2.5 Andrej Andrejevič Markov

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 50–61.

Persistent URL: http://dml.cz/dmlcz/402320

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 nepreryvnych 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.