newyslowugj.

	českých	wlaských	německých
m . .	37,5 —	22,0(mm) —	29,0 (mm)
n . .	64,0 —	90,0(nn) —	127,5 (nn)
p . .	22,5 —	27,0 —	5,0
q . .	wiz k a w		
r . .	25,0 —	43,0 —	70,0
ř . .	2,5 —	00,0 —	00,0
s . .	27,5 —	37,0 —	55,0
š . .	30,0 —	7,5(sci) —	36,5 (sch)
t . .	57,0 —	55,0 —	52,0
w . .	27,5 —	15,0(v, w w qu)	22,0
x . .			
z . .	30,0 —	wiz s —	wiz s
ž . .	10,0 —	2,5 (gi) —	00,0

Na gednu samohlásku přigde w češtině spoluhlásek 1,083, we wlaštině 0,952, w němčině 2,067. Na gednu spoluhlásku přigde w češtině samohlásek 0,923, we wlaštině 1,049, w němčině 0,483.

Obrázek 2.5: Fotokopie hláskové statistiky v časopise *Krok* (1831)

V témže časopise se nachází i druhá statistika *Český jazyk s jinými jazyky z ohledu bohatosti slov porovnaný* ([52]), o které se Grzybek již nezmiňuje. Tato statistika se týká srovnání jazyků z hlediska jejich slovní zásoby a nachází se na straně 623. Statistika se snaží vyvrátit námitky, že čeština má velmi chudou slovní zásobu na to, aby mohla fungovat jako jazyk vědy. V krátkém přehledu autor (nejspíše J. S. Presl) srovnává počet slov v němčině, španělštině, francouzštině, italštině, angličtině a češtině (viz tab. 2.15).

Jazyk	Počet slov
Němčina	29.000
Španělština	30.000
Francouzština	32.000
Italština	35.000
Angličtina	37.000
Čeština	57.000

Tabulka 2.15: Srovnání počtu slov v různých jazycích

Vedle toho je v článku uvedeno ještě zastoupení některých slovních druhů v těchto jazycích s výjimkou španělštiny – viz tab. 2.16.

Sl. Druh	němčina	francouzština	italština	angličtina	čeština
podst. jména	21 395	19 706	13 030	16 127	21 276
slovesa	2 655	5 657	7 877	10 141	13 184
příd. jména	2 475	4 803	5 843	8 444	14 184
příslovce aj.	2 475	1 834	8 250	2 888	14 356
celkem	29 000	32 000	35 000	37 000	57 000

Tabulka 2.16: Srovnání výskytu jednotlivých slovních druhů v různých jazycích

Slovní zásoba italštiny, francouzštiny a angličtiny byla získána podle autora statistiky z nejlepších slovníků těchto jazyků (není ovšem uvedeno, o které konkrétní slovníky se jedná). Slovní zásoba češtiny byla určena na základě slovníku J. Palkoviče, který nebyl ještě zcela dokončen, proto by počet slov mohl být zcela jistě daleko vyšší, budou-li přidána slova knižní a slova běžné mluvy. Přidáme-li navíc k takto vyčíslené slovní zásobě češtiny ještě českou terminologii těch oborů, pro něž dosud není vytvořena, množství slov českého jazyka se dále zvýší.

Z těchto tabulek je patrné, že slovní zásoba českého jazyka není vůbec malá (i přesto, že zde některá slova nejsou zahrnuta), jak je jí vytýkáno, ba právě naopak. Co se týče množství slov, předčí všechny uvedené jazyky. Dokonce, i pokud provedeme rozdělení slovní zásoby na jednotlivé slovní druhy, čeština ostatní jazyky převyšuje. Jedinou výjimkou jsou podstatná jména, kterých němčina vykazuje nepatrně vyšší počet. To je ale způsobeno pro němčinu tak typickým skládáním. Podíváme-li se na německá podstatná jména blíže, vidíme, že jednoduchých podstatných jmen je pouze 7 847 a složených je 13 243.

V této kapitole byly představeny dvě prakticky neznámé statistiky jazyka (hlásková a lexikální)[179], které vznikly daleko dříve než v rámci matematické lingvistiky běžně uváděné kvantitativní statistiky jazyka. Určitě by bylo přínosné hledat další takové příklady využití matematických metod v lingvistice, abychom si mohli vytvořit ucelenější přehled o prehistorii disciplíny dnes označované jako matematická lingvistika.

2.12 Augustin Seydler a rukopisný spor

Do historie matematické, přesněji kvantitativní lingvistiky, zasáhl i český astronom, fyzik a matematik Augustin Seydler svou prací z roku 1886 *Počet pravděpodobnosti v přítomném sporu*, která byla opublikovaná v časopise *Athenaeum*. Svým článkem se tak na popud přítele T. G. Masaryka zapojil do bojů o pravost *Rukopisů*. Pro nás je ovšem tato práce význačná tím, že zde vůbec poprvé v českém prostředí použil matematiky (konkrétně počtu pravdě-

[179]Rovněž jsme se zde zmínili o existenci statistiky (nejspíše grafémů), kterou okolo roku 1817 provedl Jan Kollár.

podobnosti) při řešení problému jazykové povahy. A nejspíše se jedná i o jednu z prvních (ne-li první) aplikací matematiky v lingvistice.

2.12.1 Historie rukopisných sporů

Od roku 1816 dochází k nálezům řady údajně staročeských textů, které významně měnily pohled na starou českou literaturu i kulturu. Nejprve v roce 1816 objevil Josef Linda lyrickou *Píseň pod Vyšehradem*, kladenou do 13. století. Ve věžní kobce děkanského kostela sv. Jana Křtitele ve Dvoře Králové nachází 16. září 1817 bibliotékář *Musea království Českého* Václav Hanka zlomek rozsáhlého kodexu (šest rozsáhlých epických zpěvů, dvě lyrickoepické básně, šest písní), který podle místa nálezu nazýváme *Rukopis královédvorský* (RK). I tento nález byl kladen jazykově do 13. století, ale obsahoval jak texty z období křesťanského, tak z pohanské minulosti. Nález je tvořen sedmi pergamenovými dvoulisty popsanými po obou stranách, z nichž dva listy jsou neúplné, ze tří čtvrtin odříznuté. Text je tedy umístěn na 24 celých stranách a 4 proužcích, má 783 celých řádků a 132 neúplných řádků na proužcích. V listopadu 1818 je anonymně zaslán nejvyššímu purkrabímu království Českého (tj. nejvyššímu představiteli zemské správy) hraběti Kolovratovi pro nově zřízené *Museum království Českého* zlomek rukopisu, který našel roku 1817 (jak se zjistilo až později) zámecký úředník Josef Kovář na Zelené Hoře. Rukopis je tvořen dvěma pergamenovými dvoulisty s osmi stránkami textu, celý rukopis má 129 řádků. Zlomek obsahoval část rozsáhlé skladby *Libušin soud*, projednávající spor mezi bratry Chrudošem a Šťáhlavem o dědictví po otci, proto byl zpočátku pro svůj obsah nazýván *Libušin soud*. Teprve 1859 bylo zjištěno místo nálezu a od té doby je znám pod názvem *Rukopis zelenohorský* (RZ). Časově byl kladen dokonce až do 9. nebo 10. století. V roce 1819 objevil skriptor univerzitní knihovny v Praze Jan Zimmermann *Milostnou píseň krále Václava*, roku 1827 bylo v Muzeu za přítomnosti Václava Hanky cizím badatelem objeveno množství českých glos v známém slovníku *Mater verborum* a v roce 1828 se nalezl v muzejním *Glosovaném žaltáři* meziřádkový překlad *evangelia sv. Jana* (žalm 109 a část žalmu 145). A konečně v roce 1849 je nalezeno *Proroctví Libušino*, které je kladeno do 14. století.

Největší ohlas u nás i ve světě vzbudily zejména *Rukopisy královédvorský a zelenohorský* (RKZ), zvláště když na RKZ upozornil evropské vzdělance J. W. Goethe. Tento zájem byl navíc podporovaný obrozeneckým kultem minulosti. Vždyť tyto texty dokládaly starobylost české literární tradice a posouvaly ji na úroveň takových kultur, jako byla například kultura francouzská, německá, španělská a ruská. Rukopisy se staly inspirací pro celou řadu umělců (B. Smetana, J. V. Myslbek, K. H. Mácha aj.) a byly hojně překládány do cizích jazyků. A právě o *Rukopisu královédvorském* a *Rukopisu zelenohorském* bude pojednáno v této kapitole.

Zpočátku obrozenci nepochybovali o pravosti obou památek, záhy se však začaly ozývat hlasy opačné. Spory o pravost těchto památek se vlekly celým devatenáctým stoletím a objevovaly se i ve století dvacátém. Ačkoliv ve společnosti stále existuje názor, že spor o *Rukopisy* není uzavřen, a část lidí je

dokonce považuje za pravé, pro vědce již není tato otázka aktuální. *Rukopisy* jsou chápány jako díla obrozenecká a centrem pozornosti se stává spíše to, co svou existencí vyvolaly dále. Vzhledem k rozsáhlé rukopisné literatuře se stala tato problematika pro odborníky prakticky nepřehledná. Jakousi rekapitulaci bádání o RKZ v oblasti společenských věd přináší sborník *RKZ – Dnešní stav poznání*, který byl iniciován při příležitosti 150. výročí objevení Rukopisů a který obsahuje dosud nejpodrobnější tiskem vydaný soupis odborné literatury týkající se RKZ, jehož autorem je Miroslav Laiske. Tento seznam zahrnující monografie a nejvýznamnější odborné články je tvořen 1078 položkami. Vzhledem k tomu, že tento soupis vyšel tiskem již v roce 1969, bude v současnosti toto číslo zcela jistě mnohem vyšší. Otázku pravosti považujeme tedy za vyřešenou, otázka autorství ale vyřešena není a nejspíše ani nikdy definitivně vyřešena nebude. Je téměř jisté, že hlavní podíl můžeme připisovat Václavu Hankovi a svůj podíl na padělcích má rovněž Josef Linda. V některých studiích se pak v souvislosti s paděláním RKZ objevuje jméno Aloise V. Svobody, Josefa Jungmanna či Antonína Marka. Bohužel neexistují žádné přímé důkazy, které by tuto otázku zodpověděly.

V počátečním období zažívají *Rukopisy* vrcholnou slávu, dochází k jejich vydávání, překladům i k rozsáhlé vědecké činnosti s doklady o jejich pravosti. I největší odborné autority, jako byli historik František Palacký a slavista Pavel Josef Šafařík, se vyjadřují o RKZ bez pochybností. Pochybovači jsou považováni za nepřátele českého národa. První nejistoty o pravosti těchto *Rukopisů* se objevují již roku 1824, kdy Josef Dobrovský ve svém článku *Literarischer Betrug* (*Literární podvod*) nazývá RZ „*zřejmě podvrženou mazanicí*". Rovněž Dobrovský v roce 1828 poznal nepravost zlomku z *Evangelia sv. Jana*. Hlasy o nepravosti *Rukopisů* se začínají ozývat i ze zahraničí (např. slovinský lingvista Bartoloměj Kopitar). Od roku 1856 vyjadřuje různé pochybnosti o RK a vyzývá k důkladné kritice Julius Feifalik, dále například Max Büdinger (1859). Ve prospěch *Rukopisů* se na české straně v té době vyjadřovali vedle již zmíněného Františka Palackého a Pavla Josefa Šafaříka, rovněž například Václav Bolemír Nebeský či Václav Vladivoj Tomek. Šafaříkova naznačená pochybnost o RZ týkající se dědického práva u starých Slovanů v jedné z jeho posledních rozprav nebyla českou kritikou rozváděna. Zdá se, že Šafaříkovo podezření proti RKZ bylo na sklonku jeho života mnohem větší, než naznačují jeho publikované práce, zato Palacký nikdy nezakolísal. Roku 1858 je inspirován pražskou policií veřejný útok proti RKZ (pražský policejní ředitel Päumann byl nepřítelem českého hnutí). Po prvním zamítnutí Vídní se mu podařilo v pražských *Tagesbote aus Böhmen* vydávaných pod redakcí Kuhy umístit anonymní článek *Literärische Lügen und paleographische Wahrheiten*, v němž byl RK prohlášen za falzum a Václav Hanka za jeho autora. Na obranu *Rukopisů* vystoupil v *Bohemii* František Palacký (1858). Nezůstalo ale jen u polemiky v tisku. V. Hanka byl představiteli českého hnutí přinucen k podání žaloby na odpovědného redaktora Davida Kuhu za urážku na cti, kterou se anonym v článku provinil. Kuha byl odsouzen k trestu vězení, po odvolání byl rozsudek potvrzen. Přesto si dvouměsíční trest nakonec neodpykal díky zákonu, který byl zanedlouho vydán (ale je rovněž možné, že Kuha pohrozil tím, že zveřejní pravého autora).

"kniha˙final" — 2012/7/27 — 22:04 — page 120 — #126

Tato událost se stala pro české kruhy potvrzením toho, že k útoku na *Rukopisy* nedošlo z vědeckých kruhů, ale s protičeským záměrem. A jenom to posílilo vážnost k nim. Proto i další námitky proti pravosti *Rukopisů* jsou chápány s protičeským cílem.

V období následujícím (1859–1886) se ale začínají objevovat první nepopiratelní vlastenci, kteří pochybují o pravosti *Rukopisů*. Byli to v podstatě první průkopníci, následujícími vědci nedocenění, ačkoliv tito pozdější vědci často využívali jejich výsledky. Patřil mezi ně Alois Vojtěch Šembera (1878) a středoškolský profesor Antonín Vašek (1879)[180]. Ve prospěch *Rukopisů* vystupuje tehdy ještě Jan Gebauer (zvláště 1870, od 1874 i na stránkách nově založeného časopisu *Listy filologické a paedagogické*). Ještě v roce 1879 polemizuje Gebauer s A. Vaškem, v roce 1881 s Vaškem i Šemberou, ale již o čtyři roky později se přidává na stranu odpůrců RKZ. Vysoko lze hodnotit studii A. Vaška, ve které byla výstižně formulována či naznačena velká část námitek, které pak nacházíme později u Gebauera. Mimo jiné tu byla v hlavních bodech naznačena teorie *koincidencí* s Hankou, která se pak stala jednou z hlavních částí důkazu Gebauera. Nejspíše příprava námitek proti Vaškovi se stala rozhodujícím momentem pro to, aby se Gebauer přidal na stranu odpůrců. V roce 1877 Adolf Patera a Antonín Baum v *Muzejníku*, oficiálním orgánu české vědy, publikují odhalení padělaných glos v *Mater verborum*. Stále častěji se ozývají i hlasy proti ze zahraničí (ruští učenci Petruševič a Lamanskij, chorvatský filolog Jagi aj.). Osmdesátá léta 19. století můžeme tedy charakterizovat jako jakési tušení nejasností, které si žádá nutnost vyřešení celého problému.

Za významný mezník v historii rukopisných bojů pokládáme rok 1886, kdy došlo k obnovení sporů o pravost RKZ, a to v Masarykově vědecké revui *Athenaeum* (1883–1893), později *Naše doba* (1893–1949). Tento spor byl vyvolán článkem Jana Gebauera *Potřeba dalších zkoušek rukopisů Královédvorského a Zelenohorského* ([15]), v němž právě na podnět T. G. Masaryka, který příspěvek podpořil vlastní statí, provedl podrobný filologický rozbor obou textů. Postupně bylo v *Athenaeu* uveřejněno množství vědeckých studií, polemických recenzí i glos. Jen v roce 1886, kdy započaly rukopisné boje, bylo v *Athenaeu* uvedeno 51 článků týkajících se RKZ (z celkem 113 uveřejněných článků v tomto roce), v roce 1887 to bylo v *Athenaeu* 11 prací (z celkově 28 publikovaných). Články pochybující o pravosti *Rukopisů* ihned vyvolaly ostrý spor mezi jejich obhájci a odpůrci. V tisku (hlavně Grégrovy *Národní listy* a *Osvěta*) se rozpoutala celospolečenská štvavá kampaň proti odpůrcům *Rukopisů*, zejména však proti T. G. Masarykovi a Janu Gebauerovi. Tyto útoky vedl především Julius Grégr za podpory M. Hattaly, ale i desítek středoškolských profesorů z venkova. Gebauer a Masaryk byli obviňováni z národní zrady. Sociologické a estetické rozbory RKZ vynesly Masarykovi u Grégra titul „*filosoficko-kulturně-aesthetický Paracelsus Tomáš Garrigue – Theofrastus Bombastus z Masaryků*". Tyto spory dokonce výrazně negativně ovlivnily Masarykovu kariéru, neboť o několik let odsunuly jeho jmenování řádným profesorem na pražské univerzitě. Aby nepoškodil projekt, který sám vytvořil, vzdal se rovněž redigování

[180] Vašek provedl filologický důkaz, že RKZ a zlomek *Evangelia sv. Jana* jsou podvržená díla V. Hanky.

