2.8 Information theory

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (English). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 92–101.

Persistent URL: http://dml.cz/dmlcz/402323

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

O vzdálenosti d mezi dvěma prameny můžeme říct, že:

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

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

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

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

2.8 Teorie informace

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

v druhém případě je entropie rovna

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

což po dosazení vzorce pro podmíněnou entropii upravíme na vzorec

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

z něhož dosazením vzorce

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

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

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

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

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

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

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

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

Tabulka 2.13: Hodnoty entropie různých řádů pro některé jazyky

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

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

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

(n je řád entropie).

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

$$H_\infty^{sl} \cdot \textit{počet slovních tvarů v textu} = H_\infty^{p\acute{\imath}s} \cdot \textit{počet písmen v textu}.$$

Po úpravě pak dostáváme

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Po krácení proměnnou N dostaneme výraz

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

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

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

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

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

PRIJEDU ZIBRA VEKER.

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

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

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

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

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

2.9 Glottochronologie

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

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

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