Univerzita Karlova v Praze

Filozofická fakulta

DISERTAČNÍ PRÁCE

Jiří Milička

Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů

The Theory of Communication as an Explanatory Principle for the Natural Multilevel Text Segmentation

Vedoucí disertační práce: doc. PhDr. Petr Zemánek, Csc.

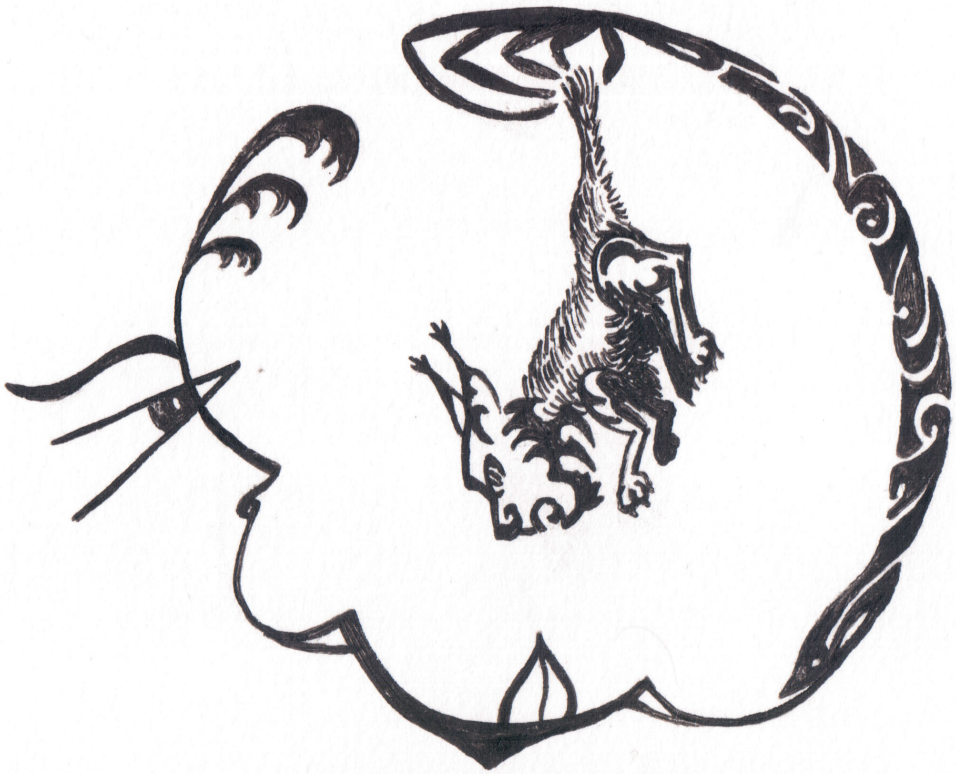
Základní součást: Ústav srovnávací jazykovědy

Studijní obor: Jazyky zemí Asie a Afriky

Praha 2015

Prohlašuji, že jsem disertační práci napsal samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, 7. 11. 2015



V prohlášení píši, že jsem disertaci napsal samostatně, což ovšem není celá pravda. Tato práce by nemohla vzniknout bez inspirace a rad mých kolegů, zejména doc. Petra Zemánka, jenž mě vedl doktorským studiem, a vůbec všech z ÚSJ FF UK, zejména Jana Bičovského a Pavla Čecha. A samozřejmě mnoha dalších kolegů, přátel i názorových oponentů, z nichž největší dík si zaslouží Hana Bednářová (která je také autorkou obrázku *Mluvčí polykající distinktivní rysy*), Zuzana Komrsková, Lenka Schindlerová, Karolína Vyskočilová, Eva Volenová, Radek Ocelák, Martin Beneš a Radek Čech. Také děkuji rodičům. A samozřejmě i daňovým poplatníkům České republiky a Evropské unie, kteří mě po dobu psaní nezištně podporovali a kterým tímto přeji hodnotný zážitek z četby.

Věnováno těm, kteří chtějí vědět, jak to, že ví, že neví, když neví.

Abstrakt

1. Na rozdíl od distinktivních rysů a morfémů nejsou hlásky, slova, věty ani souvětí logickou nezbytností jazyka.

2. Přesto je tento nebo podobný druh vnořené segmentace přítomen v různých jazycích a je pevně zakotven i v naší představě o jazyce.

3. Je tomu tak, neboť vnořená několikaúrovňová segmentace dovoluje vkládání redundance na různých úrovních, což je efektivní způsob, jak přenést informaci přes kanál obsahující dávkový šum.

4. Existuje mnoho strategií, jak redundanci vložit, a druhů redundance, které vloženy být mohou.

5. Kvůli oddělování segmentů je do segmentů vkládáno určité množství informace navíc. Množství této informace je nezávislé na délce segmentu, který odděluje. Na tomto principu je možno založit úspěšný model pro Menzerathův vztah.

Klíčová slova: Kvantitativní lingvistika, teorie informace, čeština, arabština.

Summary

1. Phonemes, words, clauses and sentences are not a logical necessity of language, unlike distinctive features and morphemes.

2. Despite this, such nested segmentation is very firmly present in languages and in our concepts of language description,

3. because nested segmentation and inserting redundancy on multiple levels is an efficient way to get the language signal through the burst-noise channel.

4. There are various strategies how redundancy can be added and what kind of redundancy can be added.

5. The segment delimiter is expressed by some additional information and the amount of delimiting information is independent from the length of the segment it delimits. This principle can serve as a basis for a successful model for the Menzerath's relation.

Keywords: Quantitative linguistics, Theory of information, Czech language, Arabic language.

Obsah

Úvod

Na počátku této disertace stálo jednoduché tvrzení: „výběr slova je omezen jeho kontextem“, a otázka, proč tomu tak je. Když jsem se do tématu ponořil hlouběji, zjistil jsem, že skutečnost, že jedna výpověď ovlivňuje vhodnost nebo pravděpodobnost užití výpovědi jiné, je tou snadněji vysvětlitelnou částí tvrzení — a s překvapením jsem si uvědomil, že lingvistika nemá odpověď na otázku, proč je jazyk vůbec dělen do slov. Slova jsou v teoriích často považována za předem dané kategorie, jakési významové celky, přičemž není jasné, v čem ona ucelenost spočívá. Dělení do slov nebo podobných segmentů považujeme za tak samozřejmé, že dospělého člověka snad ani nenapadne dožadovat se jeho vysvětlení.

Nechme tedy tuto otázku položit tříleté dítě. „Proč vůbec jsou slova?“ Otázku zobecním: „Zajímá mě, proč je mluvená i psaná řeč segmentovaná do slov, vět, souvětí, odstavců, kapitol, svazků a snad i vyšších více či méně ohraničených celků.“

„Protože jinak by to bylo nepřehledné,“ odpovíte asi.

Tedy otázku trochu přeformuluji: „Proč považujeme řeč, která není takto segmentovaná, za nepřehlednou?“ Byla snad nějaká biologická, fyzikální omezení, která zabránila tomu, aby se náš mozek vyvinul takovým způsobem, aby byl schopen produkovat i zpracovat nesegmentovaný tok symbolů? Nebo můžeme najít nějaké jiné vysvětlení, které je dáno přirozenými vlastnostmi jakékoliv komunikace?

Tato studie se pokusí ukázat, že dostatečné vysvětlení tohoto fenoménu poskytuje samotná teorie informace (nebo spíše teorie komunikace). To samozřejmě neznamená, že toto vysvětlení je jediné možné, dokonce ani to, že je nejdůležitější. To, co budeme popisovat na následujících stránkách, je nejspíš pouze jedním z mnoha faktorů, které do celého procesu zasahují.

Právě teorii informace (nebo spíše teorii komunikace) v lingvistickém kontextu představí první kapitola, která je poněkud obsáhlejší, neboť se pokouší uvést čtenáře do tématu tak, aby bylo srozumitelné i těm, kteří se s ním dosud nesetkali. Na této kapitole bude postaven celý následující text.

Druhá kapitola zpraví čtenáře o tom, co považuji za vědu a jaké místo v tomto epistemologickém systému zaujímá teorie informace. Ptáme se zde po významu a smyslu explanace, neboť vědecké vysvětlení je leitmotivem této studie.

Třetí kapitola se pokusí čtenáře přesvědčit, že dělení textu na hlásky, slova, věty,

souvětí a tak dále není nic samozřejmého, natož logicky nezbytného, a že tedy ústřední otázka této disertace je něčím, co je hodno zamyšlení.

Čtvrtá kapitola nás přivede k tomu, že přestože ona vnořená segmentace na slova, věty atd. není logickou nezbytností, rozhodně je jevem takřka univerzálním, vyskytujícím se v mnoha různých jazycích. Zároveň však kapitola ukáže, že univerzální je pouze idea vnořené segmentace, nikoli způsoby, jak je jí dosaženo, tedy že se jazyky velmi liší v tom, na jaké jednotky jsou segmentovány.

Konečně pátá kapitola, těžiště této práce, se bude zabývat modely přenosu informací mezi lidmi, které onu vnořenou segmentaci mohou vysvětlit. Začne od formalizovaného popisu prostředí, které musela a musí překonávat mluvená řeč. Přiblíží různé metody překonávání takovýchto prostředí, tak jak je známe z informatiky. Popíše, jakým způsobem tyto metody mohou být implementovány ve skutečném jazyce, a nakonec se pokusí o realistický formalizovaný model, z něhož vyplyne, že segmentace je pro efektivní přenos informací přes běžné typy kanálů výhodná.

Šestá kapitola představí konkrétní implementace, tedy jakým způsobem je redundance vkládána v reálném jazyce. Ukáže, že poesii můžeme chápat jako sekundární přenosový protokol nad jazykem, ovšem užívající obdobné metody jako jazyk. Tím se dotkneme estetiky redundantních vyjádření.

A v sedmé kapitole se podíváme na hranice segmentů. Konkrétně na to, jak je možné, že je příjemce pozná i bez vyznačení v textu. Postulujeme delimitační informaci, která je vkládána společně s redundancí k přenášené informaci. Na předpokladu, že její množství nezávisí na délce segmentu, který delimituje, postavíme model pro Menzerathův vztah. Tento model ověříme na českém a arabském textu a uvedeme do souvislostí.

* * *

V současnosti je vidět snaha o soustavný empirický přístup k mapování počátků jazyka a modelování jeho základních principů,[1] přičemž nové poznatky v antropologii, genetice, kognitivních vědách a informatice poskytují solidní znalostní bázi,

[1] V této souvislosti doporučuji edici *Studies in the Evolution of Language*, která je pod vedením Jamese Hurforda a Kathleen Gibson vydávána v Oxford University Press, zejména pak svazky (Hurford, 2007, 2012) a (Carstairs-McCarthy, 2010). Metodologicky pro mne byla velmi inspirativní studie *Self-Organization in the Evolution of Speech* (Oudeyer, 2006).

Podnětná je i kniha *Evolution of Semantic Systems* (Küppers et al., 2013) a s ní nijak nespojený, nicméně stejnojmenný velmi nadějný projekt na Institutu Maxe Plancka (http://www.mpi.nl/departments/other-research/research-consortia/eoss).

Další studnicí literatury k tomuto tématu pojednanému ze stejných pozic jako tato dizertace bude nejspíše v budoucnu edice *Computational Models of Language Evolution* v nakladatelství Language Science Press.

přičemž k přímému testování předpokladů a konsekvencí z nich plynoucích je používáno (počítačové, někdy přímo robotické) modelování pomocí interagujících agentů. Víceagentové modelování ovšem naše studie nevyužije a předpoklady budeme jednak odvozovat deduktivně, jednak testovat poněkud klasičtější metodou.

Cílem diachronních exkurzů, které tato práce skýtá, je ukázat, že segmentace textů byla obvyklá, *kam až paměť sahá*. Nebudu se rozepisovat o svých představách, jak konkrétně jazyk do tohoto stavu dospěl — vzhledem k současným znalostem by byla odpověď nutně velmi spekulativní. Můj osobní názor je, že jazyk k několikaúrovňové segmentaci nedospěl dodatečně, ale že *vznikla společně s jazykem*, postupně. Tedy že když se prodlužovaly a zesložiťovaly typické promluvy, spontánně se začaly objevovat tendence k jejich segmentaci na kratší celky. V každém případě ovšem neřeším, *jak konkrétně k tomu docházelo*, pouze *proč k tomu docházelo*, přičemž ono odůvodnění předpokládá, že jazyková změna (nespecifikovaným způsobem) z jazyka odstraňuje konstrukce, které vedly k neúspěšné komunikaci.[2] Vzhledem k naší nedostatečné znalosti konkrétních procesů jazykové evoluce můžete dojít k závěru, že moje explanace má funkcionální povahu, podobně jako explanace postavené na klasické Darwinově teorii před příchodem genetiky a neodarwinismu.[3]

Konečně vývoj jazyka je organickou součástí evoluce člověka — ti, jejichž mozky a hlasivky byly přizpůsobeny efektivní komunikaci, měli lepší šance na přežití a odchov potomků. A nebo jej můžeme chápat jako pokračování biologické evoluce jinými prostředky — *variace* (nebo chcete-li *chyba*) je pro jazykovou změnu podobná genové mutaci, úspěšný komunikační akt je ekvivalentem přežití. To ovšem chápejme jen jako metaforu, šíření jazyka se šíření genů do značné míry už z principu nepodobá a je třeba uvědomit si omezení z tohoto přístupu plynoucí.

[2] Dlouhodobě stačí i princip se slabým efektem, který nějakým způsobem znevýhodňuje některé způsoby vyjadřování, aby se tendence projevila. Enfield těmto efektům říká „transmission bias“ a chápe je jako hlavní explanatorní mechanismus jazykové změny (Enfield, 2014, str. 18). Nutno dodat, že spolu může soupeřit více principů a že s jazykovou změnou se mohou vynořovat další. Na tom je ostatně postaven Köhlerův synergetický kontrolní cyklus (Köhler, 2005).

[3] Status funkcionálních vysvětlení (*functional explanation*) je některými epistemology zpochybňován (např. Hempel, 1959). Hájí je naopak Wouters, k tématu doporučuji zejména jeho disertaci (1999); vzhledem k pojmové nejasnosti a ideologické zatíženosti pak později používá spíše termín *design explanation* (Wouters, 2007).

Pro snazší představu o předmětu sporu si představte, že na otázku „proč mají hadi rozeklaný jazyk a nikoli zašpičatělý?“ biolog odpoví „protože díky tomu mohou mít prostorově orientovaný čich, což jim pomáhá při vyhledávání kořisti a adeptů k páření“ (tato hypotéza je nyní skutečně uznávána, viz Schwenk, 1994). Takové vysvětlení je sice funkcionální, ale není plně teleologické, neboť kdesi v pozadí takové studie je předpoklad, že o hady s nevhodným tvarem jazyka se postarala evoluce. Tato explanace by byla čistě funkcionální pouze v době, kdy biologie nevěděla nic o genetice a mechanismech, jakými se přirozeného výběru dosahuje. V této době však vysvětlení tohoto druhu bylo zároveň jediné možné.

Mám pocit, že pokud se laik zeptá, proč má had rozeklaný jazyk, pak očekává právě takovouto funkcionální explanaci, a naopak kauzální explanace typu „protože před x lety došlo k mutaci chromozomu y“ by považoval za neuspokojivé.

Tato práce ovšem může přispět i k formování obecné představy o tom, jak konkrétně funguje rozpoznání neúspěšných komunikačních aktů. Tím potažmo osvětluje konkrétní prostředky, jimiž k jazykové evoluci dochází, čímž přispívá k formování kauzální explanace nejen jevu, který je hlavním námětem této práce, ale i mnoha dalších.

Kapitola 1

Teorie komunikace a jazyk

Cílem této terminologické kapitoly není podat systematický přehled o tom, jak jsou termíny užívány v relevantní literatuře, pouze se snaží navést čtenáře k tomu, jakým způsobem budou pojmy použity v této práci.

1.1 Informace a komunikace

Poněvadž některé termíny používám jinak, než jak je v lingvistice zvykem, je tato kapitola velmi důležitá. Hlavní principy, ze kterých budu odvozovat své hypotézy, jsou pevně zakotveny v algoritmické teorii informace, ovšem už u pojmu *informace* začíná první terminologická nesnáz. Samotní informatici tento pojem definují implicitně, tedy tak, že uvádějí celou teorii informace a studují systémy, ve kterých hraje roli (Sloman, 2011, str. 24). Za informaci můžeme označit prakticky jakoukoli vlastnost hmoty, která je od této hmoty abstrahovatelná. V současnosti hlavní proud informatiky (nejspíše ovlivněn nízkoúrovňovými vlastnostmi lidského mozku a počítačů) chápe informaci jako diskrétní veličinu, a tak, ať už je definice jakákoliv, nejmenší jednotkou informace je jeden bit (který v klasických strukturalistických představách odpovídá jedné binární opozici).

Gregory Bateson (1972, str. 46) definuje jednotku informace jako *difference that makes difference*, tedy (nejmenší) *rozdíl, na kterém záleží*. Tato definice se mi líbí proto, že pojímá informaci jako abstraktní entitu a zároveň do definice zahrnuje relativitu vůči zbytku systému, který zkoumáme. Technicky řečeno, jestli je abstraktní entita informací, závisí na entitě, vůči které ji vztahujeme. To, že na rozdílu záleží, se projevuje tak, že se tento rozdíl šíří (propaguje) dále skrze systém a do dalších systémů. Méně technicky a na příkladu: o tom, co je pro vás informace, si rozhodujete sami.

Onu *propagaci informace* v čase nebo prostoru můžeme nazvat komunikací, což je další obtížně definovatelný pojem. Jeden ze zakladatelů kybernetiky, W. Ross Ashby (1962), chápal produktora a recipienta jako dvě entity, které jsou propojeny vazbou,

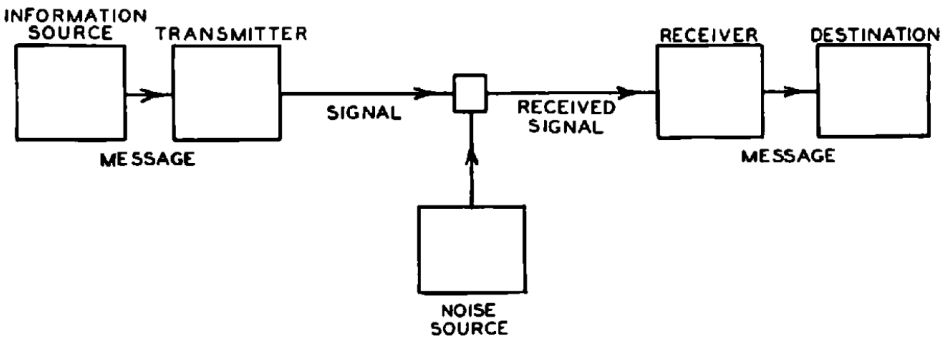
tedy že stav produktora nějakým způsobem omezuje stav recipienta, ne nutně lineárně nebo deterministicky. Informace, kterou recipient od produktora přijme, může změnit recipienta natolik, že se do něj začnou propagovat i další stavy produktora, tedy že začne záležet i na rozdílech, na kterých předtím nezáleželo. Například informace dodaná pistoli přes její spoušť se nebude propagovat dál, pokud je zbraň zajištěná, což je možno změnit dodáním jiné informace přes pojistku; podobně tato věta by pro mnoho čtenářů byla nesrozumitelná, pokud by na ni nebyli připraveni předchozím odstavcem. Informace se samozřejmě mohou týkat i kódu, kterým budou další informace předány.

Hlavní terminologická nesnáz spočívá v tom, že naše chápání pojmu *komunikace* je běžně antropocentrické, a stejně k němu přistupují i někteří vědci. Ne každé vzájemné ovlivnění dvou entit je v běžné řeči pokládáno za komunikaci. Pokud do sebe narazí dva kameny, těžko budeme onen akt nazývat komunikací, a pokud mluvíme o dvou komunikujících přístrojích, pak je to spíše metafora komunikace lidské, než že bychom vyjadřovali přesvědčení, že počítač posílající jinému počítači animovaný obrázek dvou hrajících si koťátek je principiálně odlišný od kamene, který narazí do jiného. Pokud muž A ukradne muži B láhev vodky a ten mu dá facku, pak za výsledek tohoto komunikačního aktu budeme pokládat to, že si muž A dá příště větší pozor, aby ho nikdo nechytil, nikoli to, že bezprostředně schytá druhou facku o stěnu. Zdá se tedy, že informace musí být propagována skrze centrální nervovou soustavu nebo něco, co ji připomíná, aby tato propagace byla v takovémto obecném antropocentrickém pojetí nazvána komunikací.

Mnohem pragmatičtěji ke komunikaci přistupoval Shannon (1948), který předpokládal, že komunikace je všeobecně srozumitelný pojem, a postuloval *zašuměný kanál (noisy channel)*, tedy médium, přes které signál musí projít, aby dorazil od zdroje informací k cíli. Mezi systémem, který informaci produkuje, a systémem, který ji přijímá, je přenosové médium, které může do informace vnést chyby. To znamená u některých bitů přenastavit hodnotu. Klasické schéma naleznete na obrázku 1.1 a příklad v tabulce 1.1.