"kniha˙final" — 2012/7/27 — 22:04 — page 121 — #127

Ottova slovníku naučného. Protože se rukopisný spor odehrával především na stránkách časopisu *Athenaeum*, podlehl negativnímu tlaku i jeho vydavatel J. Otto, který časopis odmítl dál vydávat, a proto jej začal vydávat vlastním nákladem sám Masaryk. Na stranu T. G. Masaryka a Jana Gebauera se však přidala řada vědců, kteří v rámci svých oborů prokázali nepravost *Rukopisů*. Vedle Masaryka, který provedl estetický rozbor obou *Rukopisů* a sociologický rozbor RZ, to byl historik Jaroslav Goll, estetik Otakar Hostinský, fyzik a astronom Augustin Seydler, chemickou analýzu provedli univerzitní profesoři Antonín Bělohoubek a Vojtěch Šafařík (syn P. J. Šafaříka), literárněhistorický rozbor provedli Jaroslav Vlček, Josef Truhlář, Jindřich Vančura či Arnošt Kraus, na poli filologie pracoval Josef Král, analýzu osobních jmen v obou *Rukopisech* provedl Jiří Polívka. Mezi obránce patřil historik Josef Kalousek, který upozornil na pochybnost *koincidencí*, jimiž se na základě Gebauerových textů zabýval Seydler. K dalším obráncům patřili například profesor slovanské filologie Martin Hattala, bratři Josef a Hermenegild Jirečkové, Vincenc Brandl, František Xaver Prusík, Adolf Patera (který ačkoliv prokázal padělanost českých glos v *Mater Verborum*, v pravost RKZ věřil) aj. Nejdůležitější roli sehrály v tomto období důkazy jazykové. V 70. a 80. letech 19. století postupovala jazykověda tak, že se pokoušela hledat odchylky jazyka RKZ od jazyka období, k němuž se hlásí. Nevýhodou bylo, že se kladl důraz na kvantitu důkazů, které se pak stávaly nepřehledné, místo rozlišení důkazů podstatných od vedlejších.

Na základě rukopisných sporů přeřadilo Gautschovo ministerstvo vyučování RKZ z literatury středověké do literatury národního obrození. Počátkem devadesátých let 19. století se tedy začalo ukazovat, že správnější názory jsou na straně odpůrců a toto mínění nezměnila ani příhoda s Václavem Flajšhansem, jedním z nejtalentovanějších Gebauerových žáků. Ten totiž roku 1896 z rozhodného odpůrce RKZ přešel na stranu obránců. Pravděpodobně se tak stalo kvůli osobním sporům s Gebauerem a sám Flajšhans svůj nezdar později přiznal. V roce 1899 oznámil Ladislav Dolanský v *Listech filologických* objevení kryptogramu „*V. Hanka fecit*“ v RZ, ale roku 1911 se zjistilo, že jde o klam. V tomtéž roce bylo vydáno prohlášení 52 realistických učenců, tzv. *silvestrovský manifest*, o tom, že rukopisný spor je uzavřen. Rukopisné spory mají v tomto období i jednu tragickou oběť. Když archeolog J. L. Píč, obránce RK, podnikl s dvěma jejich listy okružní cestu po významných francouzských a italských paleografech, vyložil si jejich zdrženlivé a rezervované vyjádření k tomuto rukopisu jako jasné vítězství RK. Po svém návratu 17. 12. 1911 Píč publikoval v *Národní politice* článek *RK před mezinárodním soudem paleografů skvěle obstál*, ale malý zájem vedení *Muzea* a hlavně posměšný článek odpůrců podepsaný „*septimán*“ jej rozrušily natolik, že spáchal sebevraždu zastřelením.

V období po roce 1918 se všichni odborníci společenskovědních oborů (filologie, literární historie, historie, dějiny státu a práva, estetika, sociologie, diplomatika, paleografie) shodli na tom, že RKZ jsou novodobé literární padělky a že se jedná o problematiku vědecky vyřešenou. Odlišně se to jevilo u přírodních věd, neboť chemické zkoumání vyznělo ve prospěch *Rukopisů*. Což ovšem nemůže popřít zjištění společenskovědních oborů, k jejichž výsledkům se často přistupuje skepticky (na rozdíl od výsledků věd exaktních). O tato

chemická zjištění se opřela poslední skupina veřejných obránců RKZ po 1918. Největší autoritou byl profesor fyziologie na lékařské fakultě Karlovy university František Mareš. Obrana Mareše a jeho stoupenců měla ale výrazně politický ráz a postupně Mareš přilnul k fašistické ideologii. Obrana našla aktivní organizační základnu počátkem třicátých let 20. století v skupině mladých fašistů *Vlajka*. Nejaktivnějším obráncem *Rukopisů* se stal její funkcionář Jan Vrzalík. Hlavní činností bylo pořádání veřejných přednášek v řadě měst, které ale byly navštěvovány zejména fašisty a jejich sympatizanty. Středoškolští studenti měli školským úřadem přednášky zakázány. *Vlajka* rozvíjela i rozsáhlou publicistickou činnost o RKZ ve svých časopisech, vydávala samostatné spisy a letáky, obsahující často naivní tvrzení, nesmysly, ale hlavně reakční nacionálně politickou ideologii. Pravděpodobně kvůli zvýšení zájmu o své hnutí Vrzalík kritiky řešil podáváním tiskových žalob na jejich autory. Podporu se pokoušel Vrzalík získat i v zahraničí, zejména Polsku.

Na obranu *Rukopisů* byla ustanovena roku 1932 v Praze zcela nepolitická společnost s názvem *Československá společnost rukopisná*. Podle stanov této společnosti, které vzalo na vědomí ministerstvo vnitra v Praze dne 26. října 1932, bylo hlavním cílem „*šíření a popularizování vědeckého bádání o starodávných rukopisech kulturních vůbec, zvláště pak o rukopisech slovanských, zejména českých*". Ve skutečnosti ale šlo jen o RKZ. Ačkoliv chtěla společnost působit celostátně, omezovala se v podstatě na Prahu a ojedinělé členy na venkově. Protože společnost viděla oprávněnost obhajoby v oblasti exaktních věd, stal se jejím předsedou roku 1935 profesor Viktorin Vojtěch, odborník na vědeckou fotografii a chemii[181]. Aktivním členem Společnosti se stal i Vrzalík z *Vlajky*. Od 15. června 1935 začala společnost vydávat své *Zprávy*, které se ale v souvislosti s působením členů *Vlajky* postupně zpolitizovaly. Podstatné je ovšem to, že všechna vzdělávací zařízení státu považovala jednoznačně otázku RKZ za vědecky vyřešenou a podvrženost za dokázanou. Rukopisná společnost pokračovala v činnosti i za okupace a své *Zprávy* vydávala až do 15. března 1941. Poslední valná hromada se konala 24. března 1941. Poúnorový režim společnosti nepřál, a proto byla její činnost na poslední schůzi likvidačního výboru 9. prosince 1952 ukončena. V té době měla tato společnost 311 členů.

21. března 1993 byla činnost *České společnosti rukopisné* znovu obnovena a začaly vycházet práce zejména samizdatových badatelů – např. J. Enderse (viz [13]), který zdůrazňuje nutnost komplexního jazykového rozboru RKZ oproti pouhému zkoumání koincidencí.

2.12.2 Život a dílo Augustina Seydlera

„Ten, co teď mluví, je také básník?"
„I ba ne! Matematik, astronom."
„Zaplať Pánbůh! Těch, kteří hvězdy opěvují,
z nich perly na břehu „mlékové řeky" činí

[181] Vojtěch se na podnět *České akademie* roku 1913 podílel na vydání přesného fotografického faksimilovaného vydání RK k 100. výročí jeho objevení. Vojtěch dospěl k závěru, že tyto památky nemohly být vyrobeny v době svého objevení a že musí být starší.

a končiny nebes květinovými pletenci spojují,
máme už tak skoro dost;
ale těch, kteří to nebe měří a váží,
hvězdy počítají, je sledují a zaznamenávají,
po čertech málo!" (Jan Neruda)



Obrázek 2.6: Augustin Seydler

Augustin Seydler (celým jménem August Jan Bedřich Seydler), významný matematik, fyzik a zejména astronom, příslušník staré pražské rodiny, se narodil 1. 6. 1849 v Žamberku manželům[182] Janu Nepomuku Seydlerovi[183] a Antonii Sukové[184] jako čtvrté dítě – sourozenci Anna[185], Marie[186], Karel[187]. Otec Jan Nepomuk Seydler v Žamberku od roku 1840 působil jako člen finanční stráže (tzv. *Spojená komorní správa důchodková v Království českém*). V květnu roku 1850 se rodina přestěhovala do Prahy, kde však 27. 10. 1852 matka Augustina Seydlera umírá na tuberkulózu. Otec se 12. 10. 1854 znovu oženil s vdovou Annou Hromátkovou z Heřmanova Městce (ovdovělou Jandoušovou[188]) a brzy se jim narodila dcera Antonie Anežka[189]. Otci se na poli profesním začíná dařit, do výslužby odchází roku 1882 s hodností c. k. vrchního finančního rady, umírá roku 1888 (o tři roky po své druhé ženě). Ke zbytku rodiny nebyl ale

[182]Sňatek byl uzavřen dne 13. 9. 1842.
[183]Narozen 12. 10. 1810.
[184]Narozena roku 1820.
[185]Narozena 17. 6. 1843.
[186]Narozena 30. 3. 1845.
[187]Narozen 14. 6. 1847.
[188]Narozena 1810.
[189]Narozena 15. 3. 1855.

osud takto štědrý, neboť nejspíše tuberkulóza předčasně ukončila životy všech Augustových sourozenců[190].

Augustin navštěvoval v letech 1860–1867 piaristické gymnázium na Novém Městě Pražském a už v této době projevoval zájem o matematiku a zejména astronomii. Po složení maturity s vyznamenáním[191] v říjnu 1867 nastoupil na filosofickou fakultu Karlo-Ferdinandovy univerzity, kde si zapsal matematiku a astronomii, dále fyziku, chemii a filosofii. Brzy zaujal profesory Macha (učitel fyziky kladoucí důraz na její experimentální stránku) a Hornsteina (profesor matematiky se zálibou v astronomii), kteří měli velký vliv na jeho následující kariéru. Oba Seydlerovi nabídli asistentská místa a on jejich nabídku přijal. Ve školním roce 1868/69 začal pracovat v laboratoři c. k. fyzikálního ústavu (E. Mach mu dokonce vymohl na dobu dvou let stipendium 100 zlatých) a od školního roku 1869/70 se stal též Hornsteinovým asistentem na klementinské hvězdárně. Zde se August Seydler brzy osvědčil a za dva roky se stal adjunktem a tuto funkci vykonával až do svého odchodu roku 1881. Vzhledem k mizivým možnostem praktické astronomie v Klementinu a pod vlivem Hornsteina se Seydler věnoval spíše astronomii teoretické a v tomto období vytvořil většinu svých prací z nebeské mechaniky, které ho zařadily k takovým osobnostem, jako byl Laplace, Lagrange či Gauss.

6. prosince 1871 získal Seydler titul doktora filosofie, ale již 30. ledna 1872 předkládá svoji habilitační práci z fyziky[192], na jejímž základě jej c. k. ministerstvo kultu a vyučování schválilo 14. 6. 1872 soukromým docentem teoretické fyziky na Karlo-Ferdinandově univerzitě v Praze a umožnilo mu tak ve 23 letech přednášet na univerzitě. Vedle studia se Seydler zajímal o filozofii a měl i značné znalosti jazykové (kromě němčiny, která byla v tehdejší době samozřejmostí, v roce 1872 vykonal na technice zvláštní zkoušky z francouzštiny, italštiny a angličtiny), které mu umožnily poznat řadu vědeckých děl v originále (např. díla Newtona, Huygense, Laplace, Lagrange). Roku 1878 byl jmenován členem zkušební komise gymnaziální v sekci české (byl nástupcem profesora Ignáce Axamita) a examinátorem fyziky zůstal i po roce 1880, kdy byla zřízena samostatná česká zkušební komise gymnaziální, i po roce 1884, kdy se tato komise spojila s českou zkušební komisí reální v jednotnou zkušební komisi pro učitelství na gymnáziích a školách reálných.

Brzy se dokázal na univerzitě úspěšně prosadit vedle profesora Macha, světově uznávané kapacity ve fyzice. Byl oceňován za svůj propracovaný, jasný výklad a rovněž za to, že přednášel česky[193]. Zdárné působení Seydlera jakožto soukromého docenta fyziky na pražské univerzitě vyvrcholilo 27. 7. 1881

[190] Roku 1859 umírá nevlastní sestra Antonie v pouhých 4 letech, 2. 11. 1867 ve 22 letech umírá Marie, 14. 8. 1883 v 36 letech umírá Karel (doktor filozofie a profesor na c. k. vyšší reálné škole), 18. 1. 1886 v 46 letech sestra Anna (představená kláštera).

[191] Dne 19. 7. 1867.

[192] Dne 29. 2. 1872 byla přijata profesorským sborem, 24. 4. 1872 vykonal předepsané kolokvium a 4. 5. 1872 přednesl habilitační přednášku na téma *O různých způsobech odvození druhé hlavní věty mechanické teorie tepla.*

[193] Vedle Seydlera přednášel česky v tehdejší době na univerzitě pouze profesor F. J. Studnička a vůbec první česká přednáška z oboru exaktních věd zazněla roku 1861 z úst Rudolfa Skuherského z deskriptivní geometrie.

jeho jmenováním mimořádným profesorem matematické (resp. teoretické) fyziky. Po rozdělení Karlo-Ferdinandovy univerzity na část německou a českou[194] přechází na českou univerzitu. Sbor české filosofické fakulty se skládal ze 14 řádných a 6 mimořádných profesorů a byl ustanoven na jaře roku 1882. Děkanem se stal František Josef Studnička. Ačkoliv měl být majetek univerzity rozdělen spravedlivě na poloviny, podstatnou část sbírek apod. získali Němci. Tak připadla německé části univerzity i hvězdárna v Klementinu, jejíž vedení přebral po Karlu Hornsteinovi Ladislaus Weinek. Seydler nevyužil možnosti nastoupit na uvolněné místo ředitele klementinské hvězdárny po Hornsteinovi a rozhodl se působit na rodící se české univerzitě, které věnoval značné množství svých sil. Stihl zde vykonat dva záslužné počiny – založit tradici systematické výuky teoretické fyziky a vybudovat český astronomický ústav.

Protože přírodní vědy jako matematika, fyzika, chemie apod. měly své vlastní katedry, jeho zájem se soustředil na rozvoj astronomie a teoretické fyziky a založení jejich samostatné katedry na české univerzitě. Ještě roku 1882 podává žádost o jmenování řádným profesorem teoretické fyziky a astronomie na c. k. ministerstvo kultury a vyučování, které by umožnilo založení této katedry. Ministerstvo se souhlasem váhalo, neboť profesura z obou těchto disciplín současně byla značně neobvyklá. Nakonec byl 1. 5. 1885 Augustin Seydler jmenován řádným profesorem teoretické astronomie a teoretické (matematické) fyziky na filosofické fakultě c. k. české univerzity Karlo-Ferdinandovy. Protože si ale již brzy začal uvědomovat náročnost tohoto spojení a sám se hodlal věnovat jen astronomii, začal usilovat o oddělení obou disciplín. K tomu došlo až po jeho smrti, kdy se řádným profesorem fyziky na české univerzitě stal profesor Koláček z Brna.

