2.7 Text linguistics

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (English). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 69–92.

Persistent URL: http://dml.cz/dmlcz/402322

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 paternitătii Cintării Rominiei*. Cercetări de lingvistică 1963, s. 329–342. Titíž: *Noi contributii statistice la studiul paternitătii „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*eště 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.

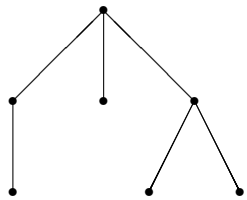
4) přesun formulace A na jiné místo, tj. Ax je přepsáno na xA.

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.

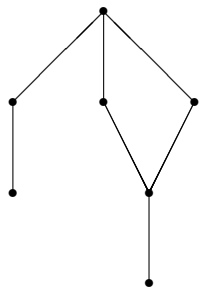
Sestavování stemmatu se skládá ze dvou fází:

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

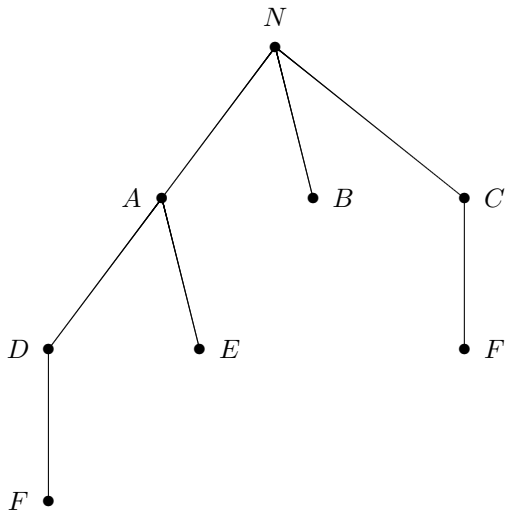
	35		36	37		38		39		
N	se	starou	Bětkou,	která	byla	její	vrstevnice	a	již u rodičů	sloužila.
A	se	starou	Bětkou,	která	byla	její	vrstevnice	a	již u rodičů	sloužila.
B	se	starou	Bětkou.	Ta	byla	její	vrstevnice	a	již u rodičů	sloužila.
C	se	starou	Bětkou,	která	byla	její	vrstevnice	a	již u rodičů	sloužila.
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):

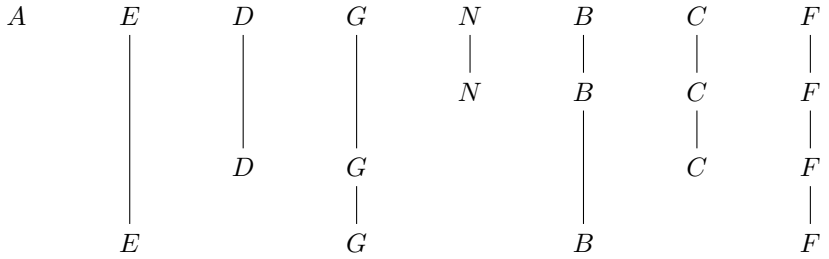
N	*A*	*B*	*C*	*D*	*E*	~~*F*~~	*G*
N		*B*	*C*			~~*F*~~	
			C			~~*F*~~	
						F	

Pro přehlednost můžeme upravit pořadí textů ve skupině:

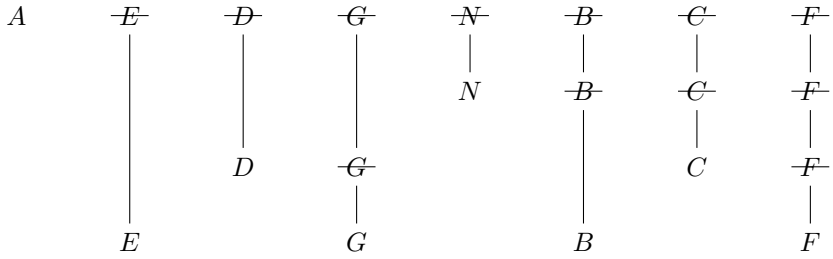
A	*D*	*E*	*G*	*N*	*B*	*C*	~~*F*~~
				N	*B*	*C*	~~*F*~~
						C	~~*F*~~
							F