Shannonovské inženýrské pojetí problému bylo přejato i do teoretických systémů a redefinice základních pojmů (před 40. lety 20. století byl pojem *informace* zakořeněn ve filozofii i lingvistice s významem navázaným na termín *význam*) vedla k velkému a vleklému terminologickému zmatení a zatemnění pojmů *informace*, *komplexita*, *entropie*, *nejistota* (*uncertainity*) a *význam*. K tomuto tématu doporučuji článek Shannonova současníka Bar-Hillela (1955).

Ještě jednou se vraťme k Batesonovi a připomeňme důležitost vztažení informace k recipientovi, neboť až ten rozhoduje o tom, co je signál a co je šum. Typickým příkladem je pohled z okna. Pokud chceme vědět, jaké je venku počasí, je zdrojem signálu to, co je za oknem, a samotné okno je zašuměným kanálem. Naproti tomu pokud chceme vědět, jestli je okno špinavé, tak zdrojem signálu je špína na okně, a naopak to, co vidíme za ním, je zdroj šumu, a tedy vlastně součást zašuměného kanálu.



Obrázek 1.1: Klasické schéma cesty informace od produktora k recipientovi přes zašuměný kanál (Shannon, 1948, str. 381).

Produktor		Recipient
1		1
0		0
1	$\longrightarrow$	1
1		1
0		0

Tabulka 1.1: Zde vidíme situaci, kdy informace došla přes zašuměný kanál nepoškozena.

Dále je třeba si uvědomit, že kanál nepřenáší informace jen prostorově, ale i časově (kniha a psaný text, ale koneckonců i mluvený text šířený vzduchem překonávají časovou vzdálenost). Lidé se během vývoje jazyka vyrovnávali s velmi rozmanitými kanály různých typů a vlastností.

1.2 Komplexita a redundance

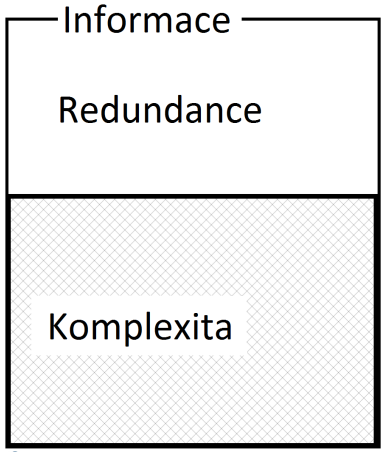
Dalšími pojmy zásadními pro další práci jsou *komplexita* a *redundance*. Oba tyto pojmy budu používat v kolmogorovovském smyslu[1].

Oba tyto termíny značí druh informace, měří se tedy také v bitech.

[1] Klasická reference ke kolmogorovovské komplexitě je (Kolmogorov, 1965), nicméně pro uvedení do problematiky doporučuji spíše kapitolu od Li – Vitányi (1990), v případě opravdu velkého zájmu pak úctyhodné kompendium Li – Vitányi (2013), kterážto kniha probírá i historické pozadí spojené s myšlenkami von Misese a obsahuje bohaté reference na literaturu k tématu.

Komplexitou rozumíme nejkratší možnou sekvenci, která popisuje danou informaci (tato sekvence je algoritmem, který přijímající systém dokáže interpretovat, proto mluvíme o algoritmické komplexitě). Pro účely této práce nepředpokládáme jakýsi univerzální minimální kód, jímž by tento algoritmus byl zapsán, naopak počítáme s tím, že onen kód musí být srozumitelný pro systém, vůči kterému se informace vztahuje (například recipient zprávy).

Redundancí pak máme na mysli tu část informace, která není komplexitou (obrázek 1.2).



Obrázek 1.2: Názorné schéma komplexity a redundance.

.

Nejlépe si tyto dva termíny osvětlíme na příkladu. Představme si, že máme informaci, která vypadá následovně:

101010101010101010101010101010101010101010101010101010101
010101010101010101010101010101010101010101010101010101010
101010101010101010101010101010101010101010101010101010101
010101010101010101010101010101010101010101010101010101010
101010101010101010101010101010101010101010101010101010101
010101010101010101010101010101010101010101010101010101010
101010101010101010101010101010101010101010101010101010101
010101010101010101010101010101010101010101010101010101010
101010101010101010101010101010101010101010101010101010101
01010

Tato informace zabírá v binárním kódování 500 bitů.
Sdělení této informace můžeme vyjádřit takto:[2]

[2]Tato notace bude fungovat, jen když příjemce bude schopen interpretovat perlovské regularní

(1010){125}

To platí za předpokladu, že recipient je schopen dešifrovat oba formáty, což není samozřejmé. V tomto případě zpráva (při zbytečně velkorysých osmi bitech na písmeno) zabírá 88 bitů. Vidíme, že nejméně 412 bitů bylo redundantních.

Zprávu můžeme zapsat ještě úsporněji:

(10){250}

V tomto případě se ukazuje, že komplexita zprávy je pouze 72 bitů.

A protože nikdy nevíme, jestli zprávu nejde zapsat ještě úsporněji, jisté je pouze to, že komplexita zprávy je *maximálně* 72 bitů. A tedy že redundance původní zprávy je *minimálně* 428 bitů. Komplexitu tedy nikdy nemůžeme změřit přesně. Ale můžeme ji aproximovat shora. Stejně tak redundanci můžeme aproximovat zdola.

K odstraňování redundance jsou určeny tzv. kompresní algoritmy. To, co jsme si právě ukázali, byla *bezztrátová komprese*. Původní signál mohl být bezezbytku rekonstruován. Někdy je ovšem účelné zbavit se redundance i za cenu ztráty určité části komplexity, čemuž se říká *ztrátová komprese*. Představte si, že nám zdroj poskytl následující informaci:

11101010101010101010101010101010101010101010101010101
01010101010101010101010101010101010101010101010101010
10101010101010101010101010101010101010111010101010101
01010101010101010101010101010101010101010101010101010
10101010101010101010101010101010101010101010101010101
01010101010101010101010111010101010101010101010101010
10101010101010101010101010101010101010101010101010101
01010101010101010101010101010101010101010101010101010
1010101010

Což můžeme v duchu předchozího příkladu komprimovat bezztrátově na:[3]

(11(10){73}){3}

Kterážto informace zabírá 120 bitů. Nebo můžeme obětovat část informace (pomineme některé detaily) a ztrátovou komprimací se dostat na velikost komplexity stejnou jako v předchozího příkladě:

(10){225}

výrazy. Pro běžného člověka můžeme informaci zapsat nějak jako *125 krát "1010"*. I tento zápis je mnohem kratší než původních 500 bitů.

[3] Opět pro běžného člověka můžeme zapsat třeba jako *3 krát (po "11" následuje 74 krát "10")*.

Shrňme si, že oba pojmy znamenají něco poněkud jiného než v lingvistice, kde se termínem *komplexita* označuje velké množství různých a vzájemně jen málo souvisejících fenoménů[4] a pojem *redundance* má poměrně úzký význam — obecně se jím označují části textu, jež jsou z pohledu toho kterého lingvisty nepotřebné. My se naopak během celé práce budeme dívat na redundanci jako na entitu nutnou, která jednak přichází do jazyka z okolního světa při jeho vnímání, jednak je nezbytná pro přenos informace přes zašuměný kanál, jak bude ukázáno v následujících třech podkapitolách. Všechna další použití těchto termínů v této práci budou vycházet z jejich běžného užití v algoritmické teorii informace, nikoli v lingvistice.

1.3 Kódy pro detekci chyb

Redundance se přirozeně objevuje ve světě tak, jak ho vnímáme svými smysly, které nám poskytují přebytek dat, a náš mozek je velmi dobře uzpůsoben k jejich ztrátové kompresi. Kromě nepotřebné redundance, které se chceme zbavit (a to i za cenu ztráty části komplexity), však někdy redundanci schválně přidáváme, a to tehdy, když chceme informaci přenést přes zašuměný kanál.

Vraťme se k prvnímu obrázku a ukažme si, co se stane, když zašuměný kanál převrátí hodnotu jednoho bitu ve zprávě (tabulka 1.2).

Produktor		Recipient
1		1
0		0
1	⟶	1
1		0
0		0

Tabulka 1.2: Zde vidíme situaci, kdy informace došla přes zašuměný kanál poškozena — čtvrtý bit během přenosu změnil hodnotu z jedničky na nulu.

V tomto případě, kdy ve zprávě není žádná redundance, příjemce nemůže vědět,

[4] Tato variabilita se odvíjí od toho, že v přirozeném jazyce je slovo *komplexita* (složitost) velmi vágní a může znamenat leccos, takže když se oprostíme od teorie informace, tak komplexitou můžeme nazvat prakticky libovolnou metriku. Navíc i když se teorie informace držíme, můžeme měřit jednak komplexitu odesílané zprávy, jednak komplexitu pravidel, kterými zprávu kódujeme, respektive komplexitu lingvistického popisu těchto pravidel. Jako příklad nám může sloužit *komplexita morfologického subsystému*, která může být chápána jako metrika pro „komplexitu“ té části zprávy, která je kódovaná morfologií (ať už to znamená cokoli) — takto ji pojímal třeba Juola (1998) a Bane (2008). Naproti tomu Vulanović (2007) ji chápal jako komplexitu morfologických pravidel jazyka. K diskusi o problematičnosti obou přístupů doporučuji (Martín, 2011).

že zpráva nedorazila v pořádku. Ukažme si jeden jednoduchý způsob, jak do zprávy dodat redundanci tak, aby příjemce poznal, že je jeden bit špatně:

Ke zprávě před odesláním přidáme jeden redundantní bit, tzv. paritní bit (*parity bit*). Pokud je počet jedničkových bitů ve zprávě lichý, přidáme ke zprávě paritní bit s hodnotou jedna, pokud sudý, pak nechť je jeho hodnota rovna nule. Když zpráva dorazí, příjemce paritu přepočítá, a pokud se liší od paritního bitu, který byl přenesen se zprávou, pak ví, že zpráva je nekonzistentní, a může podle toho jednat, například produktora požádá, abych tuto zprávu poslal ještě jednou. V tabulce 1.3, části A) zpráva dorazila správně a paritní bit odpovídá. V části B) téže tabulky čtvrtý bit dorazil s převrácenou hodnotou, což se projevilo tím, že po přepočítání parity se zpráva ukázala být nekonzistentní.

	A)			B)		
	Produktor		Recipient	Produktor		Recipient
Zpráva:	1		1	1		1
	0		0	0		0
	1	→	1	1	→	1
	1		1	1		0
	0		0	0		0
Redundance:	$\boxed{1}$		$\boxed{1}$	$\boxed{1}$		$\boxed{1} \neq 0$

Tabulka 1.3: Zprávu jsme odeslali včetně redundantního paritního bitu (označen rámečkem). V části A) zpráva dorazila správně a paritní bit odpovídá. V části B) čtvrtý bit nedorazil do svého cíle bez úhony, což se projevilo tím, že po přepočítání parity se ukázala nekonzistence zprávy.

Pokud by zašuměný kanál převrátil hodnotu dvou bitů ve zprávě současně, pak by paritní bit měl správnou hodnotu a zpráva by se jevila jako konzistentní, přestože byla změněna. Při přenosu samozřejmě může dojít i ke špatnému přenosu paritního bitu. Může pomoci, pokud zvětšíme množství redundance, ovšem pokud je pravděpodobnost změny jednotlivých bitů při přenosu kanálem nenulová, pak si (při jakkoli velkém množství redundance) nikdy nemůžeme být jisti, že zpráva dorazila v pořádku. Tuto nejistotu však přidáním vhodného množství redundance můžeme snížit na únosné minimum.

Místo paritního bitu může více či méně účelně zaujmout jakákoli informace, kterou je možno vydedukovat z vlastní zprávy. Účelnost je dána tím, jak moc se dedukovaná informace liší v závislosti na množství chyb ve zprávě. Mezilidská komunikace je samozřejmě mnohem komplikovanější, neboť různé části informace mohou mít pro odesilatele nebo pro příjemce různou hodnotu. Například si představme předpověď

počasí „zítra bude 25 °C a budou padat dvoucentimetrové kroupy“. Jsme-li majiteli auta a zároveň nevlastníme garáž, informace o kroupách bude pro nás mnohem významnější než informace o teplotě. Proto je účinná komunikační strategie, když hlasatelka přidá redundantní informaci „doporučujeme tedy zahrádkářům, aby si dali pozor na své skleníky“, neboť pokud jsme přeslechli nebo nevnímali větu o kroupách, tato redundantní informace nás přiměje, abychom se podívali na předpověď počasí ještě jednou, popřípadě si ji ověřili z jiného zdroje.

Kromě modelu, kdy je redundantní informace přidružena k odesílané zprávě, se v přirozeném jazyce zhusta používá poněkud komplikovanější opačný způsob: příjemce je odesilatelem zprávy požádán, aby na základě zprávy určitým způsobem vydedukoval určitou informaci a tu mu poslal zpátky. Pokud ona vydedukovaná informace souhlasí s tím, co stejným postupem vydedukoval odesilatel, pak je to pro něj známka toho, že k přenosu došlo správně. Typické užití tohoto modelu je ve škole, kdy učitel vhodně položenými otázkami zjišťuje, jestli žáci výkladu rozumí.

1.4 Samoopravné kódy

Vhodně vložená a dostatečně velká redundance může dopomoci nejen k detekci chyb, ale také k jejich opravě, aniž by bylo nutné ptát se původce zprávy. Tabulka 1.4 ukazuje asi nejjednodušší způsob konstrukce samoopravného kódu: modulární redundance. Každý bit příjemci pošleme čtyřikrát po sobě. Pokud mu dojdou všechny čtyři bity stejně, necháme ho předpokládat, že došly bez chyby. Pokud tři došly stejně a jeden jinak, pak je pravděpodobné, že onen jiný bit je špatně a tři stejné správně, a necháme jej, aby zprávu opravil. Pokud budou vždy dva a dva stejně, bude chyba detekovatelná, nicméně nebude možné ji opravit.

PZ	PR		DZ	Interpretace:
1	111		1101	Opraveno na 1
0	000		0000	Nepoškozená 0
1	111	⟶	1111	Nepoškozená 1
1	111		1100	Neopravitelné
0	000		1110	Chybně opraveno na 1

Tabulka 1.4: Modulární redundance (*modular redundancy*) o třech redundantních bitech. PZ — původní zpráva; PR — přidaná redundance; DZ — doručená zpráva. Čtvrtý bit původní zprávy (odshora) není opravitelný, nicméně chyba byla detekována, pátý bit zprávy dorazil natolik poškozen, že chyba nebyla ani opravena, a dokonce ani nemohla být detekována.

Při správném užití samoopravných kódů může při chybě dojít ke čtyřem situacím:

1. Zpráva dojde nepoškozená nebo redundance je dost velká na to, aby bylo možné zprávu opravit (příjemce ví a ví, že ví);

2. přenášená informace dojde nepoškozená, ale chyba v redundanci způsobí, že příjemce ji detekuje jako chybnou (příjemce ví a neví, že neví); tato situace může nastat pouze v metodách kódování, které vlastní zprávu a redundanci nějak oddělují;

3. zpráva dojde poškozená a redundance je dost velká na to, aby chybu bylo možné detekovat, ale ne dost velká na to, aby ji bylo možno opravit (příjemce neví a ví, že neví);

4. poškození je tak velké, že chybu není možno ani detekovat (příjemce neví a neví, že neví).

Pro detekci chyb je třeba méně redundance než pro jejich opravení, což favorizuje dialogické komunikační systémy, kdy je možno požádat odesilatele zpráv o opravení detekované chyby. Zřejmě i toto zjištění stálo za nedůvěrou různých kultur k uchovávání informací pomocí psaného textu, kteréžto už z podstaty opravu nedovoluje. Navzdory velké literární tradici na Blízkém východě islámští intelektuálové dlouhou dobu zdůrazňovali úlohu učitele při přenosu informace a knihu chápali jako nedokonalý odraz autora, kterého musí zastupovat učitel. Onen učitel musel mít samozřejmě také svého učitele, a to tak, že řetěz učitelů začínal u samotného autora, čímž mělo být zajištěno, že se neztratí správná interpretace knihy (Robinson, 1993, str. 235–240). Kořeny jsou samozřejmě mnohem starší, pro evropský a středomořský areál se tato představa táhne minimálně od Platóna (Faidros, 275e).

Chceme-li, aby zpráva dorazila bez chyb, velikost redundance posílané společně s původní zprávou nesmí být menší než komplexita šumu.

Idealizovaný příklad, na němž si toto tvrzení ilustrujeme, zahrnuje zašuměný kanál, který z každých sedmi bitů, které přes něj pošleme, může vybrat náhodně jeden a jeho hodnotu přetočit z nuly na jedničku nebo obráceně.

Šum, ve kterém přetočení bitu ve zprávě značíme jedničkou a ponechání bitu ve zprávě tak, jak je, značíme nulou, vypadá například takto:

Šum: 0000100 0100000 0000100 0000010

Zpráva může vypadat například takto:

Zpráva: 0101011 1110000 0100010 1101011

Po průchodu výše zmíněným zašuměným kanálem má zpráva následující hodnotu:

Doručená zpráva: 0101111 1010000 0100110 1101001

Pozice každého přetočeného bitu může nabývat hodnot mezi 0–7 (nula pro případ, že žádný bit není přetočený), což se dá zapsat pomocí tří bitů (neboť $2^3 = 8$). Komplexita tohoto šumu tak dosahuje maximálně tří bitů na každých sedm bitů zprávy, dohromady tedy ve výše uvedeném příkladě 12 bitů. Tedy šum sice zabírá 28 bitů, ovšem z toho je nejméně 16 bitů redundantních; na každé 3 bity komplexity připadají 4 bity redundance.

Abychom měli jistotu, že zpráva dojde nepoškozená, přidáme ke každým čtyřem bitům zprávy tři bity redundance, dohromady v tomto případě tedy 21 redundantních bitů, čímž se zpráva prodlouží na 49 bitů. Šum se tím taky prodlouží na 49 bitů, z nich ovšem pouze 21 bude tvořit jeho komplexitu ($\frac{49}{7} \times 3 = 21$). Délka komplexity šumu je rovna délce redundance zprávy a je možné vymyslet kódování, které nám zaručí, že zpráva dojde nepoškozená. Použijeme Hammingovo kódování,[5] redundantní bit je značen rámečkem (tabulka 1.5).

0	1	0	0	1	0	1
0	0	0	1	1	1	1
0	1	1	1	1	0	0
1	1	0	1	0	0	1
1	1	0	1	0	0	1
1	1	0	0	1	1	0
0	1	1	0	0	1	1

Tabulka 1.5: Zpráva zakódovaná Hammingovým kódováním.

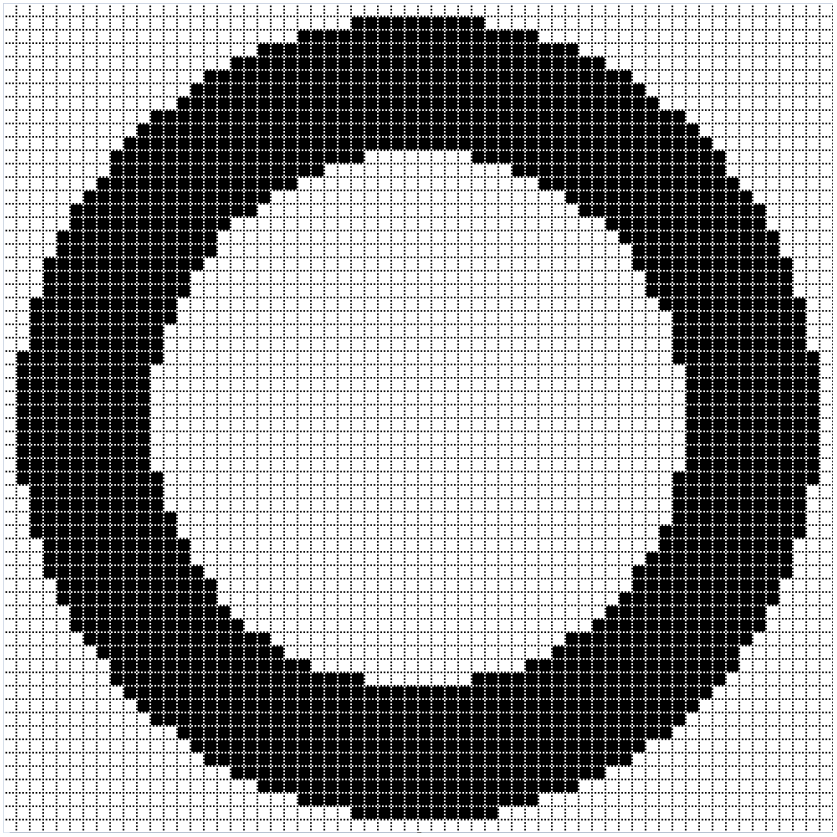
Vše by bylo složitější, kdybychom užívali Shannonova termínu *entropie*, který značí něco jako komplexitu, která systému chybí. Shannonovská entropie je definována tak, aby měla konotace k entropii, jak je chápána ve fyzice, přičemž i ona je spíše temným pojmem,[6] pročež se tomuto pojmu budeme spíše vyhýbat. Přesto doporučuji vážným zájemcům o teorii informace se s entropií seznámit podrobněji, například na stránkách knihy MacKay (2005, kapitola 2).