Praktická výuka astronomie na české univerzitě byla z již výše uvedených důvodů velmi omezená (německá univerzita měla k dispozici klementinskou hvězdárnu, na univerzitě české kupoval Seydler některé pomůcky a knihy, které by umožňovaly alespoň teoretickou práci, z vlastních prostředků). Seydler proto usiloval od roku 1886 o založení univerzitní hvězdárny. V lednu při pražské návštěvě ministra barona P. Gautsche, „*nejvyššího správce záležitostí duchovních a vyučování*", mu předložil Seydler zvláštní memorandum o zřízení astronomického ústavu, ještě v únoru téhož roku podal prostřednictvím profesorského sboru podrobnější návrhy, v nichž chtěl observatoř umístit do horní části Bubenečské obory. Počátkem roku 1887 nechal na vlastní náklady zpracovat plány a předložil je ke schválení. V dubnu téhož roku byl vyzván, aby si pro své potřeby vybral ze zrušeného c. k. dvorního astronomického a fyzikálního kabinetu ve Vídni přístroje. V červnu téhož roku mu byly přístroje zaslány do Prahy a Seydler je dočasně umístil ve svém bytě. V červnu roku 1888 získává dotaci 2 000 zlatých, ale již tehdy bylo jasné, že bude definitivně upuštěno od výstavby univerzitní hvězdárny (v tuto dobu totiž usiloval o zřízení hvězdárny v Praze i profesor Weinek a vláda poukazovala na nemožnost zřídit dvě hvězdárny, proto nakonec nezřizuje ani jednu). Seydler si tedy pronajímá v Ovenecké ulici č. 80 na Letné vilu Jana Kindla, kde zřídil provizorní observatoř (byl zde byt pro ředitele, místnost pro personál, knihovna, místnost pro instrumentální sbírky,

[194] Dne 28. 2. 1882.

učebny), kterou nazývá astronomickým ústavem. 4. 2. 1889 byl nájem vily 1 250 zlatých ročně schválen, zaplacen na pět let dopředu a byla poskytnuta dotace 2 700 zlatých na stavbu astronomického pavilonu. I toto provizorium ale Seydlera těšilo. 14. 9. 1889 získal ústav mimořádnou dotaci 10 000 zlatých[195] na dobu dvou let (1890–1891) na vybavení ústavu a byl mu přiznán roční příjem 800 zlatých. 29. září 1890 se pak na Letné začalo s výstavbou pavilonu podle Seydlerova návrhu. Vybavení pro tuto observatoř i celý ústav Seydler získal na počátku roku 1890 po pastoru Brödelovi. Požadovaná cena převyšovala dotaci určenou pro daný rok, proto Seydler dluh ručil svými vlastními státními papíry. Instalace se už bohužel nedočkal, neboť 22. června 1891 umírá. V době nemoci Seydlera a jistý čas po jeho smrti je práce ústavu ochromena. Prozatímním ředitelem je jmenován Seydlerův přítel Čeněk Strouhal, ale z důvodů jeho zaneprázdněnosti se hledá definitivní nástupce, kterým je roku 1892 jmenován Gustav Gruss, profesor astronomie na české univerzitě.

Ještě jako univerzitní student se stal Seydler ve školním roce 1867/68 členem tehdy malého *Spolku pro volné přednášky z mathematiky a fysiky* (spolek byl založen roku 1862, od roku 1869 nesl název *Jednota českých mathematiků*, dnes jej známe pod názvem *Jednota českých matematiků a fyziků*). Seydler ihned začal aktivně pracovat na povznesení této organizace. 8. 2. 1868 byl zvolen jednatelem tohoto spolku a tuto činnost vykonával i v letním běhu. Od této doby býval A. Seydler každoročně až do své smrti volen do výboru tohoto spolku. 11. února 1872 začala vydávat Jednota časopis věnovaný matematice a fyzice s názvem *Časopis pro pěstování mathematiky a fysiky* a o rok později i učebnice a vědecké monografie. Seydler působil v řadě komisí, které z Jednoty vyšly, míval na jejích schůzích časté přednášky a přispíval do zpráv i časopisu společnosti. Jako výraz vděčnosti zvolila Jednota Seydlera dne 13. prosince 1884 čestným členem (tím byl například roku 1870 zvolen i profesor Ernst Mach). Přestože v posledních letech života byl Seydlerův život ovlivněn jeho postupující nemocí a naplněn vědeckou prací, zůstal rádcem Jednoty v otázkách vědeckých a jako první knihovník dohlížel na úroveň knihovny, která by umožňovala studentům matematiky a fyziky seznamovat se s nejnovějšími světovými pracemi v těchto oborech. Jako výraz zvláštní úcty, kterou k této organizaci choval a ve které strávil 20 let, věnoval Jednotě roku 1887 Newtonovy *Principia*, které vydal vlastním nákladem při příležitosti dvoustého výročí jejich prvního vydání.

Seydler udržoval dobré vztahy s kolegy z německé univerzity i z klementinské hvězdárny. Za své vědecké zásluhy byl 5. května 1880 zvolen mimořádným, dne 11. června 1884 již řádným členem královské české *Společnosti nauk*. 3. července roku 1890 se ihned po jejím založení stal členem české *Akademie pro vědy, slovesnost a umění*. Dále byl členem lipské astronomické společnosti i členem řady spolků dobročinných a vlasteneckých.

Celý život usiloval Seydler o povznesení české vědy. Podařilo se mu vyniknout jednak na poli matematiky a fyziky, ale zejména v astronomii. Mezi jeho nejvýznamnější fyzikální práce patří třídílná učebnice *Základové theoretické fy-*

[195] Pro srovnání: vídeňská hvězdárna byla postavena s rozpočtem milion zlatých a jen dalekohled stál 100 000 zlatých.

siky[196]. Nejvýznamnějších úspěchů dosáhl na poli astronomie. Byl jejím prvním profesorem na české univerzitě. Protože u nás neexistovala žádná česká učebnice astronomie (pouze spíše populární učebnice F. J. Studničky), rozhodl se vytvořit velké původní dílo o astronomii (podobné jeho učebnici teoretické fyziky), a tím i vytvořit českou astronomickou terminologii, která se až na malé výjimky zachovala do dnešní doby. Bohužel dílo nedokončil, vypracovány byly jen některé kapitoly. Podobnou učebnici pak vydal jeho nástupce v astronomickém ústavu profesor Gustav Gruss[197]. Seydlerovy astronomické práce můžeme rozdělit do tří hlavních okruhů:

1) určení dráhy planet a komet;
2) problém tří, popřípadě čtyř těles (byl prvním, kdo v našich zemích problém tří těles řešil a k jeho řešení přispěl i novými objevy);
3) aktuální astronomické a astrofyzikální problémy – např. Keplerova rovnice (hledání nejrychlejšího, nejpřesnějšího a nejobecnějšího prostředku k řešení této transcendentní rovnice).

Jeho vědecká práce na poli astronomie získala uznání v zahraničí (jeho nekrolog byl uveřejněn v časopise *Astronomische Nachrichten*, o jeho pracích referoval francouzský časopis *Bulletin astronomique*). Vedle toho se věnoval i popularizaci astronomie. Aktivně se zajímal o filozofii a politiku.

Většinu svých vědeckých prací publikoval v *Časopise pro pěstování matematiky a fyziky*, ve *Zprávách o zasedání Císařské akademie věd ve Vídni* (*Sitzungberichte der kaiserlichen Akademie der Wissenschaften in Wien*), vědu popularizoval zejména v *Athenaeu*, jehož byl stálým přispěvovatelem, v *Květech*, Sládkově *Lumíru*, *Osvětě* a *Živě*. Jeho populární články se setkávaly se značným úspěchem, neboť dovedl věc přednést jasně a srozumitelně, přitom zajímavě, a to i čtenářům méně odborně vzdělaným. Snad se zde projevila jeho záliba v četbě populárních spisů francouzských a anglických. Roku 1871 publikoval v německém časopise spolku *Lotos*, přednášel v *Americkém dámském klubu* Vojtěcha Náprstka. Účastnil se rovněž pokusu o vznik odborného matematicko-fyzikálního časopisu *Archiv matematiky a fysiky*. Společně s Emilem Weyrem a Karlem Domalípem byl redaktorem jeho druhého svazku (1877). Z osob, s kterými přicházel nejčastěji do styku, jmenujme profesory Hornsteina, Macha, Studničku a zvláště jeho přítele Čeňka Strouhala.

V osobním životě byl pronásledován množstvím útrap. 26. dubna 1876 se oženil s Annou Weyrovou (narozena 1855), dcerou Františka Weyra, c. k. profesora na německé reálce v Praze. Jeho švagry tedy byli matematikové Eduard (1852–1903) a Emil (1848–1894) Weyrovi. Z tohoto šťastného svazku se naro-

[196] Díl první: *Všeobecný úvod a mechanika*. Slavík a Borový, Praha 1880, 394 označených stran. Díl druhý: *Theorie potenciálu, úkazů gravitačních, magnetických a elektrických*. Borový, Praha 1885, 416 stran. Díl třetí: *Mechanika molekulární, theorie vibračního a undulačního pohybu, akustika*. Sepsal August Seydler, vydal a doplnil František Koláček, Čes. akademie císaře Františka Josefa pro vědy, slovesnost a umění, Praha 1895, 416 označených stran, z toho však více než polovina z pera Koláčkova.

[197] *Základové teoretické astronomie I* (1897/8), II (1899/1900); nákladem České akademie císaře Františka Josefa pro vědy, slovesnost a umění.

dily čtyři děti: Jaroslav[198] zemřel brzy po porodu, Anna Amidea Helena[199], Antonie Marie[200] a Jan[201]. Zanedlouho po smrti bratra Karla v roce 1883 v 29 letech umírá[202] i jeho žena na ochrnutí plic. Zůstaly po ní tři malé děti, z nichž nejstarší Heleně bylo 5 let a nejmladšímu Janovi pouhé 2 roky. Brzy po ní umírá i Seydlerova nevlastní matka Antonie[203], roku 1886 starší sestra Anna ve věku 43 let (mladší Marie zemřela již roku 1869 v pouhých 24 letech) a otec[204]. V únoru 1890 se výrazně zhoršil i zdravotní stav Augustina Seydlera (tuberkulóza se u něj projevovala již od roku 1873). Přes léto příznaky nemoci ustoupily a v září roku 1890 se Seydler kvůli svým dětem znovu oženil se svou hospodyní Marií Stehlíkovou. Nemoc udeřila znovu v únoru roku 1891, v březnu Seydler odjíždí do Gorice, kde se mu vlivem nepříznivého počasí ještě více přitížilo, vrací se proto v dubnu vlakem zpět do Prahy, kde 22. června 1891 v 7 hodin ráno na tuberkulózu ve svém bytě na Letné č. 80 umírá. Pochován je v Praze na Olšanech. Na jeho počest je pojmenována planetka č. 6 586 objevená astronomem Antonínem Mrkosem (1918–1996) a jeho jménem jsou nazvány rovněž dvě ulice (v Praze a v Hradci Králové).

2.12.3 Časopis Athenaeum a Rukopisy

Jistý čas pobývala rodina Seydlerova ve vile Václava Vlčka Osvěta na Vinohradech[205] společně s rodinou T. G. Masaryka[206]. Dům stojí dodnes v ulici Jana Masaryka na pražských Vinohradech. Profesor T. G. Masaryk (1850–1937) byl povolán z Vídně roku 1882 na českou část rozdělené univerzity. Roku 1883 začal Masaryk vydávat u nakladatele J. Otty (1841–1916) kritickou vědeckou revui nazvanou *Athenaeum* (s podtitulkem *Listy pro literaturu a kulturu vědeckou*). August Seydler jako Masarykův kolega do této revue také přispíval svými články. V tomto období se rozpoutal boj zastánců *Rukopisu zelenohorského a královédvorského* za jejich pravost. Byl vyvolán především článkem [15] z roku 1886 bohemisty profesora Jana Gebauera (1838–1907), který svým jazykovědným rozborem ukázal na jejich nepravost. Masaryk se k článku přidal jako spoluautor. A právě na podnět T. G. Masaryka se do rukopisného sporu zapojil i Augustin Seydler.

Ve svém článku z roku 1886 *Počet pravděpodobnosti v přítomném sporu*, který vyšel v revue *Athenaeum*, se pokusil využít k řešení rukopisné problematiky matematický aparát. Je to vůbec první česká práce, která využívá matematiku (konkrétně teorii pravděpodobnosti) k řešení lingvistických otázek. A pravděpodobně jde i o první práci tohoto typu v celosvětovém měřítku.

[198] Narozen 13. 12. 1877.
[199] Narozena 15. 12. 1879. Ze vzpomínek Alice Masarykové se dovídáme, že to byla geniální matematička, která obdivovala svého otce, po němž zdědila své matematické nadání. Bohužel po jeho smrti nemohla studiem na univerzitě toto nadání dále rozvinout.
[200] Narozena 19. 4. 1881.
[201] Narozen 12. 9. 1882.
[202] †2. 10. 1884.
[203] †23. 3. 1885.
[204] †13. 7. 1888.
[205] Od května 1887 do Seydlerovy smrti pobývá rodina na Letné.
[206] Ti zde pobývali v letech 1886 až 1889.

Augustin Seydler vycházel z již zmíněné stati J. Gebauera a matematickými prostředky se pokusil vyčíslit jeho zjištění jazyková.

Jan Gebauer si prošel těžkým vnitřním zápasem z obránce *Rukopisů* v jejich odpůrce. První pochybnosti uveřejnil již roku 1885 v *Ersch-Gruberově Všeobecné encyklopedii*. Gebauer chtěl zpočátku projednávat problémy RKZ v uzavřeném kruhu vědeckých kapacit bez účasti veřejnosti, ale našel odezvu u T. G. Masaryka v *Athenaeu*, které sdružovalo zejména mladé profesory české univerzity, a 15. února 1886 zde vychází jeho článek *Potřeba dalších zkoušek rukopisu Královédvorského a Zelenohorského* ([15]). Článek je rozčleněn na tři části. V stěžejní první části svého článku se Gebauer zabývá podrobným jazykovým rozborem RKZ a všímá si celé řady odchylek od „*jazyka staročeského normálního*", tj. od staré češtiny vyskytující se v dochovaných písemných památkách, které nebyly prokazatelně padělány. Tyto odchylky by bylo možno vysvětlit například existencí dosud neznámého dialektu apod. Protože se zde ale vyskytuje také nápadně mnoho takových odchylek, které se současně objevují v pracích Václava Hanky vzniklých před objevením RKZ a v již prokázaných padělcích, jež byly dávány rovněž do souvislosti s Václavem Hankou, vyslovil Gebauer hypotézu o nepravosti RKZ. Všechny tyto okolnosti podle něj zavdávají příčinu k pochybnostem. Vedle toho navrhl další zkoumání, zejména paleografické a chemické. *Koincidencím* (zde máme na mysli shodné odchylky vyskytující se současně v RKZ a u Hanky, sám Gebauer tento pojem v práci nepoužívá) přikládá větší důraz z toho důvodu, že objevitel RK byl „*notorický falzátor*", a dále proto, že tyto koincidence s Hankou lze určit s jistotou, zatímco neshody se starou češtinou mohou být způsobeny neúplným poznáním jazyka tehdejší doby.