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

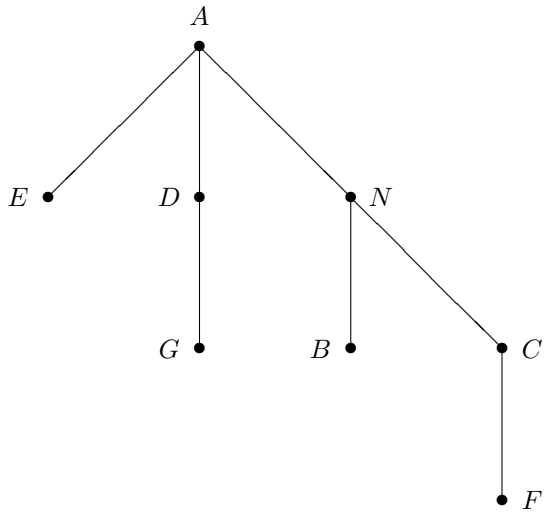
z pramene do pramene probíhal takto (spojnice mezi textovými prameny naznačují přenos změn):



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:

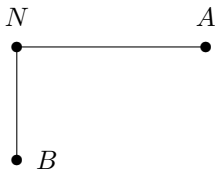
	N	A	B	C	D	E	F	G	součet
N	0	4	5	7	9	9	13	16	63
A	4	0	9	11	5	5	17	12	63
B	5	9	0	12	14	14	18	21	93
C	7	11	12	0	16	16	6	23	91
D	9	5	14	16	0	10	22	7	83
E	9	5	14	16	10	0	22	17	93
F	13	17	18	6	22	22	0	29	127
G	16	12	21	33	7	17	29	0	135

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

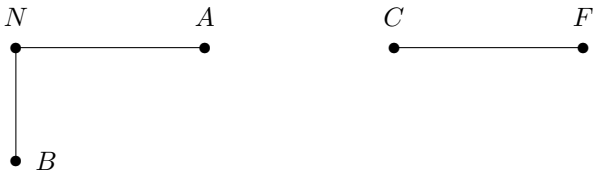
Nyní sestrojíme stemma. Pro každý textový pramen nalezneme v řádku (resp. sloupci) tabulky nejmenší vzdálenost, tj. pramen nejbližší, a tyto prameny v grafickém znázornění spojíme čarou. Tak postupujeme u všech pramenů, tedy napojujeme je na prameny již spojené.

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

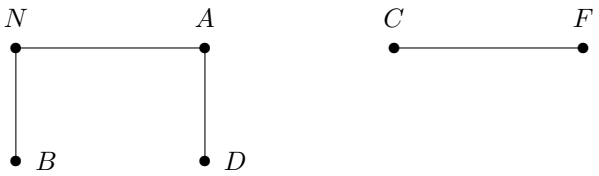
Ve třetím řádku zjišťujeme, že textovému pramenu B je nejbližší pramen N (vzdálenost 5), ke kterému byl již nejbližší pramen nalezen (pramen A), tj. graficky



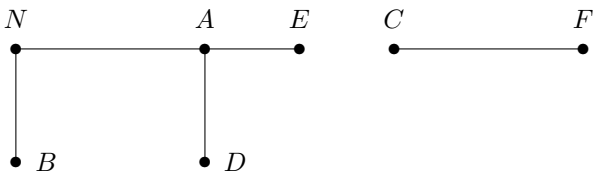
Pramenu C je nejbližší pramen F (vzdálenost 6), tj.



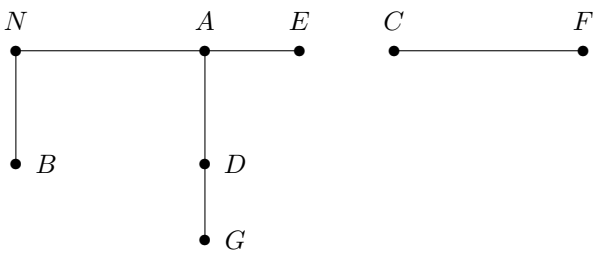
Pramenu D je nejbližší pramen A (vzdálenost 5), kterému nejbližší byl již nalezen, tj.



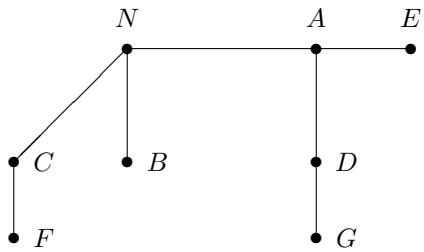
Pramenu E je nejbližší pramen A (byl již nalezen) se vzdáleností 5, tj.



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.