[5] Podrobné a snadno pochopitelné vysvětlení Hammingova kódování naleznete v MacKay (2005, str. 8). Stručně řečeno, pokud si bity v sedmici očíslujeme 1 až 7, tak 1. bit je nastaven jako paritní vůči bitům 3, 5 a 7, druhý bit je paritní k bitům 3, 6 a 7 a čtvrtý bit je paritní k bitům 5, 6 a 7. Když je tedy například 1. bit ve zprávě přenesen špatně, tehdy a jen tehdy je nekonzistentní pouze parita prvního redundantního bitu; když je přenesen špatně bit poslední, tehdy a jen tehdy jsou nekonzistentní všechny paritní bity; pokud je nekonzistentní jen první a druhý paritní bit, příjemce zprávy si může vydedukovat, že byl špatně přenesen 3. bit.

[6] Podle anekdotické zmínky je tato zastřenost dokonce záměrná: „Von Neumann told Shannon to call his measure entropy since 'no one knows what entropy is, so in a debate you will always have the advantage'." (Campbell, 1982, str. 32)

1.5 Příklad odstranění nechtěné redundance a přidání redundance potřebné pro přenos

Protože příkladů ze života není nikdy dost, ukažme si, jak funguje komprese redundance nechtěné a přidání redundance chtěné.



Obrázek 1.3: Kruh v bitmapě.

Obrázek 1.3 má na šířku i na výšku 62 černobílých pixelů, zaujímá tedy 3844 bitů (cca 480 bytů).

Pokud ho uložíme ve formátu SVG (Scalable Vector Graphics), který je do značné míry čitelný i pro člověka, zbavíme se velké části redundance, takže zabírá pouze 147 bytů:[7]

```
<svg xmlns="http://www.w3.org/2000/svg" width="62" height="62">
```

[7] První řádek značí, že se jedná o formát SVG, odkazuje na způsob, jakým se má tento kód interpretovat, a určuje, že kreslicí plátno má tvar čtverce o hraně 62 nespecifikovaných jednotek (např. pixelů). Druhý řádek značí, že interpret má na toto plátno nakreslit černou čarou o tloušťce 10 jednotek nevyplněný kruh se středem na souřadnicích 30 a 30 jednotek a s poloměrem 25 jednotek.

```
<circle cx="30" cy="30" r="25" fill="none" stroke="black"
 stroke-width="10" />
</svg>
```

Nyní do obrázku přidáme redundanci, která pomůže příjemci detekovat některé chyby, k nimž může dojít při přenosu (velikost obrázku naroste na 196 B):

```
<svg xmlns="http://www.w3.org/2000/svg" width="62" height="62">
<circle cx="30" cy="30" r="25" fill="none" stroke="black"
 stroke-width="10" />
<!-- Obsahuje 8 čísel a jejich součet je 2222 -->
</svg>
```

Nyní dodejme do poznámky k popisu obrázku nějaký neformální samoopravný kód:

```
<svg xmlns="http://www.w3.org/2000/svg" width="62" height="62">
<circle cx="30" cy="30" r="25" fill="none" stroke="black"
 stroke-width="10" />
<!-- opakuji: width=62; width=height=62; cx=cy=(width-2)/2=30;
 stroke-width=10; r=cy-stroke-width/2=25;
  r=(width-2)/2-stroke-width/2=25 -->
</svg>
```

Velikost zprávy sice narostla na 286 bytů, příjemce však může určitý počet chyb opravit a ještě větší počet detekovat.

1.6 Jazyk a text

Jazyk je metoda komunikace. Abychom byli přesnější:

1. Přirozený lidský jazyk je jedním ze způsobů, jakým člověk kóduje informaci tak, aby mohla být přijata jiným člověkem. Tento způsob nebudeme specifikovat (nejedná se o definici, ale o hyperonymum).

2. Textem rozumíme konkrétní implementaci této metody, tedy informaci zakódovanou jazykem.

Aby to nebylo tak jednoduché, na tuto metodu komunikace se lingvisté dívají nejméně dvěma způsoby. Jednak můžeme zkoumat individuální metodu komunikace každého jednotlivého člověka, chápat ji jako osobní rys, samozřejmě se měnící, zejména v závislosti na komunikačních metodách, které používají jiní lidé. Jakékoli univerzálie pak budou zobecněním těchto jednotlivých individuálních rysů, individuálních „kompetencí“. Tento přístup je příjemně přímočarý a umožňuje používat

analogie s vědami, které zkoumají méně abstraktní entity. Jazyk takto definovaný můžeme zkoumat podobně jako třeba chřipkovou epidemii, vždyť chřipka má také několik úrovní dynamického vývoje, z nichž některé jsou netriviálně ovlivněny vnějšími okolnostmi a chřipkou jiných lidí, má příčiny (ovšem ne funkce), má biologicky daná omezení, a přesto je individuální. Hypotéza, že většina lidí používá jako nepříznakový slovosled SVO, VSO nebo SOV, je strukturně podobná hypotéze, že většině lidí trvá za normálních podmínek onemocnění chřipkou sedm dní, a k jejímu testování je možno použít obdobných metod a statistického aparátu.

Pojmy jako *slovosled* nebo *příznakovost* však vznikly v dobách, kdy se za samozřejmý bral jiný přístup, který staví na tom, že jazyky jednotlivých lidí se sobě musí už z definice podobat, jinak by nemohly hrát roli metody komunikace. Pod pojmem „jazyk“ pak mají lingvisté na mysli jakýsi abstraktní konstrukt postavený nad jazyky (*idiolekty*) jednotlivých lidí, ovšem od těchto lidí do větší či menší míry abstrahovaný. Je to jazyk ve významu *langue* (který je definován jako ultimátně abstrahovaný dorozumívací kód), ovšem tento poněkud komplikovaný význam slova *jazyk* byl používán dávno před Saussurem a je běžně užíván i mimo lingvistiku.

V drtivé většině běžných užití mezi těmito dvěma pojetími není třeba rozlišovat. V normální komunikaci není rozdíl, jestli větou „chtěla bych se naučit arabsky“ myslíte něco jako „chtěla bych se nakazit stejným kmenem chřipky jako spolužáci ve třídě“, nebo jestli tím chcete vyjádřit obdobu věty „chtěla bych se naučit pravidla šachů tak, jak se hrály v Indii v 10. století“. V případě zkoumání jazyka to však velký rozdíl je, neboť různá míra abstrakce od mluvčích dovoluje různé metafory a implikuje různé výzkumné otázky a způsob jejich řešení. Najednou se zdá být validní hypotéza „většina *jazyků* používá jako nepříznakový slovosled SVO, VSO nebo SOV“, přičemž jazyk o dvou mluvčích, kteří se před lety rozhádali, a tak spolu nemluví,[8] má stejnou váhu jako jazyk, kterým spolu úspěšně komunikují stovky milionů lidí. Výzkumné otázky se pak dají koncipovat tak, že jazyky metaforicky chápeme jako jakési organismy s fylogenezí a ontogenezí, nebo dokonce seberegulující se bojler nebo zrající sýr (ke kritice podobných metafor z pohledu historické lingvistiky viz Bičovský (2010a,b, kapitola 1.4)).

Divokost těchto metafor vedla některé teoretiky (Yngve, 1986, 1996; Yngve – Wąsik, 2004) k úplnému odmítnutí onoho druhého pojetí jazyka,[9] problém je v tom, že prakticky každá generalizující hypotéza od mluvčích částečně abstrahuje, nebo alespoň redukuje některé jejich vlastnosti. Je to tedy spíše otázka míry a účelu, jakž bude

[8] Což je skutečný smutný stav jazyka Ayapaneco, alespoň pokud můžeme věřit dennímu tisku (http://www.theguardian.com/world/2011/apr/13/mexico-language-ayapaneco-dying-out).

[9] Takovémuto odmítnutí Itkonen (2003, kapitola 10) říká fyzikalistická lingvistika, kteryžto termín zde ovšem používat nebudeme, jednak proto, že je obvykle používán s pejorativním nádechem, což rozhodně nechci, jednak proto, že termíny fyzikalistický nebo naturalistický se používají v rozmanitých významech a kontextech (třeba podle Itkonena by pod něj spadala nejspíš i tato práce) a pojednáním o nich bychom příliš odbočili.

vysvětleno v kapitole o epistemologii (kapitola 2, zejména 2.3). Podobné abstrakce jsou konečně běžnou součástí našeho každodenního uvažování (Amerika zaútočila na Irák; krize ho připravila o dětství, fašismus o rodiče, stalinismus o zdraví a polistopadový vývoj o iluze…). Konečně i přírodní vědy, chtějí-li zkoumat složité vztahy většího množství vzájemně interagujících vlastností, uchylují se běžně k abstrakci od jejich nositelů nebo alespoň k zásadní redukci jejich vlastností (typické jsou v tomto například ekologické modely).

* * *

Stejným způsobem pak je problematický pojem *text*. Můžeme jím myslet jak konkrétní jazykem zakódovanou zprávu konkrétního člověka, tak (část) *parole*, jakousi instanci onoho abstraktního *langue*. Pokud texty chápeme tímto způsobem, zbavíme se nepříjemných problémů vzniklých tím, že značná část textů nemá jednoho autora, ale vznikla spoluprací několika mluvčích a přístrojů. Interpretace výsledků výzkumu je pak ovšem netriviální.

Mohlo by se zdát, že podobné otázky jsou spíše spekulativního charakteru, nicméně mají velmi reálné dopady na směřování výzkumu, konkrétní podobu a způsob testování hypotéz, operacionalizaci měření a experimentů, statistické zpracování a interpretaci výsledků. Dle mého názoru samotné množství vzájemně se ignorujících lingvistických disciplín je do značné míry dáno touto dichotomií.

Text se v lingvistice občas používá jako termín pro nějakou ucelenou, nebo dokonce ukončenou, koherentní informaci zakódovanou jazykem (například v pojmu *textová lingvistika*), nebo dokonce něco jako nejvyšší jazykovou jednotku. V tomto významu budu raději používat přímo termín *ucelený text*, i když chápu, že vždy může být něco nad textem — hypertext, diskurzivní rámec, bez čeho nebude dávat smysl, a že koherence není nikdy absolutní, že každý segment textu je potenciálně neucelený, žádná nejvyšší rovina neexistuje. Raději se budu snažit užívat přímo názvy segmentů, které mám na mysli: kniha, kapitola a podobně. Ostatně co bychom nazvali „uceleným textem“ v případě Zolova díla: kapitolu, část románu, celý román, nebo celý románový cyklus? Kniha je celek, jejíž rozsah je do značné míry určen přízemními technickými záležitostmi komunikace, proto máme tendenci s tímto pojmem v lingvistice nepracovat, ale nezapomínejme, že i nižší celky jsou omezeny technicky — třeba vyslovitelností, uchovatelností v paměti a podobně. I v mluvené podobě je omezení vyšších segmentů často dáno technickými možnostmi (třeba potřebou spánku).

Co se zašuměného kanálu týče, pro potřeby této práce ponecháme jeho přesné hranice jako otevřený problém. V každém případě za něj nebudeme považovat jen samotný přenos vzduchem nebo písmem, ale minimálně také fyzickou produkci a fyzickou percepci. Vzhledem k omezeným znalostem kognitivních procesů, které vstupují do

procesu kódování a dekódování v mozku, však těžko můžeme vhodně a bezrozporně vymezit, kde končí jazyk mezilidské komunikace a kde začíná signál překódovaný pro interní potřeby mozku a jestli je vůbec účelné takové vymezení postulovat. Vzhledem k návaznosti těchto dvou kódování a vzhledem k tomu, že jsou nejspíš utvářeny velmi podobnými prostředky, můžeme předpokládat, že tyto dva kódy budou korelovat, což je téma, které extenzivně studuje kognitivní lingvistika.

Kapitola 2

Lingvistika a věda vůbec

Věda je jedním ze způsobů získávání, uchovávání a šíření informace. Podobně jako v případě jazyka můžeme se na ni dívat jako na souhrn vlastností individuálních lidí, jako na rys, který je vlastní každému člověku zvlášť, ale také jako na abstraktní systém „nad“ těmito individuálními rysy.[1] V obou případech je věda systém pracující s informacemi, a tedy je na ni aplikovatelná teorie komunikace, jak jsme si ji popsali v předchozích kapitolách. To nám zároveň pomůže pochopit, že věda není v lidském konání nijak výjimečná a že se jedná spíše o způsob jednání, který používáme zcela běžně, neformálně a často nevědomky.

Od běžné exoterické vědy očekáváme, že informace budou dostupné komukoli, čímž vstupují do hry naše předchozí poznámky o jazyce jako takovém, v esoterických vědách, které se naopak snaží udržet informace uvnitř určité komunity (v dnešní době například vojenský výzkum), je situace ještě o něco komplikovanější, nicméně základní schéma zůstává velmi jednoduché: při vědeckém výzkumu dochází ke shromažďování informací, jejich kompresi a komunikaci. Všechny tři úzce související aspekty si nyní postupně rozebereme.

2.1 Shromažďování informací

Má podobu pozorování a experimentů a probíhá jak náhodně, tak záměrně na popud různých společenských i osobních zájmů a přibližně podle současné podoby vědy, nejen aktuálních znalostí, ale i jejích cílů, metod, klasifikací a definic. Mezi náhodou a záměrností si představme spojitou škálu.

[1] Také v tomto případě půjde o konstrukt považovaný mnohými za samozřejmý a jinými zcela zatracovaný. Jedním z mnoha zdrojů nesouladu mezi Popperem a Feyerabendem je právě to, s jakou samozřejmostí Popper tento konstrukt přijímá v rámci svého „třetího světa“ (Popper, 1979, str. 74), aniž by ovšem Feyerabend (1975) tyto zdroje nesouladu reflektoval (či vůbec Poppera citoval).

Tyto pozorované (nebo jinak vnímané) informace mají formu existenciálních tvrzení, takže je můžeme verifikovat, respektive měli bychom toho být schopni.

Obvykle se klade důraz na objektivitu/intersubjektivitu, což znamená, že tyto informace musí mít možnost vnímat/verifikovat více lidí. Vzhledem k tomu, že intersubjektivita není nikdy dokonalá, je celý proces verifikace složitý a závisí na důležitosti verifikované informace.

Různí filozofové vědy se shodnou na tom, že samotné kumulování existenciálních tvrzení není proces, který by sám o sobě mohl být nazván vědou. O tom, proč tomu tak je, bude pojednávat následující podkapitola.

2.2 Komprese (odstraňování redundance)

Prostá kumulace existenciálních tvrzení by znamenala vytváření prostého obrazu reality, ještě k tomu nedokonalého (kvůli nedokonalé intersubjektivitě). Respektive nutně pouze výseku reality, neboť informace musejí mít nějaký nosič, ten zabírá časoprostor a úplný popis vesmíru (bez komprese) by byl minimálně tak velký jako vesmír sám. Přičemž takováto deskripce by měla jen velmi omezené užití. Například by ji nebylo možno použít k predikci. Norbert Wiener tvrdí, že „nejlepším modelem kočky je další, pokud možno stejná, kočka“ (Rosenblueth – Wiener, 1945), avšak taková modelová kočka bude reagovat na podněty stejně rychle jako kočka modelovaná.

Je vhodné, aby z modelu byla odstraněna redundance, ať už kompresí ztrátovou, nebo bezztrátovou. Vzhledem k nemožnosti úplné intersubjektivity prakticky každá kompresní metoda, kterou vědci používají, je ztrátová.

Ono odstraňování redundance není nic, čím by se věda vymykala našemu normálnímu nakládání s informacemi. Naše smysly přijímají tolik informací, že brutální ztrátová komprimace je naprostá nutnost,[2] a náš mozek ji provádí, aniž bychom si to uvědomovali. Přičemž nemůžeme mluvit o nějakém dokonalém mechanismu, který by ji v mozku bez našeho vědomí zajišťoval; množství a druh komplexity, která je při takové kompresi spolu s redundancí odstraňována, je často příliš velké (tak vznikají předsudky), nebo naopak zůstane redundance a komplexity příliš mnoho (zahlcení informacemi). Současnou vědu můžeme pokládat za způsob, jak sběr a kompresi dat provozovat kontrolovatelně a explicitně.

Často vedle sebe existuje více pojetí stejného zkoumaného předmětu, která se liší právě tím, jaké druhy komprese užívají a jaké ztráty povolují. Tím máme na mysli například dvě vzájemně soupeřící hypotézy, jako byl svého času kopernikovský heliocentrický systém, který v přesnosti predikcí jen těžko konkuroval tehdy vyspělému

[2] K velmi hrubé představě může posloužit, že zrak dodává informace ze čtvrt miliardy tyčinek a čípků do mozku, který disponuje pouze cca 100 miliardami neuronů, které mají mezi sebou řádově 10 000 miliard synapsí.

systému ptolemajovskému, zato byl ale mnohem úspornější, neboť tehdy mainstreamový geocentrický systém počítal až s 80 sférami (Koperníkův systém tedy měl vyšší kompresní poměr za cenu větších ztrát komplexity, které se podařilo napravit až Keplerovi). Ale nejen to: v lingvistice vedle sebe existují různé způsoby zkoumání jazyka, které mají natolik fundamentálně jiné cíle, že je velmi obtížné vzájemně srovnávat jejich míru komprese nebo úspěšnost, což připomíná vhodnost feyerabendovského (1975) pojetí vědy. Jazykem, ať už to znamená cokoli, se zabývá zejména:

1. Lingvistika ve své klasické podobě od Pāṇiniho přes Sībawajha a de Saussura dodnes (patří sem i značná část korpusové lingvistiky), která abstrahuje jazykový systém od jeho uživatelů a zbytku světa, a tak si uvolňuje ruce k podrobnějšímu pohledu na vztahy mezi jazykovými subsystémy.

2. Lingvistika, která od jednotlivých mluvčích naopak neabstrahuje a zahrnuje je do systému, pokud možno se všemi jejich vlastnostmi a schopnostmi, zejména kognitivními (psycholingvistika, kognitivní lingvistika, neurolingvistika, sociolingvistika, altmannovsko-köhlerovská kvantitativní lingvistika).

3. Počítačové zpracování přirozeného jazyka (NLP), které abstrahuje od toho, co se děje uvnitř systému, a za „jazykový model“ považuje systém, který má stejné rozhraní jako systém jazykový a na stejné vstupy reaguje stejnými výstupy.

Samozřejmě tyto tři přístupy nejsou zcela oddělitelné a distinkce je spíše na úrovni proklamací, neboť abstrakce od mluvčích není binární veličina, stejně tak při NLP modelování jazyka se můžeme omylem dozvědět víc o tom, jak skutečně jazyk funguje, naopak NLP se inspiruje od skutečného fungování jazyka, ať již z fyzikalistického pojetí (neuronové sítě), nebo od modelů, jež jsou od mluvčích abstrahovány (dělení na subsystémy, kategorizace, zákonitosti).

* * *

Odstraňování redundance má mnoho podob, předmětem debat je obvykle kategorizace, hledání hypotéz a explanace.

2.2.1 Kategorizace

Samotná kategorizace je velmi efektivním způsobem komprimace informací. Například říci, že na komíně cihelny zahnízdili čápi, je mnohem úspornější, než ony ptáky pixel po pixelu popisovat. Popper (2005, zejména kapitola 1.4, str. 11) trvrdí, že kategorizace sama o sobě není vědou a že teprve ve falsifikovatelných hypotézách začne

hrát svou hlavní úlohu, ovšem bez sdílené kategorizace by bylo velmi obtížné už samotné shromažďování existenciálních tvrzení, a především by bylo prakticky nemožné rozhodnout o intersubjektivitě.

Je tedy velmi záhodno, aby byla kategorizace sdílená mezi lidmi, čímž se dostáváme k otázce *přirozených kategorií*. Myšlenka, že kategorie jsou jaksi předem dané, má nejspíš své kořeny v době, kdy bylo těžko představitelné, že by se sdílená kategorizace vyvinula a udržovala sama od sebe — což ovšem dnes umíme vcelku přesvědčivě modelovat (Oudeyer, 2006, str. 59). Představa, že žádné kategorie nejsou a priori dané a prakticky každá kategorizace je dána naším náhledem na realitu a větší či menší mírou konsensu a vzájemného porozumění, je v lingvistice dnes celkem běžná.[3] Ovšem to neznamená, že by každá kategorizace byla stejně dobrá. Některé kategorie a koncepty se dají velmi těžko sdílet, jiné lépe. Některé mají tradici, a tak je možné navázat na předchozí výzkum, i když pro ten náš by vyhovovaly kategorie třeba trochu jiné. Pro některé potřeby (například kvůli vyhledávání) je výhodné mít kategorie s víceméně stromovou strukturou (typicky biologická taxonomie).