Ve druhé části článku Gebauer popisuje vývoj svého vztahu k otázce pravosti *Rukopisů* ([15], str. 160n.):

„*Nejraději bych byl nalézal důkazy, že oba texty jsou úplně přesné; ale to se mi nedařilo, naopak odkrývaly se mi nové a nové známky, které budily pochybnosti. Nemohu vysloviti, jak mi nálezy tyto bývaly trapné a jaký mi způsobovaly zármutek. Bylyť úplně proti mému přání a nadání; bořily, co jsem dotud měl za pevné a co jsem se strojil upevniti ještě více; a přiváděly mi na mysl, že mi časem nastane povinnosť, předložiti tyto věci ku povážení obecnému, – povinnost zajisté vědecká a vlastenecká, ale spolu těžká a traplivá, poněvadž zpráva nepříjemná nikdy se ráda neslyší.*"

Dále se zmiňuje o práci A. Vaška z roku 1879, jehož některé námitky se mu nepodařilo vyvrátit a o žádostech profesorů A. Leskiena a T. G. Masaryka o napsání článků právě do *Ersch-Gruberovy Všeobecné encyklopedie* i posléze do *Athenaea*. A konečně se zde vyjadřuje k článku Martina Hattaly, zastánce pravosti *Rukopisů*, který vyšel 16. ledna 1886 v *Národních listech* jako reakce na Gebauerův text v *Ersch-Gruberově encyklopedii*.

Třetí část Gebauerova článku tvoří dopis redaktora *Athenaea* T. G. Masaryka (datovaný 18. 1. 1886), ve kterém žádá Gebauera o napsání článku týkajícího se jazykové problematiky RKZ a ve kterém rovněž uvádí důvody, jež ho k této prosbě vedou.

Jak již bylo zmíněno, Gebauer zdůraznil ve svém článku potřebu nového

chemického a paleografického rozboru RKZ. Dřívější chemická zkoumání RKZ totiž nevyznívaly jednoznačně pro žádnou z hypotéz. V roce 1840 bylo zjištěno, že stáří pergamenu, na kterém je napsán text RZ, opravdu odpovídá období, do kterého je rukopis datován. Jedná se totiž o *palimpsest*, jak se zjistilo později, tj. dílo vzniklé odstraněním původního textu a nahrazením textem novým. Rovněž chemický rozbor iniciál RK z roku 1880 nedokázal jednoznačně rozhodnout. Některé z původních iniciál totiž byly využity v textu novém. Protože ani tento výzkum pochybnosti jazykovědců jednoznačně nepotvrdil, trval Gebauer na rozhodujícím významu jazykových důkazů. Podle něj je při posuzování pravosti nebo nepravosti v každém konkrétním případě hlavní ta věda, která je schopna přinést pozitivní a zcela jednoznačné důvody pro jednu z hypotéz[207].

2.12.4 Seydlerův „matematický" důkaz nepravosti Rukopisů

Gebauerův článek se stal východiskem z dnešního pohledu průkopnického činu matematika, fyzika a astronoma Augustina Seydlera, který poprvé v historii české vědy (a pravděpodobně i v měřítku celosvětovém) použil matematiky k řešení otázky jazykové povahy (viz [56], [57]) a stal se tak vlastně průkopníkem dnešní kvantitativní lingvistiky. Profesor Augustin Seydler působil od prvního ročníku v *Athenaeu*, na jehož stránkách se roku 1886 znovu rozhořel rukopisný spor, jako odborný redaktor pro filosofii, a proto byl dobře obeznámen s řadou závažných materiálů týkajících se pravosti RKZ. Seydlerův článek byl publikován pouhé dva měsíce po Gebauerově studii jazyka RKZ. Gebauer zjistil, že *Rukopisy* obsahují jednak velké množství *odchylek* od normální mluvnice, tzn. od mluvnice v prokazatelně staročeských památkách, jednak se řada těchto odchylek shoduje s odchylkami nalezenými v prokazatelných padělcích spojených s osobou Václava Hanky a také v dílech Václava Hanky vzniklých před objevením RKZ. Seydler nahrazuje Gebauerův pojem „*odchylka*" výstižnějším pojmem „*zvláštnost*".

Nejprve Seydler v úvodu vysvětluje svoji motivaci ([57], str. 299):

„*Čísla mluví! Tento výrok slýcháme a pronášíme často, leč zřídka kdy se dle něho řídíme. Suchopárnou leč zřetelnou mluvu cifer nahrazujeme rádi mluvou obyčejnou i tam, kde bychom ku prvé sáhnouti a tím na př. ve vědeckém sporu sílu neb váhu důvodů a protidůvodů přesně vymeziti mohli.*"

Dále zde Seydler zmiňuje zdroje, z kterých při psaní článku čerpal. Vedle Gebauera se odkazuje na práci profesora T. G. Masaryka *Počet pravděpodobnosti a Humeova skepse* z roku 1883, v níž Masaryk uvádí některé významné představitele tohoto oboru (Gauss, Laplace, Poisson, Quetelet aj.) a v níž zdůrazňuje

[207] Další důkladnou chemickou expertízou byli pověřeni profesoři Antonín Bělohoubek a Vojtěch Šafařík, syn P. J. Šafaříka. Oba pracovali nezávisle na sobě a A. Bělohoubek objevil na jedné z iniciál RK berlínskou modř, známou až od roku 1724. Podrobnosti jsou obsaženy ve *Zprávě o chemickém a drobnohledném ohledání některých rukopisů musejních* z roku 1887. Vzhledem k tomu, že metody fyzikálního a chemického zkoumání od minulých dob značně pokročily, byly *Rukopisy* v letech 1967–1971 na podnět spisovatele literatury faktu Miroslava Ivanova znovu zkoumány pracovníky *Kriminalistického ústavu Federální kriminální ústředny* v Praze. Výsledkem bylo prohlášení, že se nejspíše jedná o palimpsesty a podvrhy.

význam počtu pravděpodobnosti pro induktivní logiku a také poukazuje na možnosti jeho využití v oborech zcela nečekaných (například v soudní praxi). Vedle toho Seydler čtenáře článku odkazuje na Poissonovu práci *Recherches sur la probabilité*[208] a na práci profesora Augustina Pánka[209].

Jak již bylo řečeno, aplikací matematiky na řešení problému jazykového se Seydler zařadil k předchůdcům matematické lingvistiky a z pera A. Seydlera zde můžeme vysledovat některé její důležité myšlenky ([57], s. 299):

„*Zde se začíná říše mathematiky; jiné vědy mohou jí předložiti své výsledky co substrat; – jsou-li takové, že je lze mathematicky formulovati, na př. v souboru jistých zjevů spočítati zjevy zákonité a zjevy nahodilé, má mathematika právo učiniti z nich material dalších dedukcí, ať jsou vzaty z přírodních věd, ze sociologie nebo třeba i z filologie.*“

Podstata Seydlerova článku spočívá v tom, že se pomocí počtu pravděpodobnosti pokouší vyčíslit následující dvě hypotézy:

1) pravděpodobnost toho, že výskyt *zvláštností* v RK je náhodný (RZ zde na rozdíl od Gebauera pro zjednodušení neuvažuje) – zvláštnosti vznikly například jako chyby písařů, opisovačů, porušením písma vlivem času apod.
2) pravděpodobnost toho, že výskyt takzvaných *koincidencí* (tj. zvláštností vyskytujících se současně v RK a spisech některých autorů 19. století) je náhodný.

Velké hodnoty těchto pravděpodobností by svědčily pro pravost *Rukopisů*, malé pro jejich nepravost. Aby ukázal, že nestraní pochybovačům pravosti *Rukopisů* (mezi které se sám řadí), ve všech případech, kde není možný přesný matematický rozbor, chce rozhodovat ve prospěch hypotézy pouhé náhody v rámci mezí počtu pravděpodobnosti.

Nejprve uvádí definici abstraktní pravděpodobnosti jevu jako „*poměr příznivých pro ten zjev případů ku všem možným případům*“ ([57], s. 300). Následuje model tzv. binomického rozdělení pravděpodobností, který můžeme dnešní terminologií vyjádřit takto: Nechť je dána posloupnost n nezávislých náhodných pokusů, z nichž každý pokus má dva možné výsledky, a to že buď nastane náhodný jev A (s pravděpodobností p), nebo nenastane. Pravděpodobnost, že v n pokusech nastane jev A právě k-krát, je vyjádřena vztahem:

$$P_k = \binom{n}{k} p^k (1-p)^{n-k}; \qquad k = 0, 1, \ldots, n.$$

Protože Seydler hodnotu p nezná, omezuje P_k alespoň shora:

$$P_k \leq \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k}.$$

[208] Pracuje s německým překladem od Schnusse *Lehrbuch der Wahrscheinlichkeitsrechnung* z roku 1841.

[209] *Pravděpodobnost a posteriori I, II.* Časopis pro pěstování mathematiky a fysiky XII, s. 227–232, s. 281–294.

Na konkrétních údajích z *Rukopisu královédvorského*, jak je získal J. Gebauer, pak počítá pravděpodobnosti jednotlivých hypotéz, tzn. pravděpodobnost náhodného výskytu zvláštností v RK a pravděpodobnost náhodného výskytu *koincidencí*. Nyní se u jednotlivých hypotéz podrobně zastavme.

Pravděpodobnost náhodného výskytu zvláštností v RK

Rukopis královédvorský je tvořen 5 400 slovy, podle Gebauera se v něm vyskytuje asi 700 *zvláštností*, pro zjednodušení (a ve prospěch hypotézy pouhé náhody) Seydler uvažuje těchto zvláštností pouze 540. Tyto zvláštnosti jsou různého druhu a každá taková skupina zvláštností by měla být posuzována samostatně. Seydler provádí výpočet vzhledem k jedné takové skupině zvláštností.

Příklad 2.4: Podle J. Gebauera se v RK vyskytuje 89 imperfekt (druh minulého času), 17 z nich se svými koncovkami liší od jiných památek druhé třetiny 14. století (např. *proměniše*, *řváše*, *blížíše*). Pro pravděpodobnost, že se tyto odchylky vyskytly pouhou náhodou, platí:

$$P \leq \binom{89}{17}\left(\frac{17}{89}\right)^{17}\left(\frac{72}{89}\right)^{72} = 0{,}1070261 < \frac{10}{93} < \frac{1}{9}.$$

Dále Seydler uvažuje pravděpodobnost společného výskytu všech skupin zvláštností. Řečeno dnešní terminologií: Mějme N skupin zvláštností s pravděpodobnostmi $P_{k_1}, P_{k_2}, \ldots, P_{k_N}$, kde $k_1, k_2, \ldots, k_N$ jsou počty zvláštností v dané skupině. Pravděpodobnost současného výskytu k_1 zvláštností v první skupině, k_2 zvláštností v druhé skupině atd. je za předpokladu nezávislých výskytů různých skupin zvláštností rovna součinu $P_{k_1} \cdot P_{k_2} \cdot \ldots \cdot P_{k_N}$.

Příklad 2.5: Kdybychom uvažovali, že každá skupina obsahuje 17 odchylek tak, jak tomu bylo v předešlém příkladu, získali bychom přes 30 skupin odchylek. Pro zjednodušení (a rovněž ve prospěch hypotézy náhody) uvažuje Seydler 10 skupin zvláštností a pravidelné rozdělení slov a zvláštností do skupin, tj. v každé skupině se vyskytuje 540 slov a 54 zvláštností. Pak pravděpodobnost toho, že ve skupině o 540 slovech se vyskytuje 54 zvláštností, je rovna:

$$P_{54} \leq \binom{540}{54}\left(\frac{1}{10}\right)^{54}\left(\frac{9}{10}\right)^{486} = 0{,}057137 < \frac{1}{17}.$$

Budeme-li dále uvažovat, že všechny pravděpodobnosti $P_{k_1}, P_{k_2}, \ldots, P_{k_N}$ jsou menší než $\frac{1}{9}$, jak tomu bylo v příkladu s imperfekty, je pravděpodobnost, že by se všechny skupiny zvláštních tvarů dostaly do RK náhodou, omezena shora hodnotou

$$\left(\frac{1}{9}\right)^{10} = \frac{1}{3486784401}.$$

„*Jinými slovy: máme 3480 millionů stejně závažných důvodů proti jednomu domnívati se, že nejsou v RK se vyskytující zvláštnosti vesměs pouhou náhodou způsobeny.*“ ([57], str. 303)

Seydler si uvědomuje, že výklad oněch zvláštností již matematikům nepřísluší, přesto se pozastavuje nad některými možnými příčinami – dialekt, poetická mluva, nevědomost původce apod.

Pravděpodobnost náhodného výskytu koincidencí v RK a spisech některých autorů 19. století

I zde Seydler uvádí několik konkrétních případů založených na Gebauerových zjištěních. Co se týče výpočtu pravděpodobnosti náhodného výskytu *koincidencí* ve dvou textech (počet shodných zvláštností), tuto pravděpodobnost pouze omezuje shora. Přesné výpočty pravděpodobnosti náhodného výskytu koincidencí provedla ve svém článku J. Zichová (viz [82]).

Nejprve Seydler seznamuje čtenáře s problematikou koincidencí. Nechť text A je tvořen $c = a + b$ tvary, kde a je počet zvláštností, b počet pravidelných tvarů, text B je tvořen $\gamma = \alpha + \beta$ tvary, kde α je počet zvláštností, β počet tvarů pravidelných. Nechť mají oba texty p společných tvarů (pravidelných i zvláštností). Jaká je pravděpodobnost, že mají oba texty právě m společných zvláštností? Při odvozování vzorce pro výpočet pravděpodobnosti koincidencí Seydler uvažuje následujícím způsobem. V c tvarech textu A se a zvláštností může náhodně vyskytnout $\binom{c}{a}$ způsoby, v γ tvarech textu B se α zvláštností může vyskytnout $\binom{\gamma}{\alpha}$ způsoby. Celkem tedy máme $\binom{c}{a}\binom{\gamma}{\alpha}$ možných způsobů pro náhodný výskyt zadaného počtu zvláštností a toto číslo bude ve jmenovateli hledaného zlomku. Čitatele zlomku odvozuje Seydler pomocí této úvahy. Z celkového počtu p společných tvarů v textech A a B lze m společných zvláštností vybrat $\binom{p}{m}$ způsoby. V textu A tedy nyní zbývá $(c - m)$ tvarů, z kterých vybíráme $(a - m)$ zvláštností vyskytujících se pouze v A, což lze učinit $\binom{c-m}{a-m}$ způsoby. V textu B zbývá $(\gamma - m)$ tvarů, z nichž vybíráme $(\alpha - m)$ zvláštností vyskytujících se pouze v B, a to lze $\binom{\gamma-m}{\alpha-m}$ způsoby. Při tomto způsobu uvažování vybíráme ovšem $(a-m)$ (respektive $(\alpha-m)$) zvláštností z příliš velkého počtu tvarů. Tak například v textu A se mezi $(c-m)$ tvary vyskytují jak zvláštnosti a pravidelné tvary patřící pouze do A, tak pravidelné tvary společné oběma textům. Proto počet příznivých případů pro objevení se m koincidencí zvláštností je zcela jistě menší nebo roven číslu $\binom{p}{m}\binom{c-m}{a-m}\binom{\gamma-m}{\alpha-m}$a pro hledanou pravděpodobnost P, že oba texty mají právě m společných zvláštností, platí, že

$$P \leq \frac{\binom{p}{m}\binom{c-m}{a-m}\binom{\gamma-m}{\alpha-m}}{\binom{c}{a}\binom{\gamma}{\alpha}}.$$

Tuto pravděpodobnost J. Zichová ve svém článku odvozuje přesně jako

$$P = \frac{\binom{p}{m}\binom{c-p}{a-m}\binom{\gamma-p}{\alpha-m}}{\binom{c}{a}\binom{\gamma}{\alpha}}.$$

Zastavme se nyní u konkrétních příkladů.

Příklad 2.6: V RK se objevuje slovo „*upěchu*" ve smyslu *spěchali*, v Jungmannově básni z roku 1808 slovo „*upěje*" ve smyslu *spěje*. V RK se vyskytuje celkem 500 různých slov, v nichž by se podobná zvláštnost mohla objevit, u Jungmanna je takových různých slov 2 000. Oba spisy mají 200 společných slov. Jaká je pravděpodobnost, že se tato „*nápadná zvláštnost*" objevuje v obou spisech v témže slově? Podle Seydlera je

$$P \leq \frac{\binom{200}{1}\binom{499}{0}\binom{1999}{0}}{\binom{500}{1}\binom{2000}{1}} = \frac{1}{5000} = 0{,}0002.$$

Přesná hodnota podle Zichové je

$$P = \frac{\binom{200}{1}\binom{300}{0}\binom{1800}{0}}{\binom{500}{1}\binom{2000}{1}} = \frac{1}{5000} = 0{,}0002.$$

Příklad 2.7: V RK i v textech Jungmannových se objevuje zvláštnost v rodu slova *vrchol* (zde rod ženský místo rodu mužského). Jestliže se v RK vyskytuje 200 substantiv, u Jungmanna 800 substantiv, z nichž 100 je společných oběma textům, pak pravděpodobnost náhodného výskytu této koincidence je podle Seydlera

$$P \leq \frac{\binom{100}{1}\binom{199}{0}\binom{799}{0}}{\binom{200}{1}\binom{800}{1}} = \frac{1}{1600} = 0{,}000625.$$

Zichová pak doplňuje přesnou hodnotu

$$P = \frac{\binom{100}{1}\binom{100}{0}\binom{700}{0}}{\binom{200}{1}\binom{800}{1}} = \frac{1}{1600} = 0{,}000625.$$

Pravděpodobnost náhodné koincidence obou těchto zvláštností z příkladu 2.6 a příkladu 2.7 je pak

$$\frac{1}{5000} \cdot \frac{1}{1600} = \frac{1}{8000000} = 1{,}25 \cdot 10^{-7},$$

„t. j. lze se založiti 8 miliony proti jedné, že pouhou náhodou se takové dvě koincidence mezi RK a básnickými spisy Jungmannovými vyskytnouti nemohou. Podobných koincidencí máme pěknou řadu: zkad, zhovadilý, jme se činiti něco; iskat něco (akk.) atd. O těch zatím pomlčíme." (Viz [57], str. 304.)

Příklad 2.8: V RK a ve spisech Hanky (*Uvedení*, *Interpolace*, *Píseň na Vyšehrad*) se objevuje přinejmenším 12 skupin zvláštností (např. odchylné tvary imperfekta, duál feminin *tě* místo *ta*, zaměňování imperfekta s aoristem). Známe tyto údaje: v RK je celkem 200 prvků, z toho je 40 zvláštností, u Hanky se objevuje 50 prvků, z nichž zvláštních je 16, prvků společných oběma textům je 40, přičemž společných odchylek je 12. Kdybychom vyšetřovali pouze 1 jedinou

shodnou zvláštnost v RK a u Hanky, a to například substantivum *pochládeček* (toto substantivum se vyskytuje pouze v RK a v *Písni na Vyšehrad*, jinak v celé české literatuře do roku 1817 neexistuje žádný jiný doklad tohoto slova), zjistili bychom, že pravděpodobnost náhodného výskytu této jediné koincidence je

$$P \leq \frac{\binom{40}{1}\binom{199}{0}\binom{49}{0}}{\binom{200}{1}\binom{50}{1}} = \frac{1}{2500} = 0{,}0004.$$

Pravděpodobnost náhodného výskytu všech 12 koincidencí v těchto textech je podle Seydlera

$$P \leq \frac{\binom{40}{12}\binom{188}{28}\binom{38}{4}}{\binom{200}{40}\binom{50}{16}} = 7{,}6615 \cdot 10^{-8} < \frac{1}{13000000}.$$

Podle Zichové je to přesně

$$P = \frac{\binom{40}{12}\binom{160}{28}\binom{10}{4}}{\binom{200}{40}\binom{50}{16}} = 1{,}6073 \cdot 10^{-12} \approx \frac{1}{622178000000}.$$

Seydler předešlé výpočty shrnuje do myšlenky, že „*lze tedy sázeti alespoň 100 billionů proti jedné, že není současně nahodilou shoda RKho s jistými před objevení RKho známými publikacemi XIX. století.*“ Součin pravděpodobností $\frac{1}{8000000}$ (viz příklad 2.7) a $\frac{1}{13000000}$ (viz příklad 2.8) nám totiž dává hodnotu $\frac{1}{104 \cdot 10^{12}}$.

Příklad 2.9: Na tomto příkladu se Seydler snaží ukázat, jak rozdělení zvláštností do skupin ovlivňuje pravděpodobnost koincidencí. Mějme prvky z příkladu 2.8, to znamená $c = 200$, $a = 40$, $\gamma = 50$, $\alpha = 16$, $p = 40$, $m = 12$. Rozdělme tyto prvky do dvou skupin tak, že v každé z nich budou poloviční počty prvků z předešlého příkladu, tj. $c = 100$, $a = 20$, $\gamma = 25$, $\alpha = 8$, $p = 20$, $m = 6$. V každé ze skupin je pak dle Seydlera pravděpodobnost náhodného výskytu 6 společných zvláštností

$$P \leq \frac{\binom{20}{6}\binom{94}{14}\binom{19}{2}}{\binom{100}{20}\binom{25}{8}} = 1{,}9926 \cdot 10^{-4} < \frac{1}{5000} = 0{,}0002.$$

Podle Zichové je přesná hodnota

$$P = \frac{\binom{20}{6}\binom{80}{14}\binom{5}{2}}{\binom{100}{20}\binom{25}{8}} = 1{,}0084 \cdot 10^{-6} \approx \frac{1}{991697}.$$

Pravděpodobnost náhodné koincidence těchto dvou skupin zvláštností je pak zcela jistě menší než $\frac{1}{25000000}$, což je ještě méně, než hodnota zjištěná v příkladu 2.8 ($\frac{1}{13000000}$). Rozdělením zvláštností do skupin tedy hodnota pravděpodobnosti ještě více klesá.

Na základě předešlých příkladů je zřejmé, že velmi nízké hodnoty pravděpodobnosti straní spíše nepravosti *Rukopisů*, a Seydler k tomu říká ([57], str. 306):

„*Mohu právem tvrditi, že se mi podařilo, oč mi v této otázce jediné šlo; vyloučiti náhodu z přítomného sporu.*“

Seydler zdůrazňuje rovněž to, že pravděpodobnosti jsou v této práci vyčísleny jen na základě pouhého počítání jednotlivých zvláštností. Podle něj by jednotlivým typům zvláštností měly být přiřazeny váhy podle jejich důležitosti. To je ale úkol filologie. Jaký vliv by tyto váhy měly na celkový výsledek, nelze však ani přibližně určit. V samotném závěru práce pak na základě svých výpočtů usuzuje, že buď autor RK znal práce literátů 19. stol., nebo čeští literáti v letech 1803–1817 znali RK.

Seydlerovo vystoupení ale nezůstalo bez vášnivých reakcí na straně zastánců pravosti RKZ. V *Národních listech* J. Grégr reagoval těmito slovy: „*K řadě dosavadních útočníkův přistoupili útočníci noví; zorganizovala se hotová literární mafia, která nesahá sice do kapes jednotlivcův, ale vztahuje své smělé ruce po nejdražších skvostech národa. Je to kotérie na mladé naší universitě, která opatřila sobě již i svého šaška, jako jej mívaly středověké „rytířské“ výpravy, když šlo o to, zaskočit a obrat kupce, vracející se s vzácným zbožím z dalekého trhu.*“ Tímto „*šaškem*“ není označován nikdo jiný, než „*šprýmovný matematik*“ Seydler. Rovněž Seydlera srovnává s „*důmyslnými badateli středověku*“, počítajícími „*kolik miliónů čertů může tančit na špičce jehly*“ a aplikaci matematiky hodnotí slovy: „*Když již jde se na naše rukopisy s takovýmito matematickými formulemi, můžeme se toho nadíti, že v nejblíže příštím čísle Athenaea rozmnoží se společnost našich „vědců“ ještě i nějakým spiritistou, který bude citovat proti našim rukopisům duchy (. . .).*“[210]

Historik J. Kalousek jako jediný z obhájců RKZ nejspíše porozuměl Seydlerovým matematickým dedukcím a poznal i jejich argumentační polemickou sílu. Ve své reakci na Seydlerův článek[211] možná proto nechává matematický problém bez povšimnutí, ale táže se, jak pravdivá jsou výchozí jazyková zjištění („*. . . o matematické dedukce ani nezavadím, není mně toho potřebí. Stojím před nimi jako kupříkladu před mlýnem, jehož strojům nemusím rozuměti. Já jen dávám pozor, co se do mlýna vozí a na koš sype. . .*“). Zpochybnil zde zejména Gebauerovy závěry o koincidencích, které se později staly slabým místem jeho práce a od kterých se v 2. polovině 20. století začalo v argumentaci proti pravosti *Rukopisů* upouštět[212]. Gebauer zde totiž směšuje dvě otázky: otázku pravosti nebo nepravosti a otázku původce, které jsou různého řádu a důležitosti, první je nadřazena druhé. Nepříznivé důsledky to má pro Gebauera zvláště v těch případech, kdy se Hankovy chybné názory podobají jazyku památek z přelomu 14. a 15. století. Řada obránců je totiž ochotna připustit vznik RK v 15. století.

Augustin Seydler reaguje na J. Kalouska v článku ([56], str. 446) těmito slovy:

„*Nechci dokazovati mathematikou něco, co může dokázati jen linguistika, palaeografie nebo historie (sociologie); vida však, jak málo se cítí váha pochybností těmito vědami, zejména první z nich, pronešených, chtěl jsem pokuď*

[210] Viz Grégr, J.: *Na obranu RKZ.* 1886, s. 74–77.
[211] *Osvěta* 16, 1886, 556–558.
[212] Komárek, M.: *Jazykovědná problematika RKZ.* In: [43].

bylo lze, váhu tu číselně vyjádřiti. Podal jsem – a to mi nikým nebylo dosud vyvráceno – důkaz, že mají odborníci povinnosť, vyložiti nejen zvláštnosti RKého, nýbrž, nač zvláštní důraz kladu, i koincidence těch zvláštností se stejnými zvláštnostmi literárních zjevů XIX. století."

Matematika zde vstoupila spíše jako prostředek zesilující hlas lingvistiky, než jako nástroj napomáhající jazykovědnému poznání. Je politováníhodné, že tato spolupráce matematiky a lingvistiky nemohla být v tomto období dále rozvíjena, neboť Seydler se tímto příspěvkem zařadil k předchůdcům matematické lingvistiky v celosvětovém kontextu. Připomeňme si, že roku 1847 jako vůbec první na využití matematiky v lingvistice upozornil ruský matematik V. Ja. Bunjakovskij v časopise *Sovremennik*. K dalším předchůdcům řadíme W. Gambleho, který roku 1861 vydal v Šanghaji pro potřeby tiskařů práci *Two Lists of Selected Characters Containing All in the Bible and Twenty Seven Other Books*. A konečně do tohoto časového období spadá i řešení tzv. *platónské otázky*, tj. určení, které *Listy* či *Dialogy* napsal skutečně Platón. K tomuto účelu byla vypracována tzv. *stylometrická metoda*, která byla založena na frekvenci některých formálních slov a hiátů. Důležité je ovšem uvědomit si to, že Seydler, na rozdíl od ostatních předchůdců matematické lingvistiky, nezůstává na úrovni pouhé deskriptivní statistiky, ale provádí pravděpodobnostní hodnocení takových statistik vzhledem k různým možným hypotézám.

Z hlediska otázky pravosti RKZ, můžeme Seydlerův postup jen těžko považovat za důkaz nepravosti. Ostatně toho si byl vědom i sám Seydler. Jeho práci nemůžeme ani zařadit k pokusům o určování autorství pomocí matematických metod. Podstata atribuce anonymního textu totiž spočívá v tom, že se nejprve provede komplexní statisticko-stylistický rozbor anonymního textu (nepravého textu) a všech textů sloužících za porovnání (to se zjišťuje zpravidla na základě různých dodatečných informací, nejčastěji literárněhistorických). Pak se rozhoduje o autorství textu vzhledem k množině porovnávaných textů. U RKZ musí být nejprve provedeno časové zařazení, pak může být teprve přistoupeno ke klasické atribuci. Podstatná a přínosná je ale hlavní myšlenka Seydlerovy práce, a to hodnocení jazykových jevů matematickými metodami.

2.12.5 Další práce týkající se RKZ

V roce 1935 R. Jakobson ve své studii o verši RKZ[213] použil metody strukturního zkoumání verše a zdůraznil nutnost aplikace nových metod, např. matematické statistiky, právě v oblasti verše a lexika.

Spolu s Jakobsonem přistupoval k verši strukturalisticky i J. Mukařovský[214], který provedl srovnávací statistiku mezislovních předělů mezi 7. a 9. slabikou a statistiku předělů před všemi lichými a sudými slabikami v desetislabičných verších RKZ a dále v básni *Ot božieho těla*, v lidových písních Bartošovy sbírky, v Jungmannově, Máchově a Vrchlického díle, u Hanky (který se pokoušel o desetislabičný verš už roku 1814), u Lindy a Svobody. Na základě

[213] *K časovým otázkám nauky o verši I. Staročeský verš a Rukopisy.* Slovo a slovesnost 1, 1935, s. 46–53.

[214] *Obecné zásady a vývoj novočeského verše.* Přetisk v *Kapitolách z české poetiky 2*, 1948.

této studie došel k závěru, že desetislabičné verše RKZ jsou skladby Hankovy družiny.

Postupně se ukazovalo, že tyto exaktnější metody by mohly tuto dosud neuzavřenou kapitolu týkající se autorství RKZ obohatit, ne-li rozřešit[215]. Jelikož má určování autorství pomocí statistických metod značnou tradici i výsledky, byl v roce 1976 v *Ústavu pro jazyk český ČSAV* podroben podrobnému rozboru text *Rukopisu královédvorského a zelenohorského*. Výzkum byl prováděn Marií Těšitelovou, Ivou Nebeskou a Janem Králíkem. Z důvodů možného srovnávání se současnou češtinou byly jednotlivé texty excerpovány podle principů FSČ. Cílem tohoto rozboru nebylo určit autorství jistých osob, ale pomocí kvantitativních metod přinést nový materiál k otázce pravosti *Rukopisů*, který by napomohl k jejímu objektivnímu posouzení. Navíc by tento rozbor mohl posloužit k prověření a rozšíření nových metod týkajících se problematiky sporného autorství.

Nejprve bylo třeba připravit vhodný srovnávací materiál. První z nich (označený symbolem $DalC$) vznikl tak, že ke každé básni RK byl vybrán odpovídající text ze 14. století, tzn. z období, do kterého se RK hlásí. Druhý soubor (ObT) byl sestaven z textů z počátku 19. století, tzn. z období, kdy se *Rukopisy* objevily. Text z 10. století, který by korespondoval s textem RZ byl pominut (z nedostatku jazykového materiálu v tomto období). Texty v těchto porovnávacích souborech byly voleny tak, aby odpovídaly délkou (N) i tematikou textům RKZ. Navíc z 19. století byly vybrány jen texty vzniklé před rokem 1817, tzn. texty, které nemohly být ovlivněny objevením RKZ. Kvantitativní analýza byla tedy prováděna celkem na čtyřech souborech: RK, RZ, ObT, $DalC$.

Při analýze těchto textů byly voleny ty charakteristiky, které respektují specifiku RKZ. Protože jsou *Rukopisy* díla básnická, a podléhají tedy vlivu rytmických pravidel, veršům o daném počtu slabik apod., ukázalo se jako nevhodné pracovat např. s délkou věty či délkou slova. Základní jednotkou této analýzy se stalo slovo a určovala se například frekvence slov, slovních druhů, gramatických kategorií, syntaktické funkce jednotlivých slov apod.

Uveďme si zde pro názornost několik příkladů.

Příklad 2.10: Analýza předložek ukazuje na shody mezi RK a RZ a na rozdíly od srovnávaného materiálu. Tyto hodnoty nám ale neříkají, že je v RKZ bohatší slovník předložek, jen se některé předložky častěji opakují.