Co je však nejdůležitější — aby měly kategorie schopnost komprese informací s co nejmenšími ztrátami, je výhodné, aby:

1. kategorizace vymezovala systémy, které mají vůči okolnímu světu co nejméně vazeb (respektive aby vazby okolí na prvky systému korelovaly tak, že jsou nahraditelné vazbou okolí na celý systém), a naopak aby co nejvíce vazeb pojilo prvky uvnitř systému;

2. a zároveň aby vazby uvnitř systému nebo vůči okolnímu světu byly obdobné pro všechny entity, které do dané kategorie řadíme;

3. těchto entit musí být tak velké množství nebo musí mít tak dlouhé trvání, aby se vyplatilo kategorii vůbec zavádět (zavedení kategorie o jedné entitě, která se okamžitě změní tak, že přestane spadat do oné kategorie, redundanci naopak zvyšuje).

Abychom zůstali u čápa z příkladu ze začátku podkapitoly, buňky náležící jednomu čápovi spolu vstupují do velkého množství interakcí a obvykle jejich vztahy s prvky mimo systém (mimo čápa) vzájemně korelují. Například poloha čapího srdce koreluje s polohou čapích jater, vztaženo třeba vůči komínu, na kterém hnízdí (pokud se játra rozletí někam jinam než srdce, čápa obvykle přestáváme nazývat čápem).

[3] I když skutečnost, že se lidé na hranici kategorií neshodnou, může vyvolat zděšení a představu, že ona kategorie je chybně definovaná, nebo že dokonce „neexistuje“ (citujme populárně naučný článek *Do syllables exist?* od Josephine Livingstone (http://www.theguardian.com/education/2014/jun/25/english-do-syllables-exist-linguists)), běžná lingvistika se s tím dokáže docela dobře vyrovnat, viz Haspelmathův článek *Pre-established categories don't exist — Consequences for language description and typology* (Haspelmath, 2007), v něm citovanou literaturu a ohlasy na něj.

Zároveň všechny systémy, které nazýváme čápem, mají dostatečně podobné vztahy jak uvnitř systému, tak vně. Díky tomu o nich lze velmi dobře vytvářet falsifikovatelné hypotézy, o kterých si povíme v následující podkapitole.

Samozřejmě v reálné lingvistice (a v biologii to není o nic lepší) nejsou žádné dvě entity totožné, a tak jakákoli tvrzení založená na kategoriích (tedy prakticky všechna) mají povahu spíše popisu tendencí než absolutních zákonů.

Kategorie vymezující konkrétní (nikoli abstraktní) entity (například čápy) jsou docela dobře definovatelné a dají se pohodlně sdílet. Můžeme se dohadovat o podstatě čápa (jeho čápovitosti) a o tom, jestli mezi čápy řadit i mrtvé čápy, a pokud ano, tak do jakého stádia rozkladu ještě, nebo jestli je potrava v čápově žaludku součástí čápa, nebo ne, ovšem valná většina lidí se velmi dobře shodne, pokud nějakého ptáka uvidí, jestli to čáp je, nebo není, a jaké má hranice (u vyšších živočichů se nejspíš shodneme na tom, že hranicí systému je kůže a její deriváty). Také je poměrně jednoduché určit, co je statistickou populací všech entit spadajících do stejné kategorie (jednoduše všichni čápi).

Abstraktní kategorie jsou vymezitelné nepoměrně hůře, už proto, že jejich vazby na okolní systémy jsou také hůře uchopitelné. Odtud pramení pochopitelná snaha chápat jazyk jako vlastnost lidí, a nikoli jako systém od těchto lidí abstrahovaný (viz kapitolu 1.6).

2.2.2 Hledání hypotéz

Pomocí non-existenciálních tvrzení[4] je možné velmi efektivně odstranit část redundance i komplexity. Tento typ výroků má výsadní postavení jak u Poppera (2005), který jim říká falsifikovatelné hypotézy a který jejich užití považuje za kritérium vědeckosti, tak u Kuhna (1997), který je také považuje za samozřejmou součást vědy.

Všimněme si, že jak kategorizace, tak hypotézy mohou mít obdobnou logickou strukturu:

V případě kategorizace říkáme: „Pokud to není bílé, tak tomu nebudu říkat čáp.“

V případě hypotéz obdobně: „Pokud je to čáp, tak je to bílé.“

Nejsou rozdílné ani tak výroky, jako spíše náš přístup k nim — když poprvé uvidíme černého čápa, tak v prvním případě máme možnost definici uchovat a prostě ho nepovažovat za čápa (i když i rekategorizace je možná), v druhém případě nám zbude vyvrácená hypotéza.

Mohlo by se zdát samozřejmé, že výroky, které jsou platné již z definice (na jejímž základě kategorizujeme), je nesmyslné testovat jako hypotézy. Nicméně kvůli zamlženosti definic v lingvistice k tomu občas dochází.[5]

[4] Tedy výroků, jež mají logickou strukturu „neexistuje čáp, který by nelétal“, což je ekvivalentní výroku „každý čáp létá“.

[5] Ještě složitější je situace, když přiřazení kategorie necháme na mechanismu, který nedokážeme

Vzhledem k logické podobnosti hypotéz a definic je otázka, jak určit, který výrok o daném objektu se stane součástí definice a který už bude hypotézou. A skutečně se mnohé školy nebo jednotliví lingvisté v tomto liší; i v případě klasických gramatik je rozlišení na definice a hypotézy implicitní a neostré.

Součástí definic bývají výroky stabilnější, zatímco testovatelné hypotézy obvykle zaujímají výroky, jejichž pravdivostí si nejsme zcela jisti (například „čápi jsou bílí“ bude součástí definice, zatímco „čápi nosí děti“ bude hypotéza).

Součástí definic jsou výroky o vlastnostech, které můžeme v typických případech přímo pozorovat (například „čápi jsou bílí“), zatímco v hypotézách jsou zmiňovány vlastnosti, které v typických případech potřebujeme predikovat (například že „maximální rychlost nenaloženého čápa je 70 km/h“).

2.2.3 Explanace

V tomto duchu budeme za kauzální explanaci považovat nikoli všechny implikace, ale pouze ty, které nějakým způsobem přispějí k odstraňování redundance nebo výpočetní složitosti. O hypotéze A (*explanans*) tedy řekneme, že vysvětluje hypotézu B (*explanandum*) jen tehdy, pokud z hypotézy A plyne hypotéza B a pokud je hypotéza A obecnější (kromě hypotézy B z ní plynou i jiné hypotézy) a nebo jinak informačně úspornější.[6]

Takováto definice vědecké explanace souhlasí s pohledem, který zastává třeba Hempel (1959), pro nějž je vědecká explanace jakýmsi svatým grálem vědeckého výzkumu. Tato kapitola však v souvislostech ukazuje, že kauzální explanace není nic jiného než efektivní způsob, jak naplnit potřebu komprimace informací.

Podobně můžeme pojímat i jednu z interpretací Occamovy břitvy: jako normativní výrok, který nás chce přimět k tomu, abychom udržovali co nejnižší množství redundance. Problém je, že prakticky každá komprese prováděná vědou je ztrátová. Z toho plyne, že Occamova břitva vždycky uřízne i kus toho, co by mohlo být dobré ponechat. Kompromis mezi odstraněním redundance a ponecháním komplexity by však byla jakási oportunistická verze břitvy, kterou bychom již onomu mnichu raději neměli připisovat. Také *estetické kritérium* používané ve fyzice a matematice má podobnou interpretaci, čemuž se budu více věnovat v kapitole 6.10.

---

přesně popsat. Například pokud statistický tagger vydedukuje z tréninkových dat pravidlo, že dvě předložky se vedle sebe vyskytují s velmi malou pravděpodobností, a následně označkuje korpus tak, že jednu ze sousedících předložek vždy označí za jiný slovní druh, pokud je to možné, pak výzkumník, který testuje hypotézu „dvě předložky se nikdy nevyskytují vedle sebe“, dojde k velmi zkresleným výsledkům. Takovou černou skříňkou může být i člověk, kterému ukážeme několik instancí kategorie a necháme ho, aby dále kategorizoval sám, per analogiam.

[6] Předmětem explanace může být samozřejmě nejen non-existenciální, ale i existenciální tvrzení.

2.3 Úloha izolace subsystémů a redukce

Jak bylo zmíněno v kapitole 2.2.1, jedním z hlavních důvodů, proč svět vně i uvnitř sebe modelujeme, je schopnost predikce jeho chování. Snižování algoritmické redundance našeho modelu nemusí vést ke snižování výpočetního času, právě naopak, v dobách, kdy nebyly k dispozici rozsáhlé výpočetní kapacity, byly i algebraické modely převáděny do tabulek, aby se snížil výpočetní čas. Dovedeno do důsledků, pokud bychom odhalili algoritmus, který stojí za vesmírem, tak pro jeho použití k predikci celého vesmíru bychom potřebovali výpočetní kapacitu stejnou nebo větší, než má vesmír.[7]

Z tohoto důvodu predikujeme jen lokálně a snažíme se jednotlivé systémy izolovat. Pro vymezení jednotlivých systémů je vhodné postupovat obdobně jako při kategorizaci (není náhoda, že systémy, které zkoumáme, jsou obvykle i entitou spadající pod nějakou kategorii, jejíž vymezení více méně kopíruje jejich hranice).

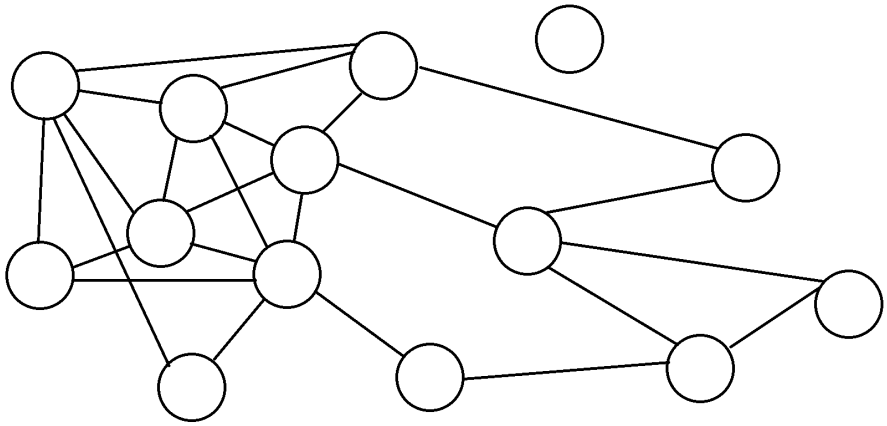
I zde se snažíme najít systémy, jejichž subsystémy mají mezi sebou vzájemně mnoho vazeb, zato nemají příliš mnoho zásadních vazeb na okolní systémy. Ideální je, když tyto vazby vně systému vzájemně korelují, a tedy jsou nahraditelné jednou vazbou systému na okolní systémy (obrázky 2.1, 2.2, 2.3 a 2.4).

Takovým systémům Ashby (1962, str. 117) říká „přirozené systémy“, nicméně je třeba si uvědomit, že systémy neexistují *a priori* a že rozdělení na systémy nám říká pouze to, že realitu takto rozdělenou mít chceme nebo potřebujeme, nikoli to, že bychom si mysleli, že je takto skutečně sama od sebe rozdělená.[8] Podobně jako u kategorií záleží i při rozdělování na systémy na tom, co vlastně chceme zkoumat a proč, zároveň je ale vhodné zachovávat kontinuitu s předchozím výzkumem.

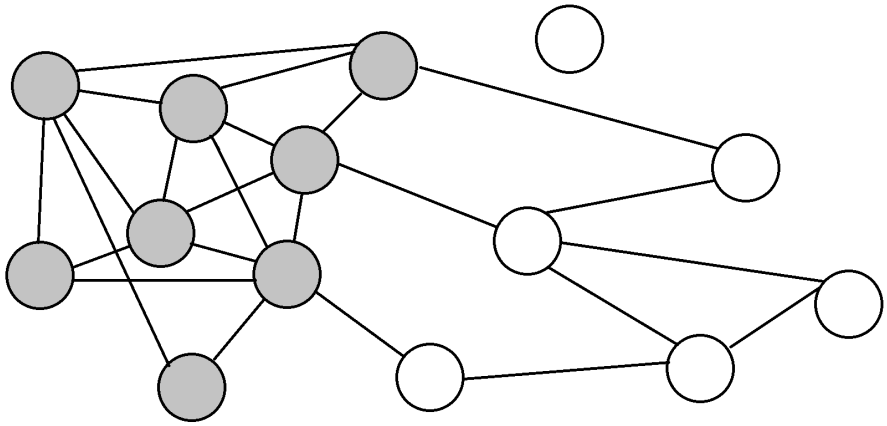
Oblíbený je redukcionistický přístup, který je založený na vnořených kategoriích se stromovou strukturou a který utvořením systému A nahrazuje všechny vazby mezi jeho subsystémy a systémy vně systému A vazbou na systém A (jak je to na obrázku 2.4), takže by se se systémem A následně jednalo jako s blackboxem, který má nějaké rozhraní a jehož vnitřek nás nezajímá. Takový čistý redukcionismus ovšem bývá v praxi (ať již záměrně, nebo nechtěně) alternován přístupem „holistickým“, který nezanedbává všechny vazby jdoucí přes hranice subsystému ani všechny subsystémy,

[7] Myšlenka, že vesmír má nějakou výpočetní kapacitu, se objevuje minimálně od 60. let (viz *Rechnender Raum* od Konrada Zuseho (2012), ostatně doporučuji celý sborník, ve kterém byl překlad této studie nově uveřejněn (Zenil, 2012)), dnes se jí seriózně zabývá Schmidhuber (1997a) a zejména Stephen Wolfram a jeho spolupracovníci (Wolfram, 2002, zejména kapitoly *Fundamental Physics*, *The Notion of Computation* a *The Principle of Computational Equivalence*).

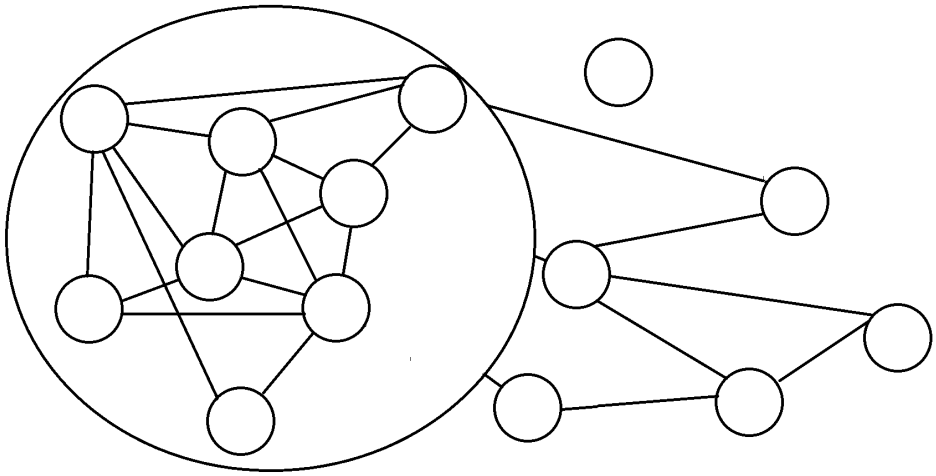
[8] Například Reinhard Köhler říká, že „[t]he statement that some object X is a system does not describe any property of X but says rather, that X is to be investigated with regard to certain aspects and by means of certain methods“(Köhler, 1987, str 241). Tedy tvrzení, že X je systémem, není tvrzením ontologickým, ale znamená, že výzkumník se chystá použít určitý druh modelování a analýzy objektu X. K tématu viz též (Milička, 2015).



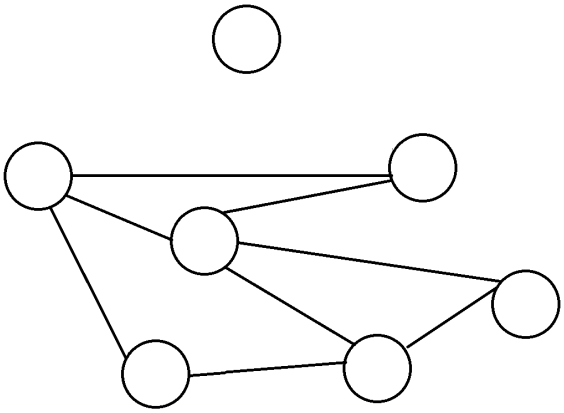
Obrázek 2.1: Jednotlivé uzly představují systémy a hrany (spojující čáry) reprezentují vazby mezi nimi. Pokud chceme zavést systém nad těmito systémy, musíme se rozhodnout, kudy vést jeho hranici. Je výhodné vést ji tak, aby hranice systému protínala co nejméně vazeb a aby byly nahraditelné vazbou na nově stanovený systém.



Obrázek 2.2: Šedé systémy mají mezi sebou mnoho vazeb a zároveň málo vazeb na systémy bílé. Pokud jsou vazby bílých systémů na šedé dostatečně korelované, pak může být výhodné jednat se všemi šedými systémy jako podsystémy jednoho ohraničeného systému, jak je ukázáno na obrázku 2.3.



Obrázek 2.3: Vazby systémů mimo nově definovaný systém na subsystémy onoho systému jsme nahradili vazbami na tento systém.



Obrázek 2.4: Tvorba nového systému završena, veškeré subsystémy jsou redukovány a je k nim přistupováno pouze přes rozhraní nově definovaného systému.

ale pouze ty, které jsou skutečně zanedbatelné nebo nahraditelné vazbou na subsystém (nebo si to autor modelu alespoň myslí). Dělení do systémů a subsystémů jednak umožňuje predikce a kontrolu výpočetní složitosti simulace, jednak zjednodušuje způsob testování modelu po částech, kdy můžeme zkoumat každou vazbu zvlášť například experimentálně, tedy tak, že měníme entity, které vazba spojuje.

2.4 Šíření vědeckých poznatků (a jejich uchovávání)

Způsob, jakým se poznatky vědy šíří a jak jsou uchovávány, spoluutváří povahu vědeckého výzkumu. Zmíněna již byla kontinuita kategorizace, kvůli které je nutné zachovávat kategorie, které jsou pro náš výzkum nevhodné, přičemž ona kontinuita často zahrnuje i nevědecký přirozený jazyk.[9] Tím také dochází k fixaci rozdělení na subsystémy.

Kategorizace je často uzpůsobena tak, aby ji bylo možno zapsat nějakým určitým formalismem. Takový požadavek může také omezit plynulou změnu kategorizace (například analytická rovina PDT (Hajič – Hajičová, 1997) má čistě stromovou větnou strukturu, kterou není možné narušit, i když různé prvky této kategorizace jsou při přechodu mezi jazyky variabilní).

Technický charakter i společenské podmínky pro šíření informací ovlivňují užité komprimační metody obdobně jako intence, s jakou k vědeckému poznání přistupujeme.

* * *

Za vědeckou činnost označuji prakticky každé poznávání, které může být sdíleno mezi lidmi.

Takový přístup k vědě je značně tolerantní a svým způsobem rezignuje na nějaké přesné kritérium vědeckosti, pouze udržuje kritérium kvality, a to je ještě vztaženo k účelu poznání. Zatímco Popper pociťoval silnou potřebu binárního rozlišení mezi Vědou a vším ostatním,[10] já v souladu s Feyerabendem tvrdím, že každé takové kritérium je nutně neostré a že je mnohem výhodnější pustit do vědy všechny, kdo tuto ambici mají, a následně nepoměřovat jejich vědeckost, ale kvalitu, přičemž kvalita se odvíjí od intersubjektivity vnímaných informací (kapitola 2.1), kompresního poměru a ztrátovosti komprese těchto informací a možnostech jejich sdílení.

[9] I když přírodním vědám se, díky všeobecnému školnímu vzdělání, podařilo mnohdy naopak přizpůsobit přirozený jazyk svým vlastním potřebám. Husarským kouskem například bylo přesvědčit společnost, že slepýš není had, pamouk hmyz a žralok ryba.

[10] To *ostatní* nazval metafyzikou.

Tato kritéria nejsou v rozporu s Popperovým dogmatem vyvratitelnosti, neboť nevyvratitelná obecná tvrzení narušují podmínku intersubjektivity. Díky tomu, že intersubjektivitu chápeme jako spojitou škálu, nemusíme rozlišovat jen mezi tím, kdy je intersubjektivní ověření principiálně nemožné („na špičku jehly se vejde maximálně deset andělů“), kdy je principiálně možné, ale prakticky neuskutečnitelné (status teorie relativity v nultých a desátých letech dvacátého století), a kdy je možné a uskutečnitelné. Můžeme totiž uvažovat i o různých mezistupních, například ne každý si může doma postavit vlastní urychlovač částic, a tak jsme závislí na víře v autority, přičemž různá tvrzení různých autorit mají pro různé lidi různou váhu, což má dle mého názoru v konečném důsledku mnohem větší vliv na utváření vědy než samotná logika zkoumání.