Text	Procentuální zastoupení předložek	Počet různých předložek
RK	13,26	26
RZ	14,66	5
ObT	9,79	30
$DalC$	9,55	20
FS	9,78	–

Tabulka 2.17: Zastoupení předložek v RKZ a v textech s nimi srovnávaných

[215] Viz Vašák, P.: *Statistika a sporné autorství*. Slovo a slovesnost 27, 1966, s. 364–370.

Příklad 2.11: Odpůrci pravosti poukazovali na výjimečné postavení spojky „*i*“ (vzhledem ke spojce „*a*“) v *RKZ*. Podívejme se na kvantitativní údaje týkající se těchto dvou spojek (viz tab. 2.18):

Text	Výskyt spojky „*i*“ (v %)	Výskyt spojky „*a*“ (v %)
RK	70,74	4,55
RZ	79,00	–
ObT	12,07	32,40
DalC	11,05	28,42

Tabulka 2.18: Výskyt spojky „*i*“ a „*a*“ v *RKZ* a textech s nimi srovnávaných

V obou sporných textech (*RK*, *RZ*) je „*i*“ nejčetnější spojkou, na rozdíl od *ObT* a *DalC*, kde je nejčetnější spojka „*a*“. V *RK* i *RZ* je v prvních 20 nejčastějších slovech jediná spojka, a to právě „*i*“, zatímco v *ObT* a *DalC* je spojek 5.

Příklad 2.12: Zajímavé je srovnání poměru V/N, kde V označuje rozsah slovníku, tj. počet různých slov, a N délku textu (ve slovech).

Text	Délka (N)	Slovník (V)	V/N (v %)
RK	5 892	1 608	27,29
RZ	573	261	45,55
ObT_{RK}	5 891	2 141	36,34
ObT_{RZ}	566	332	58,66
DalC	2 439	792	32,47
Fr. Hrubín (FSČ)	2 240	914	40,80
Fr. Halas: *Ladění* (FSČ)	5 249	2 078	39,59
V. Nezval: *Historický obraz I – III* (FSČ)	5 340	1 825	34,18

Tabulka 2.19: Porovnání poměru V/N, V a N v *RKZ* a textech s nimi srovnávaných

Vidíme, že slovník obrozeneckých textů korespondujících s *RK* je o 25 % větší než slovník *RK*. Může to být způsobeno například tím, že soubor ObT_{RK}[216] byl sestaven z textů různých autorů, zatímco *RK* má autora nejspíše pouze jednoho. Nebo se může jednat o charakteristický znak staršího období jazykového vývoje. Srovnáme-li ale zjištěné hodnoty s hodnotami pro soubor *DalC*, ukazuje se, že i zde je V/N vyšší, a tedy texty *DalC* mají blíže k textům ObT_{RK},

[216]Symbolem ObT_{RK} označujeme texty vzniklé v době národního obrození (do roku 1917) korespondující co do rozsahu materiálu a tematiky s texty *RK* (analogicky ObT_{RZ}).

než k textům *RK*, kam by s ohledem na dobu vzniku měly spíše patřit. Provedeme-li navíc srovnání se současným básnickým jazykem[217], jak byl zpracován ve FSČ, je vidět, že i zde jsou rozdíly mezi *DalC* a ObT_{RK} minimální.

Příklad 2.13: Pro vyjádření tzv. „*bohatství*" slovníku se v odborné literatuře pracuje s několika koeficienty. Zde byly počítány tyto: a) *velikost (R)*:

$$R_{80} = 100\frac{V}{0{,}80 \cdot N}, \qquad R_{70} = 100\frac{V}{0{,}70 \cdot N}.$$

Koeficient R_{80} vyjadřuje poměr počtu plnovýznamových slov k 80 % délky textu (a používá se v případě, kdy za plnovýznamová slova chápeme podstatná a přídavná jména, zájmena, číslovky, slovesa a příslovce), koeficient R_{70} vyjadřuje poměr počtu plnovýznamových slov k 70 % délky textu (v tomto případě se mezi plnovýznamová slova nepočítají zájmena a číslovky).

b) *koeficient rozptýlení (D)*:

$$D = 100\frac{V_1^{10}}{V},$$

který vyjadřuje poměr počtu plnovýznamových slov s frekvencí 1 až 10 (V_1^{10}) k celkovému rozsahu slovníku V.

c) *koeficient koncentrace (K)*:

$$K = 100\frac{N_1^{10}}{N},$$

který vyjadřuje poměr délky textu tvořeného prvními 10 nejfrekventovanějšími slovy N_1^{10} k celkové délce textu N.

Text	R_{80}	R_{70}	D	K
RK	34,50	39,42	83,27	18,86
RZ	56,94	65,07	98,47	21,12
ObT_{RK}	45,43	51,92	93,32	15,35
DalC	39,36	44,98	91,02	19,60
Halas	49,49	56,55	90,57	13,77
Nezval	42,72	48,82	88,05	13,28

Tabulka 2.20: Hodnoty koeficientů pro „*bohatost*" slovníku v *RKZ* a textech s nimi srovnávaných

První tři koeficienty (R_{80}, R_{80}, D) ukázaly, že *RK* stojí mimo oblast porovnávacích textů, čtvrtý koeficient (K) ukazuje blízkost *RK* době, do níž se hlásí (tj. *Dalimilově kronice*), což je ovlivněno vysokou frekvencí tzv. formálních slov.

[217] František Halas: *Ladění*, Vítězslav Nezval: *Historický obraz I – III.*

Příklad 2.14: Rovněž byla provedena podrobná analýza slovnědruhová (viz tab. 2.21). Ta ukázala, že procentuální zastoupení substantiv, adjektiv a sloves v *RK* se odlišuje od jejich zastoupení v *DalC* a má blíže k *ObT*. To by naznačovalo, že autorem byl zřejmě spisovatel a básník 19. století.

	RK				ObT_{RK}				*DalC*			
	lexémy		frekvence		lexémy		frekvence		lexémy		frekvence	
		v %		v %		v %		v %		v %		v %
I	575	35,36	2018	34,25	768	35,56	1810	30,70	255	33,20	639	26,20
II	297	18,27	588	9,98	410	18,98	746	12,65	119	11,98	193	7,91
III	23	1,41	523	8,88	29	1,34	583	9,89	26	3,26	354	14,52
IV	15	0,92	58	0,98	10	0,46	36	0,61	10	1,30	30	1,23
V	523	32,16	1076	18,26	655	30,32	1181	20,04	269	35,94	629	25,79
VI	132	8,12	422	7,16	200	9,26	482	8,18	67	8,33	143	5,86
VII	26	1,60	781	13,26	30	1,39	577	9,79	20	2,61	233	9,55
VIII	22	1,36	352	5,97	38	1,76	434	7,36	19	2,47	190	7,79
IX	11	0,68	71	1,21	7	0,32	13	0,22	3	0,39	5	0,21
X	2	0,12	3	0,05	13	0,60	33	0,56	4	0,52	23	0,94
Celkem	1626		5892		2160		5895		792		2439	

Tabulka 2.21: Analýza slovních druhů v *RKZ* a v textech s nimi srovnávaných

Shrneme-li na závěr předchozí statistická data zjištěná na materiálu *RKZ* a na textech s nimi korespondujících, vidíme, že v některých ukazatelích se shodují na jedné straně *RK* a *RZ*, na druhé *ObT* a *DalC*, a mezi oběma soubory je podstatný rozdíl, což by hovořilo o jisté výlučnosti *RK* a *RZ*. V jiných ukazatelích se zase *RKZ* blížil *ObT*, zatímco *DalC* vykazoval podstatně odlišné hodnoty, což by mohlo být vysvětleno dobou vzniku *RKZ*.

Závěr

Cílem předkládané publikace bylo seznámit čtenáře s historií matematické lingvistiky, disciplíny stojící na pomezí mezi matematikou a jazykovědou (v současnosti ale i informatikou a umělou inteligencí). Počátky matematické lingvistiky jako samostatné disciplíny klademe do padesátých let 20. století, zpravidla za tento mezník bývá považován rok 1957, kdy se konal VIII. mezinárodní lingvistický kongres v Oslo. Ačkoliv se tedy jedná o obor poměrně mladý, není v silách jediného člověka popsat tak široké spektrum problémů, které jsou její součástí. Matematická lingvistika se tradičně rozděluje na tři části podle toho, kterých matematických metod při studiu jazyka využívá:

1) kvantitativní lingvistika využívá kvantitativní matematické metody, jako je například statistika, teorie pravděpodobnosti aj.;
2) lingvistika algebraická pracuje s nekvantitativními matematickými metodami (teorie množin, algebra, teorie grafů, logika apod.) a
3) lingvistika počítačová využívá ke studiu přirozených jazyků počítače (v podstatě jde o praktickou aplikaci dvou předchozích částí matematické lingvistiky).

I v případě, že bychom se zaměřili pouze na jednu z těchto tří oblastí matematické lingvistiky, nebylo by možné sestavit ucelený přehled této oblasti pouze jedním člověkem. Proto bylo nutné stanovit jistá kritéria výběru témat pro tuto práci. Protože o historii matematické lingvistiky nacházíme v dostupných pramenech (viz např. [6], [49], [58], [66], [77]) jen ojedinělé a často strohé zmínky, bylo hlavním záměrem práce tyto informace nejprve shromáždit, ale především doplnit o skutečnosti zcela nové, původní. Pokud to jen bylo možné, pracovala jsem v těchto případech s primárními prameny. Dalším kritériem, podle něhož jsem volila témata příhodná ke zpracování, byla důležitost dané problematiky v rámci matematické lingvistiky. V práci se proto objevují témata, která hrála zásadní roli ve vývoji matematické lingvistiky (teorie markovských řetězců, tzv. Zipfovy zákony, metoda glottochronologie aj.) A konečně jsem chtěla představit přínos českých vědců tomuto oboru. Kromě toho, že bylo třeba shromáždit a zpracovat poměrně velké množství rozmanitého materiálu od časopiseckých článků až po příručky encyklopedického charakteru, komplikoval situaci ten fakt, že autory těchto prací byly vědci z různých, na první pohled neslučitelných oborů (např. matematika, fyzika, lingvistika, informatika, textologie aj.).

Výsledkem této snahy je předkládaná práce, která je vedle Úvodu a Závěru tvořena dvěma hlavními kapitolami. Kapitola první je spíše obecného charak-

teru a jejím cílem je čtenářům představit ve stručnosti celý obor matematická lingvistika. Dozvídáme se tu o okolnostech vzniku matematické lingvistiky jako samostatné disciplíny, o jejích cílech, metodách či aplikacích. Rovněž jsou zde blíže představeny jednotlivé obory matematické lingvistiky, na něž bývá dnes běžně dělena. Konkrétně se jedná o lingvistiku kvantitativní, algebraickou a počítačovou, i když hranice mezi nimi se stále více a více stírají. Nejzásadnější je ovšem kapitola druhá, v níž je čtenář podle výše uvedených kritérií seznámen s některými výsledky z historie a „prehistorie" matematické lingvistiky. Dočítáme se tu o poměrně jednoduchých aplikacích matematiky na jazyk (zpravidla na úrovni mystiky nebo hříčky), které byly využívány ve starověku a středověku. Ačkoliv v této době ještě zdaleka nemůžeme mluvit o matematické lingvistice, je zajímavé pozorovat, jakým způsobem se v této souvislosti formují počátky některých matematických disciplín, např. kombinatoriky aj. Navíc poznatky shromážděné v této kapitole se v tomto rozsahu nevyskytují v žádné z u nás běžně dostupných knih pojednávajících o matematické lingvistice. Skutečná matematická lingvistika (přesněji lingvistika kvantitativní) začíná až s vymezením pojmu frekvence jazykové jednotky (foném, písmeno, hláska, slovo, slabika aj.), proto je nemalá část práce věnována využití frekvence při tvorbě těsnopisných systémů, optimalizaci klávesnice psacího stroje, šifrování, kódování apod. Pozornost je zaměřena i na frekvenční a konkordanční slovníky, z nichž některé jsou blíže představeny. Samostatné oddíly jsou věnovány třem průkopníkům matematické lingvistiky, a to V. J. Bunjakovskému, A. A. Markovovi a americkému lingvistovi G. K. Zipfovi. Zejména kapitola pojednávající o ruském matematikovi A. A. Markovovi je poměrně rozsáhlá, neboť se snaží podrobně představit jeho zásadní práci [39], která byla vůbec první důslednou aplikací matematiky na jazyk a která navíc podnítila zájem matematiků o jazykovědnou problematiku. Rovněž sehrála význačnou roli v matematické teorii markovských řetězců. Při řešení některých textologických problémů (atribuce textu, stylometrie, sestavování stemmatu) se významnou měrou také uplatňovala matematika, proto jsou tyto textologické problémy, které se zasloužily o rozvoj matematické lingvistiky, představeny podrobněji. Nelze opominout samozřejmě ani Shannonovu a Weawerovu teorii informace a některé její pojmy, zejména entropii a redundanci. Zajímavou aplikací matematiky je metoda nazývaná glottochronologie (nebo také lexikostatistika), která na základě statistiky zjišťuje stáří jazyka či jazyků. Její autor Morris Swadesh se inspiroval tzv. rozpadovým zákonem, jenž na základě rozpadu radioaktivních jader uhlíku určuje stáří organických látek. Poslední tři oddíly představují přínos českých vědců matematické lingvistice. Nejprve je pozornost věnována spisu *Janua linguarum reserata*, jehož autorem je Jan Amos Komenský. Dále je pozornost upřena na dvě statistiky uveřejněné roku 1831 v časopise *Krok*, prvním českém vědeckém časopise. První z nich je statistika hlásková, jejímž cílem je vyvrátit tvrzení, že čeština je, podobně jako němčina, jazyk těžký a tvrdý na výslovnost. Ukazuje se, že se jedná o jednu z prvních českých hláskových statistik, či spíše o jednu z prvních hláskových statistik v rámci slovanské jazykovědy vůbec. Druhá statistika zase porovnává počet slov v určitých jazycích a snaží se dokázat, že český jazyk je srovnatelný s ostatními jazyky co se týče slovní zásoby. Při pří-

pravě této podkapitoly jsem náhodou narazila na zmínku o existenci statistiky, kterou si pro potřeby studia problematiky libozvučnosti sestavoval Jan Kollár. Nejednalo se sice o statistiku hláskovou či lexikální, ale nejspíše o statistiku grafémů, přesto ji lze podle dostupných informací časově vymezit rokem 1817 – byla by tedy o celých 14 let starší, než uvedené statistiky z časopisu *Krok*. Zdá se, že tato oblast matematické lingvistiky není zdaleka dostatečně prozkoumaná a že by bylo jistě zajímavé hledat další takové příklady užití matematických metod v lingvistice. A konečně poslední oddíl je věnován matematikovi, fyzikovi a astronomovi Augustinu Seydlerovi, který se svým článkem [57], v němž se pomocí počtu pravděpodobnosti snažil dokázat, že *Rukopisy královédvorský a zelenohorský* jsou novodobé podvrhy, zapojil v 80. letech 19. století do tzv. rukopisných sporů, které hýbaly celou tehdejší společností.

Doufám, že díky předložené práci mohou čtenáři zajímající se o otázky týkající se historie matematické lingvistiky získat dobrý přehled o této problematice. Hlubší zájemci pak zde mohou nalézt (v četných poznámkách pod čarou) množství odkazů na další literaturu (primární i sekundární). Kniha je psána takovým způsobem, aby byla srozumitelná a užitečná nejen matematikům (zejména z řad středoškolských učitelů a vysokoškolských studentů, kteří by zde mohli nalézt inspiraci pro svou práci), ale i zájemcům o jazyk či literaturu.

Rejstřík

A

Abulafia, Abraham, 19, 20
Achmanova, O. S., 11
Adam, Daniel z Veleslavína, 106
Agrippa z Nettesheimu, Heinrich Cornelius, 20
Ajdukiewicz, Kazimierz, 10
Aksakov, Sergej T., 54, 60
Akvinský, Tomáš, 20
al-Kindī, Abú-Yūsuf Ya'qūb ibn Ishāq, 30
Alžběta I., 69
Anchoranus, Joannes, 105
Andrejev, N. D., 11
Anubis, 69
Apollinare, Gillaume, 18
Aristoteles, 20, 21
Arnauld, Antoine, 27
Arnold, W. W., 104
Axamit, Ignác, 124
Ayers, L. P., 31

B

Bacon, Francis, 21, 69, 72, 73
Bacon, Roger, 15, 21
Bakonyi, H., 34
Balcescu, N., 74
Bales, Peter, 42
Bar-Hillel, Yehoshua, 9, 10
Bareš, Jan, 44
Bartoš, František, 137
Bartušková, M., 40
Barák, Josef, 78
Bath, William, 105
Baum, Antonín, 120
Bayes, Thomas (1702–1761), 77
Becket, Thomas, 42
Belevitch, Vitold, 66
Benkö, L., 31
Bezruč, Petr, 41, 69
Bečka, Josef Václav, 40, 41
Bečvář, J., 21
Blažek, Václav, 104
Bojnanský, Ladislav, 43
Bony, 37
Borges, Jorge Luis, 27
Bou, J. Rodriguez, 39
Brandl, Vincenc, 121
Bright, Timothy, 42
Brinegar, C., 76, 77
Brodin, D., 38
Bruno, Giordano, 25, 26
Brzobohatý, František, 44
Brödel (pastor), 126
Buchanan, M. A., 38
Bunjakovskij, Viktor Jakovlevič, 2, 8, 16, 30, 48–51, 137, 144
Bédier, J., 81
Bělohoubek, Antonín, 121, 130
Büdinger, Max, 119

C

Campbell, L., 69
Caramuel z Lobkovic, Jan, 27
Carlemann, 77
Catilinus, 42
Caton, 42
Cicero, Marcus Tullius, 41
Clavius, Chrisoph, 26

Cohen, M., 8
Cordovero, Moše, 19
Coseriu, E., 104
Cotteret, J. M., 38
Crăciun, C., 10, 11
Curry, H. B., 10
Czekanowski, Jan, 17, 82

Č

Čapek, Karel, 40, 66
Čebyšev, Pafnutij Lvovič, 51–54
Čech, Eduard, 77, 78
Čejka, Mirek, 104
Čigorin, M. I., 52
Čuprovyj, A. A., 54
Čáda, F., 70
Čáp, Jindřich, 46

D

Dalimil, 140
Davidovitch, C., 38
de Courtenay, Jan Baudouin, 8, 16
de Gaulle, Charles, 38
de Saussure, Ferdinand, 108
Dee, John, 25
della Mirandola, Pico, 25
della Porta, Giambattista, 26
Descartes, René, 27
Dewey, G., 36
Dittenberger, W., 69
Dobrovský, Josef, 110, 119
Dobrušin, R. L., 10, 11, 61, 95
Dolanský, Ladislav, 121
Domalíp, Karel, 127

E

Eldridge, R. C., 31, 64
Ellegård, Alvar, 72, 74, 75, 77
Embletonová, Sheila, 104
Enders, Julius, 122
Erk, H., 35
Estoup, Jean-Baptiste (1868–1950), 43, 63

F

Faigl, Josef, 44
Faucett, L., 36
Feifalik, Julius, 119
Fitialov, S. J., 10, 11
Flajšhans, Václav, 121
Fodor, Jerry Alan, 10
Formánková-Kubínová, A., 38
Fourier, Jean Baptiste Joseph, 48
Francis, Philip, 75, 76
Francis, W. N., 36
Friedman, E. A., 77
Frobenius, Ferdinand Georg, 51
Froger, J., 81, 82, 87
Frumkina, R. M., 11
Fuchs, Eduard, 21
Fuchs, W., 98
Förstemann, Ernst, 16, 109
Fügner, Jindřich, 44, 48

G

Gabelsberger, Franz Xaver, 43, 44
Gamble, W., 31, 137
Gauss, Carl Friedrich, 124, 130
Gebauer, Jan, 120, 121, 128–133, 136
Gerson, Jean, 73
Gerwig, G. A., 73
Ginzburg, R. Z., 36
Gladkij, A. V., 10, 11
Goethe, Johann Wolfgang, 31, 70, 118
Gogol, Nikolaj Vasiljevič, 60
Goll, Jaroslav, 121
Gorkij, Alexej Maximovič, 52
Gruss, Gustav, 126, 127
Grzybek, Peter, 109, 116
Grégr, Julius, 120, 136
Gröber, G., 81
Guiraud, P., 66, 73
Guldin, Paul, 26
Gába, František, 44

H

Hadamard, Jacques, 54

Halas, František, 140
Hamerský, Th., 40
Hamilton, A., 76, 77
Hanka, Václav, 110, 111, 118–121, 129, 130, 134–138
Harary, F., 11
Harkin, D., 104, 105, 108
Harsdörffer, Georg Philipp, 25
Hattala, Martin, 8, 120, 121, 129
Havránek, Bohuslav, 45
Heger, Hynek Jakub, 44
Henmon, V. A. C., 38
Herdan, G., 8, 66, 68, 73, 74, 93
Herout, Alois, 45–47
Hiller, Heinrich, 27
Hjelmlev, Louis, 11
Hockett, Ch. F., 93
Hoffmann, L., 34, 35
Holas, Alois, 40
Holinová, Anna, 78
Horn, E., 36
Hornstein, Karel, 124, 125, 127
Horák, Josef, 46
Hostinský, B., 77
Hostinský, Otakar, 121
Hoze, V. G., 39
Hradilík, František, 45
Hume, David, 130
Huygens, Christiaan, 124

CH

Chang-Rodriguez, E., 39
Chidekel, S. S., 36
Chmela, Josef, 112
Chomsky, Noam, 9, 10, 16, 27

I

Isou, Isidor, 18
Ivanov, Miroslav, 69, 130

J

Jagi, Vatroslav, 120
Jakobson, Roman, 45, 93, 137
Janovská, Helena, 47
Jarník, V., 77
Jay, J., 76
Jehuda z Wormsu, Eleazar Ben, 19
Jelínek, J., 40
Jireček, Hermenegild, 112, 121
Jireček, Josef, 112, 121
Jodl, Jan, 111
Josef, František (císař), 127
Josselson, H. H., 37
Joyce, James, 64
Juilland, Alphonse, 38, 39
Junek, Jaroslav, 47
Jungmann, Antonín, 110, 111
Jungmann, Josef, 110, 111, 119, 134, 137
Junius, 72, 74–76

K

Kalousek, Josef, 121, 136
Katz, Jerrold, 10
Kempenský, Tomáš, 73
Kepler, Johannes, 127
Kindl, Jan, 125
Kiselevskij, A. I., 34, 37
Klimov, G. A., 104
Knowles, J., 31
Kollár, Jan, 111, 145
Kolmogorov, Andrej Nikolajevič, 10, 11, 54, 97
Kolovrat, hrabě, 118
Koláček, František, 125, 127
Komenský, Jan Ámos, 2, 13, 16, 17, 25, 27, 28, 42, 43, 104–108, 144
Komárek, Miroslav, 136
Konrad, A., 44
Kopecký, Jan Dalibor, 43
Kopitar, Bartoloměj, 119
Korejs, J., 40
Korejsová, H., 40
Korkin, Alexandr Nikolajevič, 51
Kouble, Josef, 111
Kovář, Josef, 118
Kramář, Josef, 111
Kraus, Arnošt, 121

Kraus, Jiří, 45, 46
Krondl, Antonín, 44
Král, Josef, 121
Králík, Jan, 138
Králík, O., 78
Krámský, Jiří, 9
Kuha, David, 119
Kulaginová, O. S., 10, 11
Kumprecht, K., 34, 38
Kuraszkiewicz, W., 73
Kučera, Henry, 36
Käding, F. W., 8, 31–34, 36, 43

L

Lachmann, K., 81
Lagrange, Joseph Louis, 124
Laiske, Miroslav, 119
Lamanskij, Vladimír Ivanovič, 120
Lamprecht, Arnošt, 104
Lancelot, C., 27
Laplace, 48, 124, 130
Lees, Robert B., 101
Leibniz, G. W., 2, 16, 21, 25, 27
Leskien, A., 129
Lesskis, G. A., 73
Levison, M., 71, 77
Libby, Willard Frank, 102
Lichačev, D. S., 81
Linda, Josef, 118, 119, 137
Ljapunov, A. M., 53
Llull, Ramón (viz také Lullus), 21
Lodwick, Francis, 16, 25
Lorge, I., 36
Ludvík, Josef Myslimír, 112
Ludvíková, M., 45
Lullus, Raimundus, 2, 15, 21–26
Lutosawski, W., 71

M

Macek, E., 78
Mach, Ernst, 124, 126, 127
Madison, J., 76, 77
Maki, Itsu, 36
Malíř, F., 37
Mandelbrot, Benoît, 17, 64–66
Marcus, Solomon, 10, 11
Marek, Antonín, 111, 112, 119
Mareš, František, 122
Markov, Andrej Andrejevič, 2, 8, 16, 50–55, 57, 59–61, 65, 71, 144
Markov, Andrej Andrejevič mladší, 52
Markov, Andrej Grigorjevič, 50
Markov, Michail Andrejevič, 50
Markov, Pavel Andrejevič, 50
Markov, Petr Andrejevič, 50
Markov, Vladimír Andrejevič, 50, 51
Markovová, Anna Josifovna, 50
Markovová, Jekatěrina Andrejevna, 50, 51
Markovová, Jevgenie Andrejevna, 50, 51
Markovová, Lýdie Andrejevna, 50, 51
Markovová, Marie Andrejevna, 50
Markovová, Naděžda Petrovna, 50
Masaryk, Jan, 128
Masaryk, Tomáš Garique, 117, 120, 121, 128–130
Masaryková, Alice, 128
Mathesius, Vilém, 9, 18, 45
Matoušek, Karel, 43, 46
Matula, Miloš, 43, 45, 46
Mačák, Karel, 21
Meier, H., 33, 34
Melčuk, I. A., 11
Mendenhall, T. C., 72, 73, 76
Mersenne, Marin, 26
Meyer-Eppler, Werner, 61, 95
Mikuláš II., 52
Mikulík, Svojmír, 45–47
Miller, G. A., 77
Mitrík, Jozef, 39, 43
Mittag-Lefler, Gösta, 77
Mochinger, J., 105
Moreau, R., 38
Morgan, B. Q., 33, 34
Morozov, N. A., 60, 61, 70, 71
Morse, Samuel, 8, 16, 29, 30, 99

Morton, A. Q., 71, 73, 77
Mosteller, F., 76, 77
Mrkos, Antonín, 128
Mukařovský, Jan, 137
Muller, Ch., 8
Myslbek, J. V., 118
Mácha, Karel Hynek, 30, 69, 83, 89, 118, 137

N

Nebeská, Iva, 138
Nebeský, Ladislav, 10, 11
Nebeský, Václav Bolemír, 119
Neiescu, I., 74
Neruda, Jan, 69, 78, 123
Newman, E. B., 77
Newton, Isaac, 124, 126
Nezval, Vítězslav, 18, 140
Nirasawa, T., 77
Novotná, Petra, 103, 104
Novotný, Eduard, 44
Novotný, F., 70
Novotný, Josef, 44
Novák, Pavel, 6, 61, 66, 93
Náprstek, Vojtěch, 127
Němcová, Božena, 83, 90

O

Olbracht, Ivan, 67
Osiris, 69
Otto, Jan, 121, 128

P

Padučeva, J. V., 11
Palacký, František, 110–112, 119
Palkovič, Jiří, 111, 117
Paper, H., 11
Paris, G., 81
Patera, Adolf, 120, 121
Paul, Herman, 16
Pelcl, František Martin, 110
Perebejnos, V. I., 36
Petruševič, Malorus Antonij, 120
Petrásek, Jan, 46
Piotrovskij, R. G., 34, 37
Plath, W., 6, 8, 63
Platón, 69, 70, 137
Poisson, Siméon Denis, 48, 130, 131
Polívka, Jiří, 121
Posse, Konstantin Alexandrovič, 51
Presl, Jan Svatopluk, 109–112, 116
Presl, Karel Bořivoj, 110, 111
Prusík, František Xaver, 121
Pujmanová, Marie, 66
Purkyně, Jan Evangelista, 110–112
Puškin, Alexandr Sergejevič, 16, 31, 54, 55, 60
Pánek, Augustin, 131
Päumann, Anton von, 119
Píč, Josef Ladislav, 121
Příhoda, Václav, 18, 40

Q

Quentin, Dom, 81
Quetelet, Lambert Adolphe Jacques, 130

R

Rabelais, François, 18
Remak, Robert, 51
Revzin, I. I., 10, 11
Rešel, Tomáš, 106
Rhenius, M. I., 107
Ritter, C., 70
Roberts, A. Hood, 36
Rosa, Václav, 106
Rosengrenová, I., 35
Russo, Al., 74

Ř

Řeháček, L., 93

S

Sankin, A. A, 36
Scherrerová, Marie, 78
Schleicher, August, 16, 109
Schneider, Z., 105
Schnuss, 131

"kniha˙final" — 2012/7/27 — 22:04 — page 152 — #158

Schrey, Ferdinand, 32, 43
Sedláček, Josef, 40, 45, 46
Sedláček, Josef Vojtěch, 110
Seidel, J., 105
Seifert, Jaroslav, 18
Selenus, Gustavus, 27
Seydler, Augustin, 2, 9, 17, 77, 117, 121–137, 145
Seydler, Jan, 128
Seydler, Jan Nepomuk, 123
Seydler, Jaroslav, 128
Seydler, Karel, 123, 124, 128
Seydlerová, Anna, 123, 124, 128
Seydlerová, Anna (roz. Hromátková, ovdovělá Jandoušová), 123
Seydlerová, Anna (roz. Weyrová), 127
Seydlerová, Anna Amidea Helena, 128
Seydlerová, Antonie (roz. Suková), 123, 128
Seydlerová, Antonie Anežka, 123, 124
Seydlerová, Antonie Marie, 128
Seydlerová, Marie, 123, 124, 128
Seydlerová, Marie (roz. Stehlíková), 128
Sgall, Petr, 10
Shakespeare, William, 69, 72, 73
Shannon, Claude Elwood, 7, 66, 92, 93, 95, 97, 98, 144
Shannon, G., 61
Shermann, L. A., 73
Simon, H. A., 66
Skuherský, Rudolf, 124
Sládek, Josef Václav, 127
Smetana, Bedřich, 118
Snodgrass, Q. C., 76
Sochockij, Julian V., 51
Sokolov, N. N., 43
Somers, H. H., 66
Spang-Hansen, H., 6
Stan, A., 74
Stan, I., 74
Starostin, Sergej, 104
Stolze, Wilhelm, 32, 43
Strouhal, Čeněk, 126, 127
Struve, P. B., 50
Studnička, František Josef, 112, 124, 125, 127
Stáně, Alois, 44
Střížová, A., 37
Sušil, František, 112
Svoboda, Václav Alois, 112, 119, 137
Swadesh, Morris, 101–104, 144

Š

Šafařík, Pavel Josef, 110, 111, 119, 121, 130
Šafařík, Vojtěch, 121, 130
Šaumjan, S. K., 9, 10
Šembera, Alois Vojtěch, 120
Šmilauer, Vladimír, 6, 9, 18
Štejnfeldtová, E. A., 37
Štorch, Karel Boleslav, 112
Šumavský, Franta Josef, 112
Šír, František, 112

T

Tablic, Bohuslav, 112
Tauer, František, 44
Taylor, Brook, 66
Thorndike, E. L., 35, 36, 38
Tilbury, John of, 42
Tiro, Marcus Tullius, 41
Tolstoj, Lev Nikolajevič, 53, 60
Tomek, Václav Vladivoj, 119
Tomíček, Jan Slavomír, 112
Tormasov, A. P., 48
Trithemius, Johannes, 26
Trnka, Bohumil, 9, 17, 18, 33, 41, 44–46, 67
Trubeckoj, Nikolaj Sergejevič, 45
Truhlář, Josef, 121
Trut, 69
Turgeněv, I. S., 60
Twain, Mark, 76
Těšitelová, Marie, 9, 17, 32, 40, 66–68, 73, 75, 104, 108, 138

U

Uldall, Hans J., 11

V

Vachek, Josef, 9, 18
Valvatjevová, Jekatěrina Alexandrovna, 51, 52
Valvatjevová, Marie Ivanovna, 52
Vander Beke, G. E., 38
Vančura, Jindřich, 121
Vašek, Antonín, 120, 129
Vašák, Pavel, 72, 78, 79, 81–83, 89, 138
Vetešník, František, 112
Vincenc, Pavel Žák, 112
Vinogradov, V. V., 71
Vlček, Jaroslav, 121
Vlček, Vácslav, 128
Vodička, F., 78
Vojtěch, Viktorin, 122
Volkova, N. O., 36
Vrchlický, Jaroslav, 137
Vrzalík, Jan, 122
Vrátný, Jaroslav, 47

W

Wallace, D. L., 76, 77
Ward, Seth, 25
Weawer, Warren, 92, 93, 144
Webster, John, 16, 25
Weinek, Ladislaus, 125
Wengerscius, A., 105
West, M., 36
Weyr, Eduard, 127
Weyr, Emil, 127
Weyr, František, 127
Whatmough, Joshua, 6
Whitney, W. D., 8, 16, 109
Wiener, Norbert, 7, 54, 93
Wilkins, John, 16, 25, 27
Williams, C. B., 73, 74, 76
Willis, John, 42
Wolf, D., 36
Wolf, František, 77
Woronczak, J., 73
Wängler, H.-H., 34

Y

Yule, G. U., 73, 74, 77–79

Z

Zasorinová, L. N., 37
Zeller, E., 70
Zichová, J., 133–135
Zimmermann, Jan, 118
Zipf, George Kingsley, 2, 8, 16, 17, 43, 62–68, 144
Zolotarev, Jegor Ivanovič, 51

Literatura

[1] Bečvář, J., Fuchs, E. (ed.): *Matematika v 16. a 17. století*. Prometheus, Praha 1999. ISBN 80-7196-150-7.