Tím se opět dostáváme k limitům, na které narazíme, pokud budeme chtít analyzovat vědu nebo jazyk abstrahovaně od vlastností lidí, kteří je používají.

Kapitola 3

Logická nezbytnost morfémů a segmentace na vyšší celky

Tato krátká kapitola klade si za cíl přesvědčit vás o platnosti jedné jazykové univerzálie a jednoho spekulativního tvrzení. První říká, že jazyk nutně musí obsahovat distinktivní rysy (tedy alespoň jeden distinktivní rys) a morfémy, respektive něco, co označuje kategorie a jejich instance, jak jsme si je vymezili v kapitolách 2.2.1 a 2.3; symboly, chcete-li. Těmto nezbytným součástem jazyka budeme říkat minimální segmenty. Druhé tvrzení pak říká, že ostatní celky naopak logickou nezbytností nejsou.

3.1 Nejnižší segmenty

3.1.1 Distinktivní rysy

Distinktivní rysy jsou nejnižší prvky jazyka, které nesou informaci, a jsou oním *rozdílem, na němž záleží* — jednotkou informace, která je podobně jako v digitálních technologiích shodou okolností nespojitá a bez níž bychom se už z podstaty věci neobešli (na rozdíl od hlásek či slabik). Distinktivní rys je obvykle vymezen na spojité škále a je zajímavé, že se nakonec na této škále ustálí jen několik málo opozic (nejčastěji jenom jedna binární opozice, jako třeba nazální/nenazální, ta pak vyjadřuje právě jeden bit); tuto vlastnost považujeme za samozřejmost, nicméně nijak samozřejmá není a není zas tak jednoduché vytvořit vývojový model, který by k takovým výsledkům sám od sebe dospěl (jeden takový model vytvořil Oudeyer (2006)).

Přitom opozic by mohlo být, vzhledem ke schopnostem lidských smyslů, na jedné škále mnoho. Trénovaní lidé s absolutním sluchem jsou schopní rozlišovat několik desítek půltónů,[1] což by stačilo na reprezentaci všech různých segmentů odpovídajících hláskám, přesto dokonce ani hvízdavé jazyky, které používají výšku tónu jako hlavní

[1] Předpokládám, že by nebyl problém, aby se absolutní sluch rozšířil na celou populaci, kdyby to

škálu pro distinktivní rysy, mívají pouze dvě až čtyři opozice.[2] Výsledkem je tolerance k chybám a variacím už na této úrovni, samozřejmě nejen ve hvízdavých jazycích.

3.1.2 Morfémy

Morfémy bývají definovány jako nejmenší jednotky, které nesou význam.[3] Možná bude lepší toto tvrzení mírně modifikovat a považovat je za nejmenší jednotky, které označují nějakou kategorii nebo její instanci.

Vraťme se ke kapitole 2.2.1, která pojednává o tom, že kategorizace je efektivním způsobem komprese informací, a ke kapitole 2.3, která ukazuje, jak hranice kategorií souvisí s hranicemi systémů, které zkoumáme. Již tam jsem připomínal, že takové komprimování informací není nic specifického pro vědecké poznání, právě naopak, že je nám vlastní a že ho podstupujeme, aniž bychom o tom dlouze přemýšleli.

V každém jazyce tedy najdeme nějaké nejnižší segmenty, které reprezentují tyto kategorie, přičemž tím rozhodně nechci říci, že by kategorizace, jakou používáme v jazyce, musela kopírovat kategorie, tak jak je používáme interně při přemýšlení.

Tezi, že každý jazyk nějak vyjadřuje kategorie (kategorie ve smyslu kapitoly 2.2.1), můžeme obhájit jednak empiricky, jednak prostou dedukcí z teorie informace. Neexistence takových jednotek by totiž vedla k nereálným důsledkům: i kdyby mluvčí místo prosté věty „viděl jsem červeného čápa“ popisoval barevné vjemy na každé tyčince a čípku svých očí, asi jako když se přenášejí data z digitálního fotoaparátu do počítače, neubránil by se kategorizaci, neboť barevnou škálu, která je spojitá, by musel nějakým způsobem kategorizovat (běžný monitor používá 2^{18} nebo 2^{24} kategorií).[4] Dokonce i kdyby bylo technicky možné přeložit spojitou škálu barev na spojitou škálu nějakého distinktivního rysu (například kdyby se frekvence světelných vln nějak namapovala na frekvenci zvuku, který vydáváme), neubránili bychom se kategorizaci vzhledem k nespojité povaze našeho smyslového aparátu (ony zmiňované tyčinky a čípky, nervová zakončení).

Když se podíváme na skutečný text, zjistíme, že v mnoha jazycích je segmentace slov naznačena v psaném projevu graficky, v mluveném pak na jejích hranicích mohou být různé pauzy, nádechy, hranice též bývají naznačeny intonací. Oproti tomu hranice morfémů nemusí být naznačeny vůbec a pro běžného mluvčího indoevropského nebo

bylo evolučně výhodné. Někteří autoři dokonce tvrdí, že absolutní sluch není nic výjimečného a že schopnost dlouhodobě si pamatovat absolutní výšky tónů je v populaci běžná (Levitin, 1994).

[2] Mluvím o jazycích Riallandem kategorizovaných jako *formant-based whistled languages* (Rialland, 2005, str. 238).

[3] Slovo *význam* je silně zatíženo a je používáno v mnoha vzájemně málo souvisejících úlohách. Solidní úvod do různých významů *významu* v dějinách lingvistiky i „v jejich současnostech“ naleznete v (Geeraerts, 2010). Přesto doufám, že čtenáři tomuto tvrzení porozumí tak, jak jsem ho zamýšlel.

[4] Otázka, jakým způsobem dochází ke kategorizaci barev přímo při percepci a jak při komunikaci, je dnes intenzivně studována. Na toto téma doporučuji (Ocelák, 2015), který přináší kritický náhled na poslední vývoj na tomto poli.

semitského jazyka není otázka, jak a proč segmentovat text na slova, věty a souvětí, ale spíše (pokud nemá lingvistické vzdělání) pro něj paradoxně může být neintuitivní myšlenka, že by mohlo existovat něco jako morfémy. Tím se liší od mluvčího čínštiny, který na morfémy segmentuje při zápisu textu v čínském písmu.

Pro myšlenku neintuitivnosti a nepůvodnosti morfematického členění hovoří to, že v lingvistickém popisu bylo vytvořeno velké množství vzájemně nekompatibilních metod, jak slova na morfémy rozčlenit,[5] přičemž variabilita je skutečně ohromující, zvláště v české lingvistické tradici, kde různé proudy postulovaly různé nulové morfy.[6] Arabská nonkonkatenativní morfologie se svou stabilní tradicí k již zmiňovanému Sībawajhovi je proti tomu příjemně logickým systémem. Proti hovoří naopak fakt, že běžný mluvčí morfologickou derivaci bez problémů spontánně používá a slovotvorba není záležitostí několika vyvolených lingvistů nebo dlouhodobého vývoje jazyka, naopak i na poměrně malém množství textů (řádově stovky milionů slov) se velmi zřetelně projevuje, že počet slovních různých slov je nekonečný, tedy že se asymptoticky neblíží nějaké konstantě (Milička, 2013).[7]

Ostatně přístupnost morfologické slovotvorby při běžné komunikaci je velkým tématem psycholingvistiky od 80. let, kdy byla nastolena otázka, jestli si člověk pracuje s celými slovy, jak jsou, nebo jestli je skládá z morfů přímo při produkci textu, respektive do jaké míry kterého z těchto dvou přístupů používá. Pro běžného indoevropského studenta arabštiny je těžko představitelné, že by semitská nonkonkatenativní morfosyntax mohla být v myslích rodilého mluvčího aktivně užívána a že by na místě vkládal jednotlivé morfémy podle složitých pravidel do sebe.[8] Nicméně jak případové

[5] Metodologicky ne právě čistá, nicméně velmi inspirativní je v tomto ohledu bakalářská práce Květy Mrštíkové (2014), která dokládá, že pokud necháme 100 českých rodilých mluvčích (ať již lingvisticky školených, nebo ne) rozčlenit 1000 slov na morfémy, dočkáme se velmi různorodých výsledků.

[6] Pro představu doporučuji například (Faltýnek, 2011, zejména od strany 96).

[7] Kromě matematického přístupu mohu nabídnout i anekdotický doklad, kdy se během studií arabistiky ve třídě spontánně vytvořil arabistický slang, který spočíval v derivaci nových českých slov z arabských kořenů. Například slovo „okraj“, „periferie“ se arabsky řekne *hāmiš*; prosím laskavého čtenáře, aby si tipnul, jak se v našem slangu řeklo „odstranit někoho na okraj“. Pokud hádáte, že *vyhámišovat*, máte pravdu. Jedna spolužačka byla před magisterskými státnicemi přesvědčená, že ono slovo je úplně normální součástí češtiny, a používala ho i před nearabisty (což byl u některých podobných slov i můj problém). Zajímavé je, že sloveso *vyhámišova*t sémanticky odpovídá arabskému slovesu *tahāmaša*, které je odvozeno od stejného kořene. Jednoslovné české vyjádření pro tento význam mi dodnes občas chybí.

[8] Velmi krátce shrnu arabskou morfologii: podle tradičního pohledu arabských gramatiků (od již několikrát zmiňovaného Sībawajha po dnešní učebnice arabštiny) se typické arabské plnovýznamové slovo skládá z kořene, který má 2–5 souhlásek, nejčastěji 3. Tento kořen, který nese poněkud mlhavý význam (například kořen KTB má něco společného s psaním, kořen ḌRB pak s udeřením), je kombinován s *šablonou*, která určuje slovní druh a blíže specifikuje význam, například šablonou $\text{m}\alpha\text{R}_1\text{R}_2\alpha\text{R}_3\alpha$ se vytvoří podstatné jméno značící místo, které má s daným kořenem něco společného; *maKTaBa* = knihovna. Teprve k tomuto celku se přidávají gramatické koncovky, popřípadě může být afigován i slovotvorně.

studie na afaticích, tak zkoumání dětských jazykových her ukazují, že semitští mluvčí nejspíš skutečně skládají slova z morfémů tak, jak je chápe arabská gramatika.

V tomto ohledu nejpřesvědčivější mi přijde studie (Prunet et al., 2000). Ta podrobně rozebírá případ arabsko-francouzského afatika, který ve svých arabských jazykových projevech přesmykával hlásky a písmena v arabských slovech pouze v rámci trojkonzonantního kořene, tedy například namísto slova *iḥtimāl* vyslovil slovo *iḥtilām* (přesmyčka je v rámci trojkonzonantního kořene ḤML a nezasahuje do *šablony* $\mathrm{i}\mathrm{R}_1\mathrm{t}\mathrm{i}\,\mathrm{R}_2\bar{\mathrm{a}}\mathrm{R}_3$). Naopak ve svých projevech francouzských žádnou šablonu nerespektoval a přesmykával souhlásky se samohláskami. Dále se zabývá jedním afatikem — rodilým mluvčím hebrejštiny, který naopak zaměňoval šablony, avšak kořeny zůstávaly netknuty.

Tato zjištění není možné odbýt tím, že jde o nemocné a že běžní mluvčí nakládají s morfologií jinak, neboť studie dále uvádí, že podobným způsobem přesmykávání se vytvářejí argotické jazyky odvozené od arabštiny a že podobných záměn se v rámci přeřeknutí dopouštějí i zdraví mluvčí.

Dalším argumentem pro to, že mluvčí aktivně používají dekompozici na morfémy, je schopnost dodržovat pravidla, která platí na morfematických švech. Například české protetické *v* je přítomno i vevnitř slova, pokud obsahuje hranici morfému, který začíná na *o* (například slovo *zvorat*). Podobně v německém pravopise se *st-* na začátku morfu vyslovuje jako *št-*, aniž by toto pravidlo mluvčím činilo potíže, takže když vidí napsané neznámé slovo, řekněme Johannestal, jsou schopní správně určit morfematický šev, respektive se na něm shodnou s dalšími mluvčími němčiny.[9]

3.2 Ostatní segmenty

Na rozdíl od segmentů minimálních, ostatní segmenty logickou nutností nejsou. Toto tvrzení nebudeme dokazovat na základě příkladu jazyka, který takové jednotky nemá (a v příští kapitole 4 si vysvětlíme proč), ale tím, že si ukážeme, že se bez takové segmentace dá docela dobře obejít.

Naopak neobejdeme se bez několikanásobné vnořené sémantické interakce nejnižších segmentů (syntaktici hovoří o rekurzi), což je způsob, jak vytvářet libovolné množství výpovědí za použití omezeného množství nejnižších jednotek — vlastně se jedná o vnořenou kompozicionalitu.

[9] Samozřejmě netvrdím, že je tato schopnost absolutní nebo že je jí užíváno vždy, příkladem budiž etymologicky průhledné slovo *objed* (předpona *ob-*, kořen *jíst* — Fiedlerová (1978)), jež se, navzdory tomu, že český pravopis systematicky zachovává písmeno *j* na počátcích morfémů, běžně zapisuje jako *oběd*.

3.2.1 Kompozicionalita distinktivních rysů

Dokážeme si představit jazyk, který se obejde bez kompozicionality na úrovni distinktivních rysů? Když dovedeme do důsledků úvahy v kapitole 3.1.1, je možné si představit hvízdavý jazyk, jehož mluvčí jsou schopní produkce i percepce rozdílů v setinách půltónu, takže výška tónu by byla distinktivním rysem rozdělujícím přímo morfémy. Kompozicionalita distinktivních rysů by tedy splývala s kompozicionalitou morfémů. Například hvízdnutí okolo 440 Hz by znamenalo „delfín“, kdežto 450 Hz „skočit“, přičemž i škála by mohla zůstat spojitá a i mezistupně by mohly nést význam, například 440,1 Hz by znamenalo něco jako „mohl by to být delfín, ale nejsem si tím tak docela jistý“.

Takovému jazyku by logicky nic nebránilo v implementaci, ovšemže s biologicky danými omezeními, neboť počet morfémů by se odvíjel od percepčních schopností. Snad proto žádný takový lidský hvízdavý jazyk neexistuje a v hvízdavých i v běžných jazycích se morfémy vyjadřují kompozicí distinktivních rysů — a to nejen sekvenčně, ale i paralelně kompozicí různých distinktivních rysů na různých škálách, přičemž došlo k velmi kreativní exaptaci dutiny ústní a přilehlých oblastí.

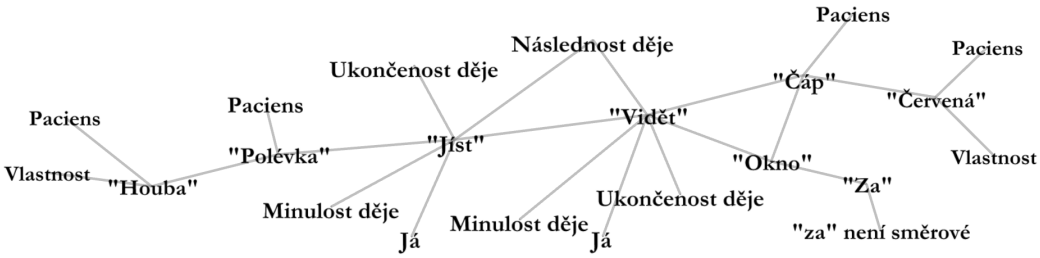
Všimněme si, že oba dva přístupy se obejdou bez dělení na hlásky a slabiky. Ke kompozicionalitě distinktivních rysů by stačila jediná binární opozice na jedné škále, například tón od určité frekvence vnímaný jako vysoký a od určité jako nízký. Variace těchto tónů v čase by vyjadřovala přímo morfémy.

3.2.2 Kompozicionalita morfémů

Kompozicionalita na úrovni morfémů je schopnost významu jedné jazykové jednotky interagovat s významem jiné jazykové jednotky, a tak se vzájemně zpřesňovat (například už zmiňovaný pojem *červený čáp* vznikl spojením kategorie červené barvy a kategorie čápa) nebo úplně měnit (například *pomazánkové máslo*, které není máslem, nebo *bílé víno*, které není bílé). Přičemž jestli jde o zpřesnění nebo změnu významu, není binárně rozdělitelná, ale spojitá škála, vzpomeňme nejasné hranice mezi koncovkami (které význam morfému, ke kterému se vážou, zpřesňují) a příponami (které význam mění). Nebo neurčité hranice mezi idiomy (které jsou definované jako sousloví, které má jiný význam než slova, ze kterých se skládá) a jinými ustálenými spojeními.

Představme si poněkud zjednodušenou síť vztahů mezi nejnižšími segmenty, asi jako na obrázku 3.1. Je to jen jeden z mnoha způsobů, jak tuto síť nakreslit, předpokládám, že byste ji konstruovali úplně jinak (včetně výběru toho, co je nejnižší segment). Přestože jsem se snažil, aby síť nepřipomínala žádný syntaktický formalismus, určitě jsem byl ovlivněn pohledem na syntax, jak mi byla předkládána již od základní školy.

Samotná síť je myslím srozumitelná a při troše cviku by bylo možné číst ji jako



Obrázek 3.1: „Když jsem dojedl houbovou polévku, uviděl jsem za oknem červeného čápa.“

normální text, přičemž by ji bylo možné bez problémů rozšířit. Zkusme si takový jazyk předstzavit. Jazyk, který se skládá pouze ze symbolů a markerů vztahů mezi nimi, ať už vztahů významových — kompozičních, jak je naznačeno v grafu, nebo třeba deixe. Vím, je to těžké, neboť je to proti všem vašim dosavadním zvyklostem. Takovému jazyku logicky nic nebrání v existenci, naopak, ušetřily by se informace nutné pro značení hranic segmentů (není možné určit přesně, která hláska do těchto informací patří, nicméně už z principu takovéto hranice musí být přítomny, více v kapitole 7) a na všech úrovních by se pro stejné vztahy mezi symboly mohly používat stejné markery.[10]

Obdobný způsob kódování se používá v již zmiňovaném formátu SVG (v kapitole 1.5) , který také nemusí být uvnitř nijak strukturovaný. Pokud vám připadá kódování SVG příliš omezené, jako další příklad lze uvést jazyk symbolických adres (jazyk pro assembler), který se též obejde bez jakékoli vnořené segmentace, lze jím vyjádřit prakticky každou myšlenku, kterou byste chtěli svému procesoru sdělit (respektive kterou by procesor mohl přijmout), a komplexita zpráv jím kódovaných není limitována. Oba tyto kódy mají něco společného: nemají za úkol chránit data při průchodu zašuměným kanálem (k tomu slouží jiná kódová vrstva) a reální lidé je neradi používají napřímo. Raději kódují své zprávy určené procesoru ve vyšších programovacích jazycích, které segmentují zdrojový kód do procedur a funkcí, objektů, modulů, projektů…

A právě o tom, jak univerzální je potřeba vnořené segmentace v lidských jazycích, bude pojednávat následující kapitola.

[10] V češtině jsou různá pravidla a markery vztahů ve větné syntaxi, syntaxi a morfosyntaxi, ačkoli kódované informace si odpovídají. Další sadu pravidel a markerů máme pro nadvětné struktury, které jsou též obdobné strukturám v rámci vět, v této souvislosti nemůžeme nezmínit klasickou práci na tomto poli (Halliday – Hasan, 1976). Například vztah příčiny a následku můžeme vyjádřit jak na úrovni jedné věty (*Kvůli časté absenci dostala dvojku z chování.*), tak na úrovni souvětí (*Dostala dvojku z chování, protože často chyběla.*), tak s použitím dvou souvětí *(Dostala dvojku z chování. Často totiž chyběla.*), aniž by se nějak radikálně změnil význam.

Kapitola 4

Vnořená segmentace — její univerzálnost a variace

V minulé kapitole jsme se zamysleli nad tím, že v jazyce musí existovat nejnižší segmenty a že jakákoli další segmentace logicky nutná naopak není. Navzdory tomu se vyšší segmentace v jazycích vyskytuje, a to jak z pohledu jejich uživatelů, tak z pozice těch, kdo ony jazyky popisují.

Univerzálnost vyšší segmentace v reálném jazyce bude pojednána v následující podkapitole, nyní se krátce zastavme u univerzálnosti kategorizace vyšších segmentů v jazykovém popisu.

4.1 Segmentace podle lingvistů

Kategorie, jak je používají lingvisté, sice vychází z kategorií, jak je vnímají mluvčí (konec konců i lingvisté jsou mluvčími nějakého jazyka), nicméně tyto kategorie se shodovat nemusejí a často neshodují.

Důvod je jednak snaha o univerzalizmus kategorií nebo alespoň funkcí, které prvky spadající do té které kategorie mají plnit (k těmto snahám doporučuji Haspelmath, 2011), nebo jednoduše to, že lingvisté popisující ten který jazyk jsou mluvčími jazyka jiného (angličtiny v případě prvních generativistů, perštiny v případě klasických arabských gramatiků…),[1] popřípadě jsou ovlivněni kategoriemi jazyka, jehož sice nejsou rodilými mluvčími, ale jehož popis znají (typicky latina a řečtina pro evropské lingvisty až do začátku dvacátého století).

[1] Toto ovlivnění je samozřejmě přirozené a objevuje se i mimo lingvistický popis. Například rodilí mluvčí angličtiny mají, navzdory arabskému písmu, tendenci přepis arabských jmen segmentovat na co nejnižší jednotky, například *Abd al Rahman*, zatímco čeští rodilí mluvčí, zvyklí na flektivní jazyk se slovy skládajícími se z mnoha morfů, naopak píší tato jména spíše celá dohromady, tedy *Abdarrahmán*, popřípadě klitika oddělují pomlčkou, podobně jako české klitikon „-li“. K této otázce se mi nepodařilo dohledat žádnou empirickou studii, proto tuto poznámku berte spíš jako můj osobní dojem.

Zde je zajímavé, že se ke kategorii něčeho jako slova a něčeho jako věty dostaly i lingvistické tradice neevropské a na Evropě nezávislé, například už zmíněný Pāṇini respektoval hranice slov (Keidan, 2007).

My se zastavíme u zajímavé tradice lingvistů, kteří konstruovali gramatiky a lexikony arabštiny, jejíž nejvýznamnějšími představiteli byli lexikograf al-Ḫalīl ibn Aḥmad al-Farāhidī a gramatik Sībawajhi (Zemánek, 2007, str. 136), pro bližší informace o těchto osobnostech a jejich způsobu vědeckého myšlení doporučuji Versteegh (1997, kapitoly 3 a 4).

Vzhledem k povaze arabského písma, jež má tendenci spojovat písmena, která patří do stejného slova (z pozice mluvčích), nás nepřekvapí, že jak v al-Ḫalīlově slovníku Kitāb al-ʿajn, tak v Sībawajhově gramatice (Bakr, s.d., s. 21–116) je *slovo* (*kalima*) pojímáno stejně jako dnes, tedy graficky, přičemž je striktně rozlišován rozdíl mezi gramatickými afixy a klitiky, která se ke slovu přidružují. Mnohem zajímavější je, že podobná tradice se drží i v případě vět, respektive celků, které zhruba odpovídají dnešnímu pojetí věty, jež v té době nebyly značeny interpunkčními znaménky.[2]

V každém případě podobně jako v evropské tradici (Haspelmath, 2007) hranice segmentů nejsou uspokojivě definovány. Přesto můžeme najít rys, který je společný různým tradicím: vnořenost. Tedy hranice vyšších celků respektují hranice celků nižších, například hranice věty nemůže být uprostřed slova, hranice souvětí uprostřed věty. Tato univerzálie jazykového popisu je platná i pro segmentaci, jak ji uznávají přirozené jazyky. A právě na segmentaci, jak ji používají přímo mluvčí a pisatelé, se zaměříme v následující podkapitole.

4.2 Segmentace podle mluvčích

Konečně můžeme odstoupit od segmentů, jak jsou definovány v lingvistice (nebo spíše jak definovány nejsou), a přistoupit k segmentům tak, jak je vnímají přímo mluvčí a pisatelé. Jejich náhled bude lingvistickým pohledem jistě ovlivněn, například téměř všichni rodilí mluvčí češtiny se učili ortografii na státních školách, kde byli indoktrinováni učiteli, kteří měli jakés takés lingvistické vzdělání. Nicméně lingvistické kategorie jsou taktéž do značné míry ovlivněny kategorizací běžných mluvčích, takže je třeba si přiznat, že lingvisté nestojí nad jazykem, ale jsou přirozenou součástí jeho vývoje (podobně jako ekologové často spoluutvářejí biotopy, které následně zkoumají, a jako ekonomové zasahují do ekonomiky).

Hlavním důvodem, proč dělám tak výrazný předěl mezi lingvistickým pojetím segmentů a jeho pojetí v přirozeném jazyce, je to, že v přirozeném jazyce segmentace mohla existovat nezávisle na tom, jestli vůbec existovala kategorie pro její popis.

[2] I to je nicméně pochopitelné, pokud si uvědomíme, že Sībawajhi byl a je chápán jako stěží kritizovatelná autorita ve všech ohledech, co se arabské gramatiky týče (Versteegh, 1997, str. 38).

Například ugaritština byla ve svém zápise rozdělována do slov, jak je přibližně chápeme dnes, v dobách, kdy slovo pro *slovo* (*hwt*) znamenalo něco jako *promluva* (Ward, 1969). Podobně arabské slovo *kalima* označovalo lingvistický termín pro slovo už od časů Sībawajha, ovšem v běžné arabštině mělo též význam *promluva*, *úsek řeči* a třeba ve frázi *ʾalqā kalimatan* (přednést projev) se tak používá dodnes.

Stejně jako segmenty lingvistického popisu, ani segmenty přirozeně se vyskytující v textu nejsou jednoznačně dané. Variabilní jsou jak v rámci jednoho jazyka (například slovo *například*, které někteří mluvčí češtiny chápou jako slova dvě a také ho tak bez problémů napíší), tak samozřejmě mezi jazyky, kdy jsou jednotlivé hladiny vlastně nesrovnatelné — českému *slovu* by běžný mluvčí nejspíš přiřadil arabský pojem *kalima*, neboť to je také segment, jehož hranici tvoří mezera v textu a rozpojení spojitého písma. Těžko ovšem říct, čemu *slovo* odpovídá třeba v čínštině. Nicméně ani v případě zcela srovnatelných a příbuzných jazyků, jejichž pojetí slov i lingvistická tradice jsou si blízké, si segmentace odpovídat nemusí, ať už máme na mysli segmentaci v psaném textu, nebo mluveném (naznačenou třeba přízvukem), viz segmentaci záporky *ne* / *ne-* ve slovanských jazycích.

Segmentace může být nejen variabilní, ale i stabilně nesystematická, například německé odlučitelné předpony jsou chápány jako afixy, jsou-li v pozici před slovem, ke kterému se sémanticky váží, a jako (přinejmenším graficky) samostatná slova, pokud jsou kdekoli za ním. Také stojí za zmínku, že grafická segmentace v zápisu textu se může systémově lišit od segmentace v mluveném projevu, jak ji můžeme odhadovat třeba podle slovního přízvuku a melodie věty.

Segmentace na vyšší celky, nikoli nutně na slova, věty a souvětí, ale na nějaké podobné vnořené segmenty, byla vlastní už mluvčím antických jazyků, jak můžeme usuzovat z jejich zápisu.

Egyptský hieroglyfický písemný systém sice hranice slov explicitně nevyznačoval, nicméně úlohu markerů hranic mezi slovy částečně plnily determinativy (Coulmas, 1999, str. 550). Tuto úlohu ostatně determinativy do značné míry plní všude, kde se používají, přičemž elamský klínopisný systém v pozdější fázi svého vývoje jeden determinativ nadužíval právě proto, aby plnil funkci markeru hranice slov (Coulmas, 1999, str. 142).

Mezopotámské klínopisné soustavy hranice slov sice běžně neznačily, nicméně z lexikálních seznamů a hranic řádků můžeme usuzovat, jakým způsobem mluvčí koncept slova chápali (Coulmas, 1999, str. 550). V případě akkadštiny můžeme tento koncept srovnávat se semitskými jazyky zaznamenávanými písmem hláskovým[3] a konstatovat,

[3] Už tabulky psané ugaritským písmem zaznamenávají hranice slov pomocí svislých čar (Coulmas, 1999, str. 523).

že je s mírnými variacemi prakticky ekvivalentní s konceptem, který se prostřednictvím arabštiny a hebrejštiny dostal až do dnešních dnů. Snad prvním písemným systémem, který zaznamenával hranice slov i vět, bylo v devátém století př. n. l. písmo Moábců, kteří slova oddělovali tečkami a věty pomlčkami (Coulmas, 1999, str. 340).

Jinak je grafické značení hranic vět v tomto areálu spíše výjimka, nutno ovšem dodat, že semitské jazyky naznačují hranice vět lexikálně a třeba klasická arabština disponuje systematicky užívanou partikulí *wa-*, která naznačuje novou větu souslednou s větou předchozí, a partikulí *fa-*, která značí následnost nebo následek, popřípadě změnu podmětu. V akkadštině je, vzhledem k málo frekventovanému užívání spojek, situace poněkud složitější, k rozlišování můžeme užívat konektivní (respektive ukončovací) partikuli *-ma*. V novoasyrštině, kde se tato partikule nevyskytuje, pak najdeme partikuli *-mā*, která segmentuje text na zajímavé celky, které leží někde mezi naším dnešním pojetím věty a souvětí, zároveň také uvozuje přímou řeč (Worthington, 2006).

Hranice souvětí, respektive něčeho, co bychom dnes nazvali odstavcem, byla naproti tomu v areálu běžná. Jak takové segmenty vypadaly se můžeme dodnes přesvědčit pohledem na verše ve Starém zákoně,[4] verše v Koránu (*ʾājāt*) nebo jednotlivé články Chammurabiho zákoníku. Tyto celky respektovaly hranice vět, jak je chápeme dnes.

Hranice slov srovnatelné s tím, na co jsme zvyklí z arabštiny, syrštiny nebo hebrejštiny, byly v písemnostech areálu značeny v různé míře a různě. S revolučním přístupem ke značení slov přišli pisatelé nabatejského písma (ze kterého se vyvynulo písmo arabské), které neznačí hranice slov, ale naopak spojuje písmena, která patří do jednoho slova (pro ucelený podrobnější pohled na vývoj viz Gruendler, 1993). Inventář ligatur se rozšiřoval postupně a hranice slov nejsou tímto způsobem jednoznačně a vždy určitelné ani v arabském písmu, pročež je tento systém doplňován mezerami mezi slovy.

Zásadní však je, že hranice slov značil písemný systém fénický (Coulmas, 1999, str. 401), ze kterého se vyvinula alfabeta a latinka. Značení segmentů původně převzali jak Řekové, tak Římané. Rané řecké texty měly také rozvinuté značení segmentů vyšších než slova: dvojtečka byla občas užívána pro značení konce věty, svislice pro konec odstavce a pomlčka oddělovala repliky jednotlivých postav. V helénské Alexandrii byly dále vyvinuty značky pro pauzy; obdobné značení bylo v různých variacích přejato pro zápis latiny (Coulmas, 1999, str. 421).

Nicméně v obou jazycích se postupně ujal způsob zápisu, kterému říkáme *scriptura continua*, tedy souvislý text bez jakýchkoli náznaků konce slov (Coulmas, 1999, str. 456).

Tento vývoj můžeme vysvětlit tak, že v semitských písemných soustavách zaznamenávání hranic segmentů kompenzovalo ztrátu informace, ke které docházelo při

[4] K problematice dělení na verše, *parašot* a vyšší významové celky ve Starém zákoně v celé své složitosti doporučuji (Korpel et al., 2007; de Hoop et al., 2009) a další svazky z edice *Pericope*.

odstraňování samohlásek (zapsání textu písmem bez samohlásek můžeme interpretovat jako ztrátovou kompresi). Oproti semitským písmům té doby měly alfabeta a latinka výhodu v tom, že samohlásky zapisovaly, a tak takovouto kompenzaci nepotřebovaly — oba jazyky mají charakteristické koncovky u téměř všech významových slov a další značení hranic segmentů v grafické reprezentaci je do značné míry redundantní).

Nicméně grafické značky pro hranice slov se do latinky vrátily už v raném středověku,[5] nejspíše kvůli zhoršení zašumění kanálu mezi produktorem a recipientem textu (jednak se prodloužila doba mezi produkcí a recepcí, jednak recipienti začali méně ovládat kód, který produktor používal) a nutností přidání větší redundance, přičemž segmentace na slova je jeden z velmi mála způsobů přidávání redundance, které můžeme použít při přepisování již hotového textu, aniž bychom ho radikálně měnili. Podobným způsobem a z podobného důvodu byla do Koránu přidávána vokalická znaménka a další značky.

4.3 Variabilita segmentace — příklad

Segmentace vyšších celků je v různých jazycích různá, což by nás nemělo překvapit. Například v anglických textech slova obsahují průměrně málo morfů a jejich hranice často s hranicemi morfů splývají, naproti tomu v turečtině mohou slova reprezentovat celé věty, u čínštiny pak o kategorii slov jako takové nejspíš není vhodné hovořit vůbec.

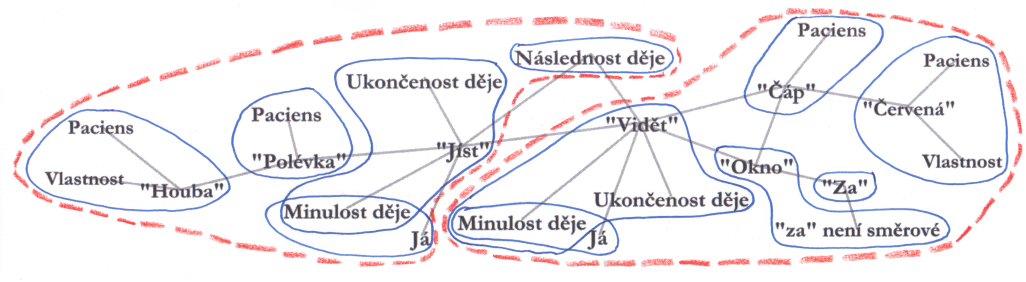
Podívejme se nyní na obrázek 4.1, který je odvozen z obrázku 3.1, do kterého jsme v grafu naznačili hranice slov (modře tenkou plnou čarou) a vět (červeně širší čárkovanou čarou). Na takto krátkém fragmentu textu jsou viditelné zejména hranice slov.

Srovnáme-li českou verzi věty o čápovi s jejím německým (obr. 4.2) a arabským (obr. 4.3) ekvivalentem, zjistíme, že společná sémantika se projevila v tom, že struktura grafu je velmi podobná, liší se pouze v přítomnosti symbolů pro některé vlastnosti, například v češtině nejsou vyžadovány symboly pro určenost, v němčině se zase neobjevují žádné speciální markery toho, jestli je daný symbol vlastností, nebo ne; dále jsou různé symboly různou měrou redundovány, což si ještě rozebereme v kapitole 6.3.[6]

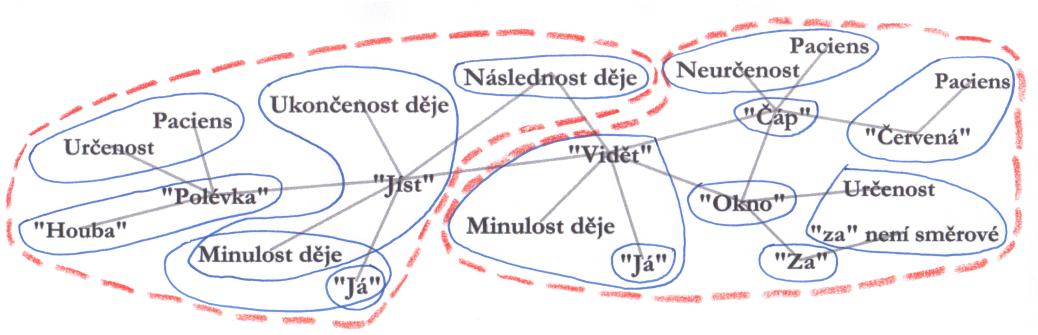
Naproti tomu rozdělení do slov a vět se i v této ne příliš složité větě radikálně liší,

[5] K tématu doporučuji Saenger (1997), přestože celkové vyznění díla je podle mě třeba brát s rezervou.

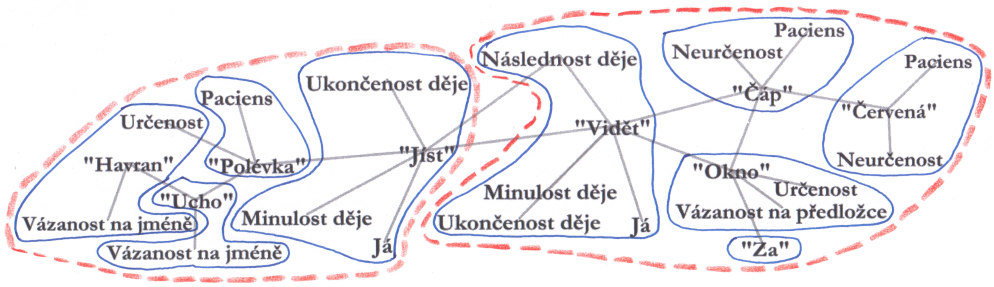
[6] Ještě jednou připomínám, že tato síť si neklade nároky na nějakou „pravdivou“ nebo ucelenou (neřku-li jedinou možnou) analýzu věty, ať už si pod tím představíme cokoli. Rozhodně nechci zavádět nějaký nový formalismus pro popis toho, jak spolu významy interagují. Je to jen ad hoc zobrazení a předpokládám, že by vás napadla spousta jiných a lepších zobrazení, například si vůbec nejsem jistý, jestli (v němčině) zahrnutí symbolu pro „já“ do více slov je správné řešení, jestli neměl být spíše opakován, neboť se objevuje fyzicky jako více morfémů, zvlášť když marker pro paciens opakuji. Důvod,



Obrázek 4.1: „Když jsem dojedl houbovou polévku, uviděl jsem za oknem červeného čápa.“



Obrázek 4.2: „Als ich die Pilzsuppe aufgegessen habe, sah ich hinter dem Fenster einen roten Storch.“



Obrázek 4.3

ʾAkaltu šūrbata ʿušši ʾl-ġurāb fa-raʾaytu laqlaqan ʾaḥmaran warāʾa ʾl-nāfiḏati.

أَكَلتُ شُوربَةَ عُشِّ الغُرَاب فَرَأَيتُ لَقلَقًا أَحمَرًا وَرَاءَ النَافِذَةِ.

vlastně jediné, co zůstává stejné, je, že oba symboly pro děj (které formují slovesa) jsou odděleny v různých větách. Kdybychom tímto způsobem rozebrali více konstrukcí (pokud možno ve více jazycích), zjistili bychom, že se jazyky neliší jenom v míře, v jaké seskupují morfémy do vyšších segmentů (tedy tradiční skaličkovská škála syntetičnost-analytičnost), ale že se fundamentálně liší v tom, jaké druhy symbolů se spolu seskupují do stejných segmentů. Například v češtině je morfém, který desambiguuje význam předložky „za“, zahrnut do stejného slova s morfémem, ke kterému se předložka váže („za okn-*em*“), zatímco v němčině je morfém se stejným významem zahrnut do stejného slova s morfémem, který má za úkol označit určenost celé konstrukce („hinter d-*em* Fenster“). Je zajímavé, že ani v jednom případě není zahrnut do jednoho slova s předložkou, s jejímž významem jeho význam interaguje.

Tedy: vágní je pojem nejnižší jednotky; nejasné je, jestli má smysl předpokládat jednotku nejvyšší; a variabilní a neostré je i samotné členění na segmenty v hladinách mezi nimi a počet hladin. Tato variabilita a vágnost je ovšem v jazyce obvyklá a spíše by nás překvapil opak.

4.4 Důvody segmentace

Abychom shrnuli dosavadní text: vyšší vnořená segmentace není logicky nezbytná, přesto se s větší či menší variabilitou vyskytuje jak v povědomí mluvčích, tak v lingvistických popisech jazyků. Konečně se nyní dostáváme k otázce, proč tomu tak je.

Používá se segmentace kvůli rozpoznání vztahů mezi symboly? Když se podíváme na segmentaci na obrázcích 4.1–4.3, zdá se, že skutečně sdružuje symboly, které k sobě patří, které spolu souvisí a které vzájemně významově interagují. Ale není to tak vždy, třeba pád zpřesňující význam předložky je v češtině i v němčině spojen nikoli s touto předložkou, ale v případě češtiny se slovem, s jehož významem předložka interaguje, v případě němčiny pak se slovem, které má za úkol informovat o určenosti slova, se kterým daná předložka interaguje. V případě arabštiny není marker určenosti spojen do jednoho slova se segmentem, který má být určen, ale se segmentem, který s tímto segmentem interaguje. Je možné, že segmentace do slov může rozpoznávání vztahů napomáhat, ovšem pokud tomu tak je, je ono napomáhání systematicky nedůsledné.

Spojujeme do vyšších celků shluky symbolů, které dostávají ustálený význam? Slova přece mají ustálený význam, často odlišný od toho, co bychom čekali z pouhé kombinace morfémů, ze kterých se skládají. Taktéž věty mají často ustálený význam. Jenže některá slovní spojení jsou taky ustálená a hranice idiomů rozhodně nekopírují hranice slov, vět nebo souvětí.

proč jsem zavedl nové a nezvyklé zobrazení vztahů, byl ten, že jsem nechtěl použít nějaký již existující (třeba syntaktický) formalismus, který by s sebou nesl nějaké teoretické a historické implikace a omezení. Grafy mají spíše ilustrovat základní myšlenku, než že by měly ambici ji nějak dokazovat.

A není vyšší segmentace užitečná kvůli deixi? Deixe nemusí být jenom na úrovni slov. Můžeme poukazovat na celý odstavec, nebo dokonce kapitolu (příklad pro koreferenci na větu nebo odstavec: „tato myšlenka byla oblíbená v anarchistických kruzích…“ ; koreference na celou kapitolu: „…což bylo hlavním tématem předchozí kapitoly…“).