[2] Blažek, V., Novotná, P.: *Glottochronologie a její aplikace pro slovanské jazyky*. In: Sborník prací Filozofické fakulty brněnské univerzity A 53. Masarykova univerzita, Brno 2005, s. 51–81.

[3] Bunjakovskij, V. Ja.: *O vozmožnosti vvedenija opredelitelnych mer doverija k rezultatam nekotorych nauk nabljudatelnych, i preimuščestvenno statistiki*. Sovremennik 1847, tom 3, otdel 2, s. 36–49.

[4] Confortiová, H.: *Český retrográdní slovník*. Naše řeč 72, 1989.

[5] Čermák, Fr., Blatná, R.: *Manuál lexikografie*. Nakladatelství a vydavatelství H&H, Jinočany 1995. ISBN 80-85787-23-7.

[6] Černý, J.: *Dějiny lingvistiky*. Votobia, Olomouc 1996. ISBN 80-85885-96-4.

[7] Červená, V.: *O lexikálním archívu Ústavu pro jazyk český ČSAV*. Naše řeč 64, 1981, s. 14–22.

[8] Červenka, J. (red.): *Johannis A. Comenii: Ianua linguarum reserata (Editio synoptica et critica quinque authenticos textus Latinos necnon Janualem Comenii textum Bohemicum continens. Praecedit commentatio de Comenii Janua linguarum reserata)*. SPN, Praha 1959, 344 s.

[9] *Čeští spisovatelé 19. a počátku 20. století*. Československý spisovatel, Praha 1973.

[10] Dobrušin, R. L.: *Matematické metody v lingvistice*. In: Teorie informace a jazykověda, Praha 1964, s. 24–53.

[11] Eco, U.: *Hledání dokonalého jazyka v evropské kultuře*. Nakladatelství Lidové noviny, Praha 2001. ISBN 80-7106-389-4.

[12] Eco, U.: *Jak napsat diplomovou práci*. Votobia, Olomouc 1997. ISBN 80-7198-173-7.

[13] Enders, J.: *Jazykovědný rozbor Rukopisu Královédvorského, Zelenohorského a dalších staročeských textů s nimi spojovaných.* Neklan 1993. ISBN 80-900884-6-5

[14] Filipec, J.: *Z prehistorie a historie Ústavu pro jazyk český ČSAV.* Naše řeč 64, 1981, s. 225–232.

[15] Gebauer, J.: *Potřeba dalších zkoušek rukopisu Královédvorského a Zelenohorského.* Athenaeum, roč. III (1886), 152–168.

[16] Glivický, J.: *Profesor Dr. František Wolf šedesátiletý.* Časopis pro pěstování matematiky 90, 1965, s. 247–249.

[17] Grodzenskij, S. J.: *Andrej Andrejevič Markov (1856–1922).* Izdatelstvo Nauka, Moskva 1987.

[18] Grzybek, P.: *A Very Early Slavic Letter Statistic: Jan Svatopluk Presl (1791–1849).* In: Glottometrics 12, 2006, s. 88–91.

[19] Grzybek, P.: *Viktor Jakovlevič Bunjakovskij (1804–1889). A forerunner of quantitative linguistics.* In: Glottometrics 6, 2003, s. 103–106.

[20] Havránek, B., Horálek, K.: *Osmý mezinárodní kongres lingvistů v Oslo.* Slovo a slovesnost 19, 1958, s. 47–52.

[21] Hostinský, B.: *Počet pravděpodobnosti, první část.* Jednota československých matematiků a fyziků, Praha 1950.

[22] Hostinský, B.: *Počet pravděpodobnosti, druhá část.* Přírodovědné nakladatelství, Praha 1950.

[23] Hrabák, J. a kol.: *Dějiny české literatury I (Starší česká literatura).* Nakladatelství Československé akademie věd, Praha 1959.

[24] Hrabák, J., Jeřábek, D., Tichá, Z.: *Průvodce po dějinách české literatury.* Panorama, Praha 1978.

[25] Jelínek, J., Bečka, J. V., Těšitelová, M.: *Frekvence slov, slovních druhů a tvarů v českém jazyce.* Praha 1961.

[26] Komenský, J. A.: *Dvéře jazyků odevřené.* In: Opera omnia [11], Academia, Praha 1973.

[27] Králík, S. a kol.: *Otázky současné komeniologie.* Academia, Praha 1981.

[28] Kratochvil, A.: *Stručné dějiny československého těsnopisu.* Praha 1937.

[29] Kraus, J.: *Frekvenční slovník slovenštiny.* Naše řeč 53, 1970, s. 273–277.

[30] Kraus, J.: *Stroje v jazykovědě.* Naše řeč 51, 1968, s. 295–297.

[31] Kraus, J.: *Vztah těsnopisu a jazykovědy.* Naše řeč 55, 1972, s. 1–9.

[32] Lehár, J., Stich, A., Janáčková, J., Holý, J.: *Česká literatura od počátků k dnešku*. Nakladatelství Lidové noviny, Praha 1998. ISBN 978-80-7106-963-8.

[33] Ludvíková, M., Kraus, J.: *Kvantitativní vlastnosti soustavy českých fonémů*. Slovo a slovesnost 27, 1966, s. 334–344.

[34] Mačák, K.: *Počátky počtu pravděpodobnosti*. Prometheus, Praha 1997. ISBN 80-7196-089-6.

[35] Mačák, K.: *Vývoj teorie pravděpodobnosti v českých zemích do roku 1938*. Ústav pro soudobé dějiny AV ČR, Praha 2005. ISBN 80-7285-052-5.

[36] Majstrov, L. J.: *Teorija verojatnostej, istoričeskij očerk*. Nauka, Moskva 1967.

[37] Markov, A. A.: *Izbranyje trudy (Teorija čisel, teorija verojatnostej)*. Izdatelstvo Akademii nauk SSSR, Leningrad 1951.

[38] Markov, A. A.: *Ob odnom primenenii statističeskogo metoda*. In: Izvestija Imperatorskoj Akademii nauk, série 6, díl 10, 1916, č. 4, s. 239–242.

[39] Markov, A. A.: *Primer statističeskogo issledovanija nad tekstom "Jevgenija Onegina", illjustrirujuščij svjaz ispytanij v cep*. In: Izvestija Imperatorskoj Akademii nauk, série 6, díl 7, 1913, č. 3, s. 153–162.

[40] Mazlová, V.: *Jak se projevuje zvuková stránka češtiny v hláskových statistikách*. Naše řeč 30, 1946, s. 101–111, 146–151.

[41] Neruda, J.: *Podobizny IV (1889–1891)*. Knihovna klasiků, Státní nakladatelství krásné literatury, hudby a umění, Praha 1957.

[42] Novotná, P.: *K aplikaci glottochronologie na (nejen) indoevropské jazyky*. Rigorózní práce, Brno 2008.

[43] Otruba, M. (red.): *Rukopisy královédvorský a zelenohorský – Dnešní stav poznání*. Academia, Praha 1969.

[44] Otruba, M., Vašák, P.: *První užití matematiky v české jazykovědě*. Pokroky matematiky, fyziky a astronomie, ročník XIX, 1974, str. 86–91.

[45] Panevová, J.: *Lexikální zásoba češtiny v číslech*. Naše řeč 60, 1977, s. 31–34.

[46] Panevová, J.: *Využití strojů při zkoumání češtiny*. Slovo a slovesnost 23, 1962, s. 222–227.

[47] Petr, J., Tylová, M.: *M. Těšitelová (Bibliografický soupis)*. Praha 1986.

[48] Petrásek, J.: *Dějiny těsnopisu*. Státní pedagogické nakladatelství Praha, 1973.

[49] Plath, W.: *Matematická lingvistika.* In: Teorie informace a jazykověda, Praha 1964, s. 24–53.

[50] Pohorský, M.: *Intimní Karel Hynek Mácha.* Československý spisovatel, Praha 1993. ISBN 80-202-0393-1.

[51] Pohorský, M. a kol.: *Dějiny české literatury III (Literatura druhé poloviny devatenáctého století).* Nakladatelství Československé akademie věd, Praha 1961.

[52] Presl, J. S.: *Český jazyk s jinými jazyky z ohledu bohatosti slov porovnaný.* In: Krok, 1831, s. 623.

[53] Presl, J. S.: *Srovnání spoluhlásek a samohlásek v češtině, vlaštině a němčině.* In: Krok, 1831, s. 624–625.

[54] Rulfová, M.: *Kniha o určování autorství.* Naše řeč 64, 1981.

[55] Sedlačíková, B.: *Matematická lingvistika.* Učitel matematiky 10, 2001–2002, s. 30–36, 80–88, 174–181, 226–234.

[56] Seydler, A.: *Dodatek k mé úvaze o pravděpodobnosti.* Athenaeum, roč. III (1886), 446–449.

[57] Seydler, A.: *Počet pravděpodobnosti v přítomném sporu.* Athenaeum, roč. III (1886), 299–308.

[58] Sgall, P. a kol.: *Cesty moderní jazykovědy.* Orbis, Praha 1964.

[59] Stich, A.: *Český frekvenční slovník.* Naše řeč 46, 1963, s. 80–86.

[60] Strouhal, V.: *Dr August Seydler I. Nástin životopisný.* Časopis pro pěstování matematiky a fysiky 21, 1892, s. 193–202.

[61] Strouhal, V.: *Dr August Seydler IV. Závěrek.* Časopis pro pěstování matematiky a fysiky 21, 1892, s. 212–217.

[62] Struve, P. B.: *Kto pervyj ukazal na primenenije statistiki k filologičeskim issledovanijam.* Izvestija rossijskoj akademii nauk, 1918, s. 1317–1318.

[63] Šmilauer, V.: *Kvantitativní lingvistika a Marie Těšitelová.* Naše řeč 64, 1981, s. 195–201.

[64] Šolcová, A., Šolc, M.: *Profesor August Seydler, astronom, fyzik a humanista.* Pokroky matematiky, fyziky a astronomie 42, 1997, 188–209.

[65] Těšitelová, M.: *J. A. Komenský a racionalizace osvojování si slovní zásoby.* Slovo a slovesnost 31, 1969, s. 187–191.

[66] Těšitelová, M.: *Kvantitativní lingvistika.* SPN, Praha 1987.

[67] Těšitelová, M.: *O využití výsledků kvantitativní lingvistiky.* Naše řeč 70, 1987, s. 225–236.

[68] Těšitelová, M.: *Otázky frekvence slov (zvláště v češtině)*. Univerzita Karlova, Praha 1951.

[69] Těšitelová, M.: *100 let od vydání prvního frekvenčního slovníku*. Naše řeč 80, 1997, s. 190–194.

[70] Těšitelová, M. a kol.: *Kvantitativní charakteristiky současné češtiny*. Praha 1985.

[71] Těšitelová, M. a kol.: *O češtině v číslech*. Academia, Praha 1987.

[72] Těšitelová, M., Nebeská, I., Králík, J.: *On the quantitative characteristics of the Czech texts of disputed authorship – RKZ*. In: PSML 5, 1976, s. 119–147.

[73] Těšitelová, M., Petr, J., Králík, J.: *Retrográdní slovník současné češtiny*. Praha 1986.

[74] Trnka, B.: *K výstavbě fonologické statistiky*. Slovo a slovesnost 11, 1948–49, s. 59–64.

[75] Trnka, B.: *Kvantitativní lingvistika*. Časopis pro moderní filologii 34, 1951, s. 66–74.

[76] Trnka, B.: *Pokus o vědeckou teorii a praktickou reformu těsnopisu*. Praha 1937.

[77] Vašák, P.: *Matematika, exaktnost a literatura*. Čs. spisovatel, Praha 1986.

[78] Vašák, P.: *Metody určování autorství*. Praha 1980.

[79] Vašák, P.: *Statistika a sporné autorství*. Slovo a slovesnost 27, 1966, s. 364–370.

[80] Vašák, P., Čornej, P., Kolár, J., Mazáčová, S., Pražák, E., Víšková, J.: *Textologie (teorie a ediční praxe)*. Praha 1993. ISBN 80-7066-638-2

[81] Vodička, F. a kol.: *Dějiny české literatury II (Literatura národního obrození)*. Nakladatelství Československé akademie věd, Praha 1960.

[82] Zichová, J.: *Teorie pravděpodobnosti a rukopisný spor*. Pokroky matematiky, fyziky a astronomie 49, 2004, str. 95–103.

[83] Zubatý, J.: *Vědecké methody realismu zvláště v otázce Rukopisů*. Naše řeč 11, 1927, s. 149n.

Internetové zdroje

http://astro.mff.cuni.cz/historie/defaultse5.html (4. 11. 2011)

http://cs.wikipedia.org/wiki (25. 9. 2010)

http://is.muni.cz/el/1421/jaro2008/CJBB105/matematicka_lingvistika.pdf ?fakulta=1421;obdobi=3704;kod=CJBB105 (5. 10. 2010)

http://kix.fsv.cvut.cz/rkz (4. 11. 2011)

http://sirrah.troja.mff.cuni.cz/ mira/astronomie/Hvezdarna/historie/ historie.html#seydler (4. 11. 2011)

http://www.jcmf.cz/hist.html (15. 9. 2010)

http://www.muzeumtgm.cz/cz/osobnosti/Podrobny-zivotopis (4. 11. 2011)

http://www.nuov.cz/statni-tesnopisny-ustav (6. 11. 2011)

http://www.phil.muni.cz/fil/scf/komplet/seydlr.html (6. 11. 2011)

http://www.pozitivni-noviny.cz/cz/clanek-2009060022 (6. 11.2011)

http://www.quido.cz/osobnosti/seydler.html (6. 11. 2011)

http://www.tesnopis.cz/ (6. 11. 2011)

http://www-history.mcs.st-andrews.ac.uk/Biographies/Markov.html (6. 11. 2011)

http://ziva.avcr.cz/ (6. 11. 2011)