Hranice segmentů mohou mít i technické vysvětlení – třeba kniha má svou typickou velikost, protože by se jinak nedala dobře držet v ruce. Některých hranic segmentů se používá třeba k nadechnutí se (slova, klauze). Obecně lze říci, že čím vyšší je segment, jehož hranici používáme k hezitační pauze, tím je hezitační pauza přijatelnější.

V následující kapitole se vás pokusím přesvědčit, že všechny tyto aspekty sice mohly podpořit vznik vnořeného segmentování v jazyce, nicméně že nejdůležitějším důvodem pro jeho emergenci byla nutnost vkládání redundance (jak jsme si ji popsali v kapitole 1.3) na několika úrovních, kteréžto vnořené vkládání je efektivním způsobem jak překonat různé typy zašuměných kanálů. A že díky tomu může jazyk být tím, čím je: kódem vhodným pro mezilidskou komunikaci za (skoro) všech podmínek.

Kapitola 5

Modely přenášení informace jazykem

V předchozích dvou kapitolách jsme si ukázali, že segmentace textu na hlásky, slova a věty není samozřejmostí, naopak že je jevem, jenž si zasluhuje explanaci. Nuže nyní se pokusím jedno vysvětlení nabídnout.

Budu se snažit, aby vysvětlení bylo srozumitelné i těm z vás, kteří se s teorií informace setkáváte poprvé, pročež může tato kapitola (spolu s kapitolou 1) posloužit jako stravitelný úvod do teorie informace pro lingvisty. Nicméně čtenářům s hlubším zájmem o toto užitečné téma doporučuji nějakou dedikovanou publikaci, například úvod, který napsal MacKay (2005), a jeho přednášky.[1]

Nejprve prozkoumáme prostředí, v jakém mezi sebou lidé obvykle komunikují, a namodelujeme zašuměné kanály, které musí a musel jazyk obvykle překonávat. Následně si ukážeme, jak si s těmito kanály poradí různé metody přidávání redundance, a nakonec se pokusíme určit obecný model strategie, kterou jazyky pro překonávání zašuměného kanálu skutečně používají.

5.1 Modely zašuměných kanálů

V této kapitole si probereme klasické modely zašuměných kanálů a zhodnotíme, se kterými z nich se musí potýkat přirozený jazyk, následně navrhneme modely nové, které budou lépe odrážet podmínky, v jakých se mezilidská komunikace obvykle odehrává.

Vzhledem k tomu, že jazykovou komunikaci pojímáme diskrétně jako sled realizací distinktivních rysů (o čemž bylo pojednáno v kapitole 3 a ještě bude detailněji na stranách 84 a 97), budeme se zde zabývat pouze digitálními (diskrétními) modely.

[1] Přístupné na adrese `http://www.inference.phy.cam.ac.uk/itprnn/Videos.shtml`.

5.1.1 Zašuměný kanál bez paměti

Nejprve si ukážeme jednoduché zašuměné kanály, které nemají paměť (*memoryless noisy channels*), tedy jejich chování nezávisí na předchozích datech a na předchozím chování kanálu — do kanálu vstupují jednotlivé bity a u každého zvlášť se kanál rozhodne, jakou operaci na nich provede, přičemž ono rozhodnutí je nezávislé na předchozích rozhodnutích. Asi nejznámější model, se kterým pracoval už Shannon a na kterém jsme si ukazovali základní pojmy v kapitole 1.3, je kanál, do kterého přitékají nuly a jedničky a každá jednička může projít nepoškozena, nebo se s určitou pravděpodobností změní na nulu, a naopak každá nula se s určitou pravděpodobností změní na jedničku. Pokud jsou tyto dvě pravděpodobnosti stejné, mluvíme o symetrickém kanálu (kvůli jednoduchosti je model vhodný pro teoretické úvahy), pokud ne, tak o kanálu asymetrickém, ten je schematicky představen na obrázku 5.1. Příklad šumu, který tento model generuje, naleznete na obrázku 5.4.[2]

Tento model reprezentuje situace, kdy jednotlivé distinktivní rysy slyšíme v převrácené variantě, například místo krátkého *i* slyšíme *í* dlouhé, místo emfatického arabského *ṣ* neemfatické *s* (tradiční to problém indoevropských arabistů). V mezilidské komunikaci není ovšem předem určeno, jaký čas nebo prostor zabírá jeden bit, takže distinktivní rys může nejen změnit svou hodnotu, ale může zcela vymizet.[3] Otázka je, jestli ztrátu třeba takové nazality chápat jako změnu bitu z jedničky na nulu, nebo úplné odstranění bitu. To samozřejmě platí i na vyšší úrovni a zakládá kontroverzi ohledně postulování nulových znaků a podobně. V každém případě tento model kanálu je v literatuře popsán[4] (tzv. *deletion channel*, my si zavedeme pojem *vymazávací kanál*) a jeho schéma najdete na obrázku 5.2.

Kromě úplné ztráty bez náhrady se může stát, že recipient ví, že něco chybí, například když vidí pohybující se rty, ale kvůli hluku neslyší, nebo když najde v sešitě zbytky po vytržených stranách. Tuto situaci modeluje tzv. *erasure channel* (jemuž budeme říkat česky *přemazávací kanál*), který s určitou pravděpodobností přenášený bit smaže, přičemž příjemce se dozví, že bit byl smazán (viz obrázek 5.3).

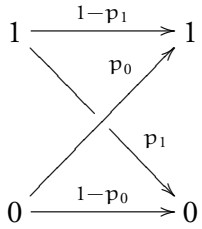
5.1.2 Zašuměný kanál s pamětí

V reálném světě se často vyskytují situace, kdy šum neovlivní právě jeden bit předávané zprávy, ale více bitů najednou — hladina šumu je kolísavá, rušivé vlivy mají tendenci

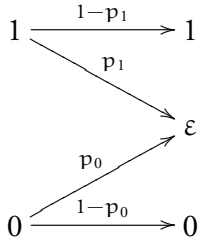
[2] Tato a všechny následující ilustrace zašuměných kanálů vznikly jako produkt vlastní implementace těchto modelů. Pokud máte zájem s nimi dále experimentovat, jsou volně dostupné na adrese `http://milicka.cz/disertace.zip`. V tomto archivu rovněž naleznete i implementace kódování pomocí multidimenzionální redundance, jakož i rozhraní pro jejich interakci. Zdrojový kód je záměrně napsán tak, aby bylo snadné modely kanálů i kódování volně modifikovat a rozvíjet.

[3] To platí zejména v mluvené řeči; v hebrejském kvadrátním písmu nebo ve strojové (*monospace*) latince máme představu o zamýšleném množství posílané informace přesnější.

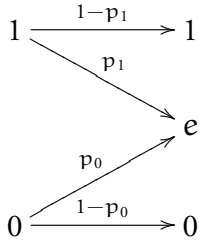
[4] Asi nejsrozumitelněji o soudobém vývoji pojednává Mitzenmacher (2009).



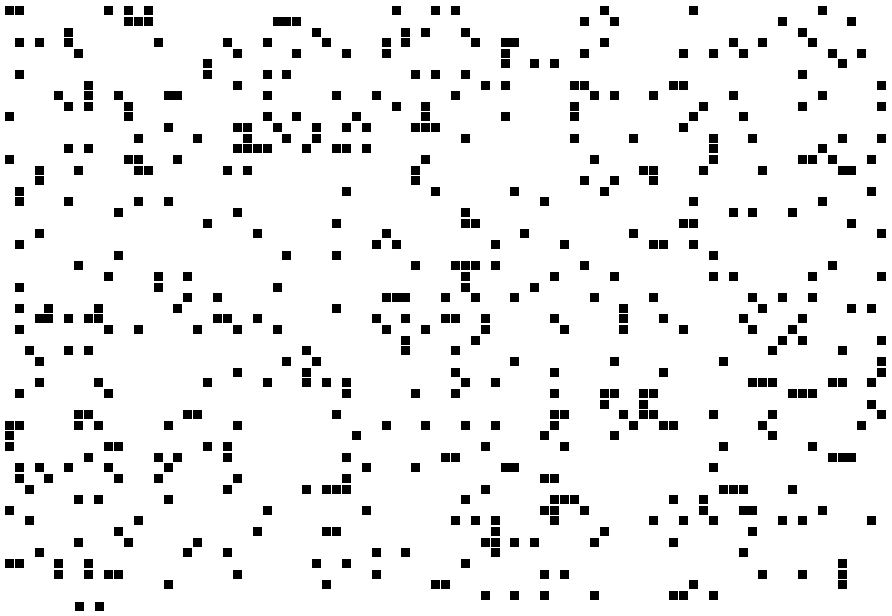
Obrázek 5.1: *Shannonovský asymetrický kanál:* na vstupu i na výstupu tohoto kanálu je sekvence skládající se ze znaků $\{1, 0\}$, p_1 je pravděpodobnost, že jednička se změnila na nulu, a p_0 je pravděpodobnost, že nula se změnila na jedničku.



Obrázek 5.2: *Vymazávací kanál (deletion channel):* na vstupu i na výstupu tohoto kanálu je sekvence skládající se ze znaků $\{1, 0\}$, přičemž ε značí, že znak byl vymazán bez náhrady, a p_1 a p_0 jsou pravděpodobnosti, že jednička, respektive nula jsou takto vymazány.



Obrázek 5.3: *Přemazávací kanál (erasure channel):* na vstupu tohoto kanálu je sekvence skládající se ze znaků $\{1, 0\}$ a na výstupu sekvence skládající se ze znaků $\{1, 0, e\}$, přičemž e je znak pro vymazanou hodnotu a p_1 a p_0 jsou pravděpodobnosti, že jednička, respektive nula jsou na tuto hodnotu změněny.



Obrázek 5.4: Příklad shannonovského symetrického šumu (generováno po řádcích); $p_0 = p_1 = 0{,}05$.

být pospolu. Mluvíme o „šumu s pamětí“, neboť stav kanálu je ovlivněn tím, jaký byl stav kanálu v minulosti.

V praxi si můžeme takový šum s pamětí představit jako poryvy větru nebo zvuk příboje, sousto v ústech mluvčího, řev raněného hrocha, řinčení nádobí nebo zbroje, paralelní mluvu někoho, koho neposloucháme, a podobně.[5]

Doplňovací cvičení v učebnicích cizích jazyků budiž důkazem, že pro recipienta není problém původní zprávu po průchodu takovým kanálem zrekonstruovat.

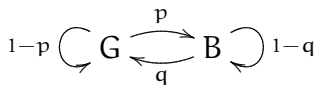
Když jsem uvažoval nad tím, jaký model by nejlépe popisoval šum, kterým musí zpráva kódovaná přirozeným jazykem projít, přemýšlel jsem nejprve nad nějakým druhem fraktálního šumu, například odvozeného od náhodně konstruované Cantorovy množiny. V analogových modelech zašuměných kanálů se fraktálních množin zhusta užívá (například růžový šum) a koneckonců fraktální šum pozorovaný samotným Mandelbrotem stál u zrodu zkoumání fraktálů.

V literatuře se však operuje s mnohem jednodušším modelem, tzv. *burst error noise* (který my budeme nazývat *dávkovým šumem*) (Gilbert, 1960), který je jednoduchý a přitom realistický. Model spočívá ve střídání dvou stavů, tradičně označovaných jako

[5] Kanál může být ovlivněn nejen svým vlastním stavem, ale i daty, která jím procházejí. Tak například může filtrovat výběrově různé distinktivní rysy (například rýma vytváří zašuměný kanál, který mění nazalitu, při šeptání se ztrácí znělost). Kanály s pamětí na data, která jím prošla, však v této práci modelovat nebudeme.

G a B. Ve stavu G je pravděpodobnost, že kanál převrátí hodnotu přenášeného bitu, g a ve stavu B je tato pravděpodobnost b. Zároveň pravděpodobnost, že se kanál dostane ze stavu G do stavu B, značíme p, zatímco pravděpodobnost, že se naopak dostane ze stavu B do stavu G, značíme q. Vlastně se jedná o střídání dvou symetrických shannonovských bezpaměťových kanálů.

Na obrázku 5.5 zobrazíme toto střídání stavů jako markovovský diagram.



Obrázek 5.5: Schéma generování Gilbertova dávkového šumu: pravděpodobnost přechodu mezi stavy $G \to B = p$, pravděpodobnost obráceného přechodu je $B \to G = q$.

Obrázek 5.6: Příklad Gilbertova dávkového šumu; $p = \frac{1}{72}$; $q = \frac{1}{8}$; $g = 0$; $b = \frac{1}{2}$. Stejně jako na obrázku 5.4 je celkový podíl chyb 5 %.

Důležitá charakteristika, která nám může pomoci určit adekvátnost tohoto modelu, je distribuce délek dávek. Hustotu distribuce délek dávek udává následující vzorec, ve kterém $f(n)$ značí relativní frekvenci dávek o délce n (odvození naleznete v příloze A.1):

$$f(n) = q(1-q)^{n-1} \tag{5.1}$$

Z čehož vyplývá, že pokud náhodně vybereme velké množství vzorků daného šumu a změříme pro každý vzorek průměrnou délku dávek, tak hustota distribuce průměrných délek bude dána následujícím vzorcem:

$$f(n) = qnq(1-q)^{n-1} \tag{5.2}$$

Průměrný podíl přetočených bitů ve zprávě pak udává vzorec 5.3:[6]

$$P(1) = \frac{\frac{g}{p} + \frac{b}{q}}{\frac{1}{p} + \frac{1}{q}} \tag{5.3}$$

V následující kapitole se podíváme na empirická data, abychom zjistili, zda je tento model realistický.

5.1.3 Realistický model kanálu pro mluvenou řeč

Inspirací k této kapitole byla asi čtyřicetiminutová nahrávka mluvené řeči, kterou mi laskavě poskytla Karolína Vyskočilová.[7] Nahrávka byla pořízena v chalupě v jedné české vesnici v Banátu, což simuluje prostředí, ve kterém po staletí žili naši předkové, neboť obsahuje zejména rušivé vlivy zvuků hospodářského stavení na venkově před industrializací: hlas hospodářských zvířat volně se pohybujících venku i vevnitř místnosti, jakož i zvířat nedomestifikovaných, kroky a dupot, praskání rozvrzaného nábytku a nenamazaných dveří, ťukot nádobí, šplouchání vody, překrývající se řeč a různé pracovní zvuky.

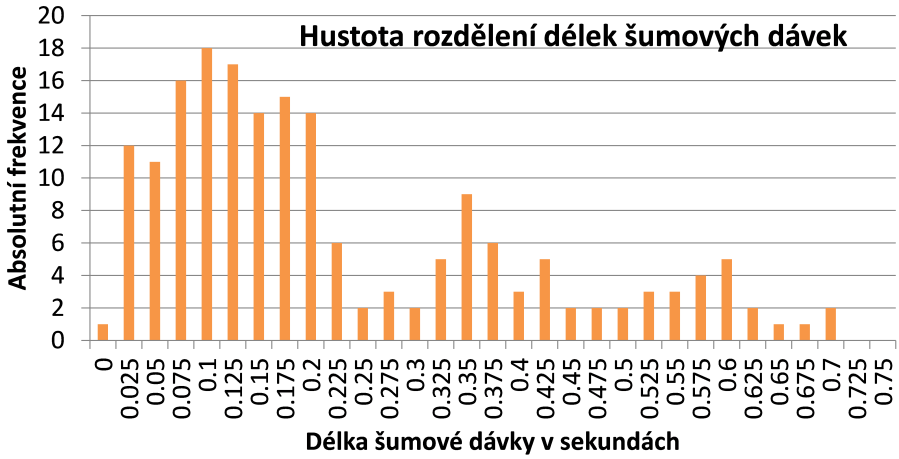
Pokud se na nahrávku podíváme podrobněji, zjistíme:

1. že rušivé zvuky se vyskytují v dávkách,
2. že hlasitost rušivého zvuku v dávce není rovnoměrná, přičemž předpokládám, že pravděpodobnost chyby přenosu na této hlasitosti závisí,
3. že distribuce délek těchto zvuků neodpovídá tomu, co předpovídá zmiňovaný Gilbertův model,
4. a že do konverzace zasahuje jen několik málo specifických a rozlišitelných zašuměných kanálů, které se vzájemně překrývají, Tyto kanály je možné separovat a modelovat přesněji než obecnou Gilbertovou aproximací.

[6] Tento vzorec je obecnější, než jak ho udává Gilbert (1960, str. 1255, vzorec 1), který předpokládá, že parametr $g = 0$, nicméně byl konstruován stejným způsobem. Také vzorce 5.7, 5.8 a 5.9 jsou založeny na stejném principu.

[7] S ní jsem o tomto fenoménu napsal podrobnější článek, který nejspíš vyjde pod jménem *Models of the Noisy Channels that Speech Gets over*, zatím neznámo kdy a kde. Karolína Vyskočilová nejenže dodala nahrávku (o vzniku banátského korpusu podrobněji ve Vyskočilová, 2014), ale také prováděla analýzu nahrávky (segmentaci na zašuměné a bezšumové části), a to pomocí programu Praat (Boersma – Weenink, 2015).

Než se na tyto jednotlivé zvuky a jejich modely podíváme podrobněji, nejprve si ukážeme, jak reálným datům odpovídá Gilbertův model. Distribuci délek jednotlivých šumových dávek ve vzorku naleznete na obrázku 5.7.



Obrázek 5.7: Hustota distribuce délek dávek šumu ve vzorku.

Už na první pohled je asi jasné, že tento graf predikované geometrické distribuci neodpovídá, což si názorně ukážeme na obrázku 5.8. K fitování modelů používám program Eureqa od firmy Nutonian (Schmidt – Lipson, 2009). O mnoho si nepolepšíme, když se pokusíme nafitovat několik vzájemně se překrývajících Gilbertových modelů, což odpovídá několika vzájemně heterogenním zdrojům šumu. Asi nás nepřekvapí, že když nafitujeme 10 vzájemně se překrývajících Gilbertových modelů, dostaneme se k R^2 jen o málo vyššímu ($R^2 = 0{,}51$). Takový model pak bude mít hustotu distribuce charakterizovanou vzorcem:[8]

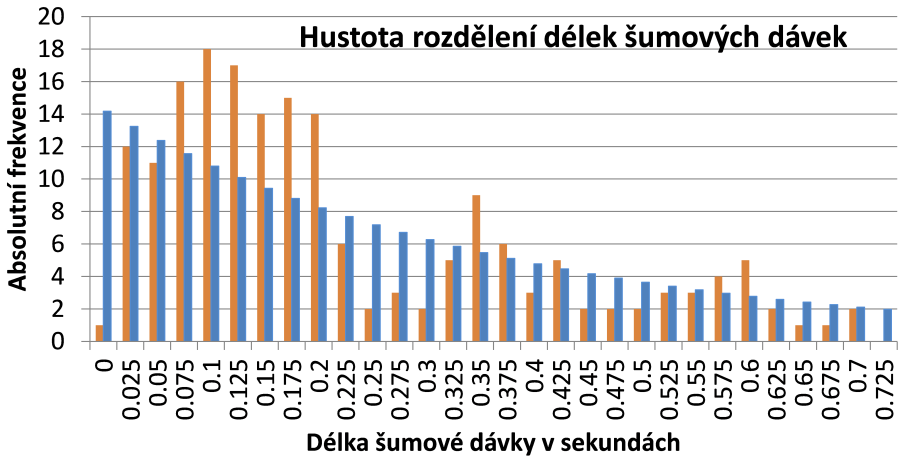
$$f(n) = \sum_{i=1}^{10} q_i (1 - q_i)^{n-1} \tag{5.4}$$

Zarážející je, že nejen na distribuci délek šumových dávek, ale ani na distribuci tichých mezer mezi těmito dávkami (obrázek 5.9) se Gilbertův model nehodí (obrázek 5.10).

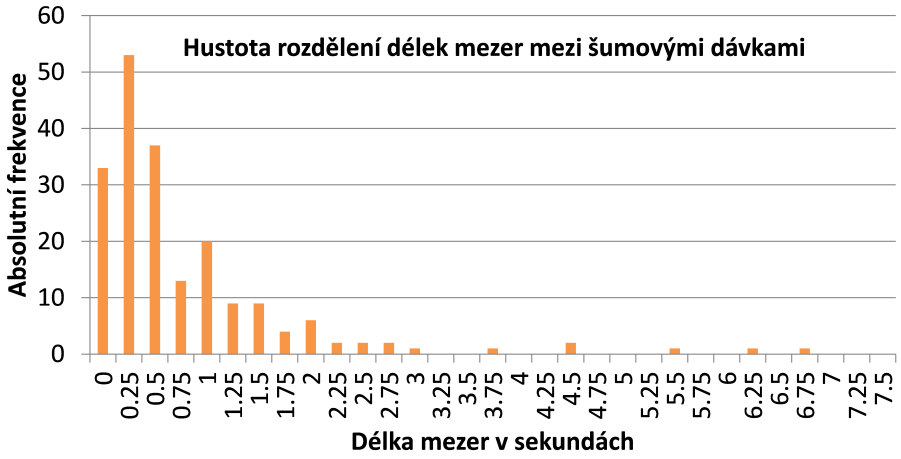
V případě mezer mezi dávkami se však situace trochu zlepší, když postulujeme 10 vzájemně se překrývajících Gilbertových modelů, v tomto případě bude možné nafitovat model s $R^2 = 0{,}7$.

Nicméně zřejmá nevhodnost tohoto modelu nás vede k tomu, abychom se poohlédli po něčem realističtějším.

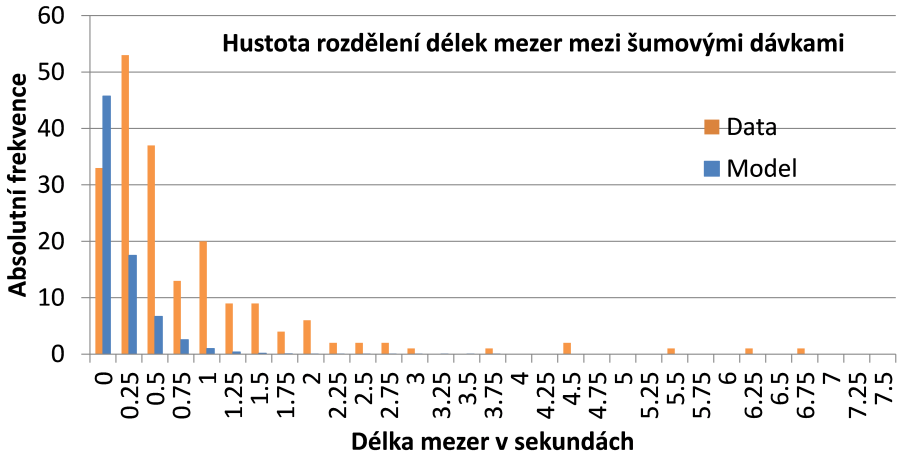
[8] Vychází z vzorce 5.1, kde q_1 až q_{10} jsou parametry jednotlivých překrývajících se Gilbertových modelů.



Obrázek 5.8: Hustota distribuce délek dávek šumu ve vzorku a Gilbertův model nafitovaný s parametrem $q = 0{,}934$; $R^2 = 0{,}475$



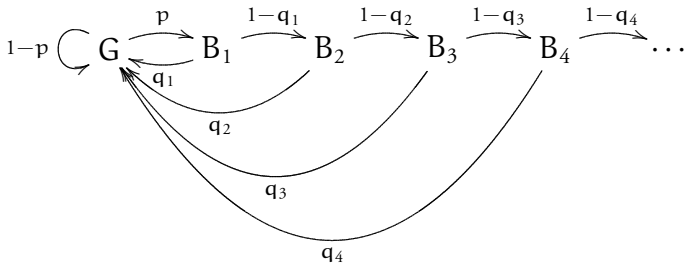
Obrázek 5.9: Hustota distribuce délek mezer mezi šumovými dávkami.



Obrázek 5.10: Hustota distribuce délek mezer mezi šumovými dávkami a na ni nafitovaný Gilberův model s parametrem $p = 0{,}979$, $R^2 = 0{,}36$.

Model pro nepravidelně se opakující zvuky stabilní délky (dále model M1).

Tento model si klade za cíl popsat zvuky, které se vyskytují nepravidelně, jež ale mají nějakou typickou délku, kolem níž variují. Takové rozdělení délek můžeme aproximovat normálním rozdělením. Takovými zvuky je například kravské bučení a vůbec různé zvířecí zvuky. Tento model, podobně jako Gilbertův, zahrnuje stav G (tzn. *good*), který má nízkou pravděpodobnost vzniku chyby g (v nejjednodušším případě je tato pravděpodobnost nulová). S pravděpodobností p se pak model dostane do stavu B_1, z něhož se model může následně dostat do stavu B_2 a tak dále (obecně B_i), přičemž pravděpodobnost vzniku chyby ve stavech B_i si označíme b_i (pro opravdu rovnoměrně hlasité zvuky můžeme aproximovat b_i jako konstantu, tedy $b_i = b$). Podobně jako v případě Gilbertova modelu si můžeme ukázat názorné schéma, které ho reprezentuje (obrázek 5.11).



Obrázek 5.11: Schéma generování šumu typu M1: pravděpodobnost přechodu mezi stavy $G \to B_1 = p$, pravděpodobnost obráceného přechodu je $B_i \to G = q_i$.

Co se týče funkce, která přisuzuje pravděpodobnost q_i, záleží na tom, jaké distribuce délek dávek chceme docílit. Pokud je q_i konstantní, distribuce bude stejná jako v Gilbertově modelu, neboť tento model je zobecněním Gilbertova modelu (kterýžto z tohoto modelu dostaneme dosazením $q_i = \mathrm{const.}$ a $b_i = \mathrm{const.}$).

Pokud bychom chtěli aproximovat rozdělení délek dávek normálním rozdělením, nabývaly by pravděpodobnosti změnu ze stavu B_i do stavu B (které značíme jako q_i) následujících hodnot (μ značí střední hodnotu; σ směrodatnou odchylku; odvození této rekurentní rovnice naleznete v příloze A.2):

$$q_1 = f_1 = \frac{e^{-\frac{(1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{5.5}$$

$$q_i = \frac{q_{i-1}\, e^{\frac{\mu - i + \frac{1}{2}}{\sigma^2}}}{1 - q_{i-1}} \tag{5.6}$$

Pokud vzorek obsahuje jiné rozdělení délek dávek než normální, je možno postupovat při vytváření vzorce pro q_i stejně jako v příloze A.2.

Průměrný podíl přetočených bitů ve zprávě pak pro normální rozdělení udává vzorec 5.7:

$$P(1) = \frac{\frac{g}{p} + b\mu}{\frac{1}{p} + \mu} \tag{5.7}$$

Podobně jako pro Shannonův a Gilbertův model, i pro tento si ukážeme typický šum, najdete ho na obrázku 5.12.

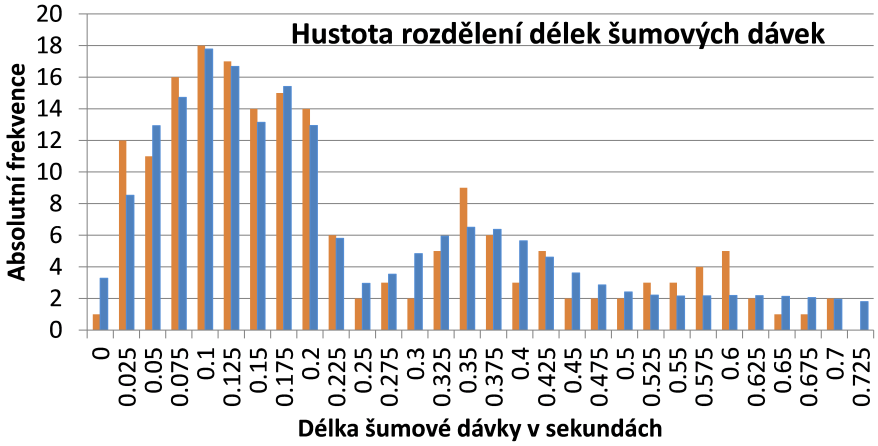
Nyní si zkusíme nafitovat model, který předpokládá pět vzájemně se překrývajících kanálů s normálně rozděleným dávkovým šumem. Výsledek vidíte na obrázku 5.13.

Tento model řeší distribuci délek šumu, zatímco distribuce délek mezer je popisována stochastickým procesem stejným jako v případě Gilbertova modelu. To se pokusíme změnit na následujících řádcích.

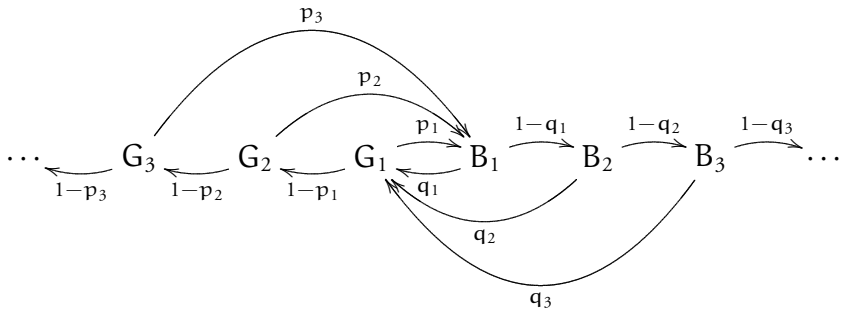
Model pro pravidelně se opakující zvuky stabilní délky (dále M2).

Tento model zachycuje zvuky, které nejenže trvají nějakou typickou dobu, ale také jsou mezi nimi nějak typicky dlouhé rozestupy, samozřejmě s variacemi. Pod takovými zvuky si můžeme představit hlas některých zvířat, například sovy, která houká ve vcelku pravidelných periodách, vytí vlka, zvuk veslování nebo otáčení rezavého rumpálu, porážení stromu sekyrou, řezání dřeva… Model je znázorněn na obrázku 5.14. Oproti předchozímu modelu přibyla série stavů G_i s pravděpodobností p_i, že model přejde do stavu B_1.

Obrázek 5.12: Příklad šumu, model pro nepravidelně se opakující zvuky stabilní délky (M1); $p = \frac{1}{72}$; $\mu = 8$; $\sigma = 2$; $g = 0$; $b = \frac{1}{2}$. Stejně jako na obrázku 5.4 je celkový podíl chyb 5 %.



Obrázek 5.13: Hustota distribuce délek mezer mezi šumovými dávkami a na ni nafitovaný model o pěti překrývajících se normálních rozděleních. $R^2 = 0{,}923$.



Obrázek 5.14: Schéma generování šumu typu M2: pravděpodobnost přechodu mezi stavy $G_i \rightarrow B_1 = p_i$, pravděpodobnost obráceného přechodu je $B_i \rightarrow G_1 = q_i$. Jak p_i, tak q_i můžeme přiřadit funkci 5.6, čímž docílíme toho, že intervaly setrvávání modelu v obou stavech budou normálně rozdělené (model M2).

Průměrný podíl přetočených bitů ve zprávě pak pro normální rozdělení udává vzorec 5.8:

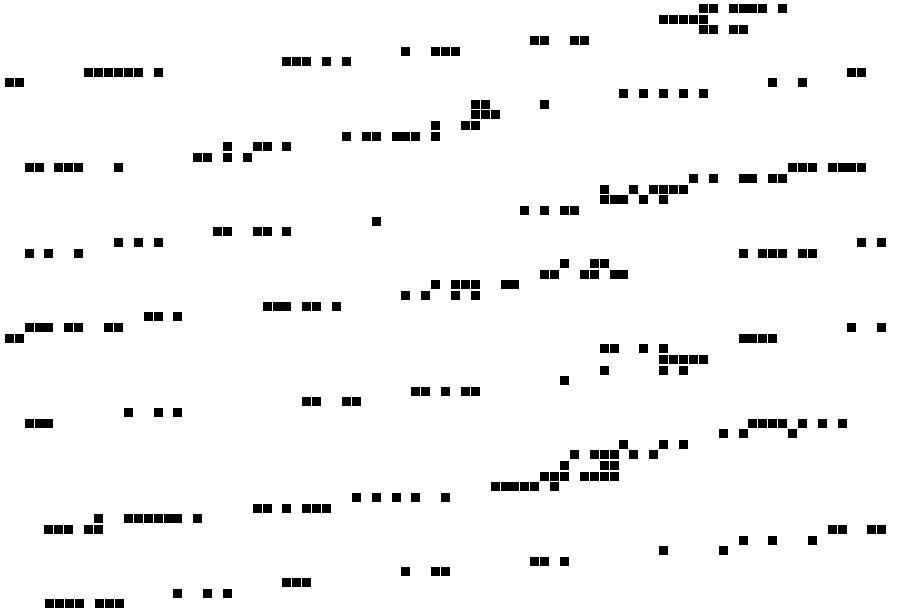
$$P(1) = \frac{g\mu_g + b\mu_b}{\mu_g + \mu_b} \tag{5.8}$$

Podobně jako pro Shannonův a Gilbertův model, i pro tento si ukážeme typický šum, najdete ho na obrázku 5.15.

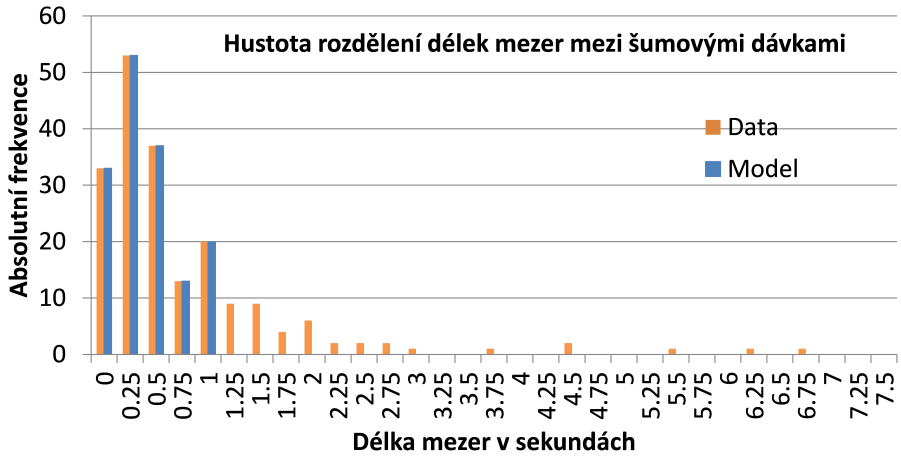
Opět nafitujeme model, který předpokládá pět vzájemně se překrývajících kanálů s normálně rozdělenými délkami mezer mezi dávkami šumu. Výsledek vidíte na obrázku 5.16. Samozřejmě můžeme první a druhý model kombinovat: na obrázku 5.17 naleznete výsledky pro překrývající se čtyři geometrické distribuce a tři normální.

Model pro nepravidelně se opakující série pravidelně se opakujících zvuků stabilní délky (vnořený model, dále M3).

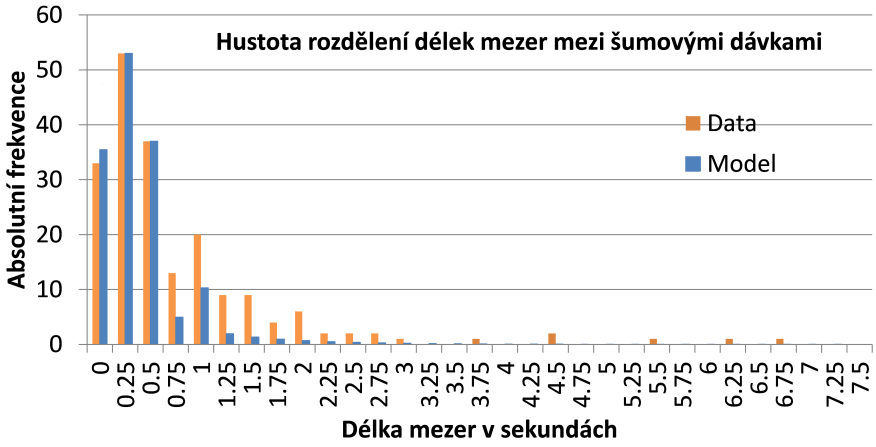
V reálném světě se často setkáváme s generátory šumu, které v nepravidelných intervalech přecházejí ze stavu delších období ticha do sérií pravidelně se opakujících krátkých dávek hluku a ticha. Tento model zachycuje série zvuků, které se vyskytují nepravidelně — takové jsou zvuky a hlasy některých zvířat, například datla nebo žáby, zvuky kroků při občasném přecházení po místnosti, naklepávání kosy a podobně. Tyto série se skládají z pravidelně se opakujících chvil ticha (stavy G_i) a zvuků, které také trvají nějakou typickou dobu, samozřejmě s variacemi (stav B_i). Po skončení série se model vrací do stavu H (stav mimo sérii). Model je znázorněn na obrázku 5.18.



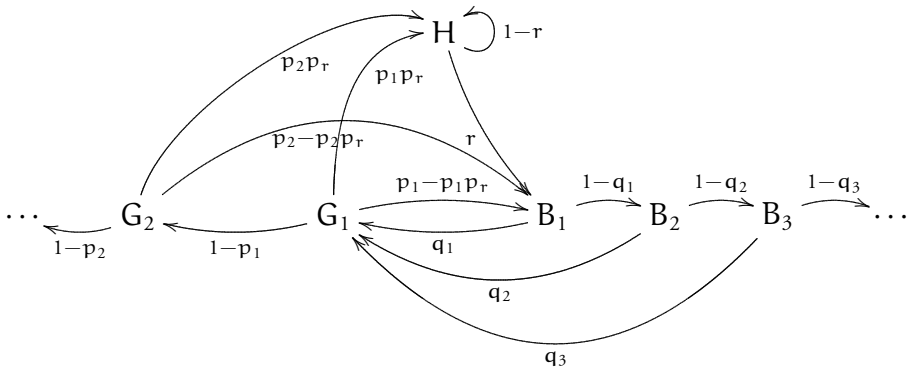
Obrázek 5.15: Příklad šumu, model pro pravidelně se opakující zvuky stabilní délky (M2) $\mu_g = 72$; $\sigma_g = 5$; $\mu_b = 8$; $\sigma_b = 2$; $g = 0$; $b = \frac{1}{2}$. Stejně jako na obrázku 5.4 je celkový podíl chyb 5 %.



Obrázek 5.16: Hustota distribuce délek mezer mezi šumovými dávkami a na ni nafitovaný model o pěti překrývajících se normálních rozděleních. $R^2 = 0{,}95$.



Obrázek 5.17: Hustota distribuce délek mezer mezi šumovými dávkami a na ni nafitovaný model o čtyřech překrývajících se geometrických rozděleních a třech normálních rozděleních. $R^2 = 0{,}93$.



Obrázek 5.18: Schéma generování šumu typu (M3): pravděpodobnost přechodu mezi stavy $G_i \to B_1 = p_i$, pravděpodobnost obráceného přechodu je $B_i \to G_i = q_i$. Jak p_i, tak q_i můžeme přiřadit funkci 5.6, čímž docílíme toho, že intervaly setrvávání modelu v obou stavech budou normálně rozdělené. Tento model disponuje kromě stavů G_i a B_i stavem H, který reprezentuje delší pauzu mezi jednotlivými sériemi, přičemž z tohoto stavu se model může dostat do stavu B_1 a zpět se může dostat ze stavů G_i. S pravděpodobností p_i model opustí stav G_i a následně se rozhodne, jestli se vrátí do stavu H (pravděpodobnost p_r), nebo jestli bude pokračovat v sérii $(1-p_r)$.

Průměrný podíl přetočených bitů ve zprávě udává následující vzorec 5.9:

$$P(1) = \frac{\frac{h}{r} + \frac{g\mu_g + b\mu_b}{p_r}}{\frac{1}{r} + \frac{\mu_g + \mu_b}{p_r}} \tag{5.9}$$

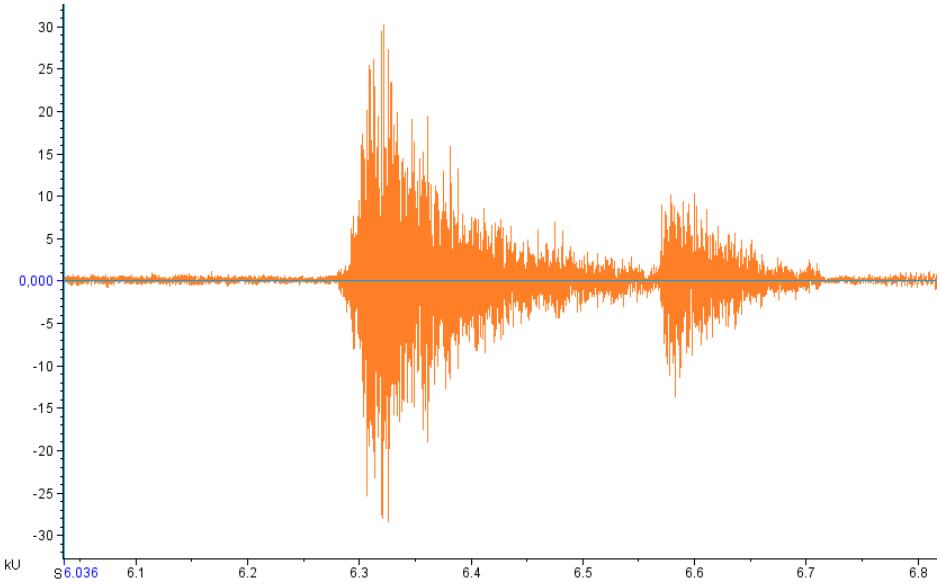
Podobně jako pro všechny ostatní modely, i pro tento si ukážeme typický šum, najdete ho na obrázku 5.19.

Obrázek 5.19: Příklad šumu, model pro nepravidelně se opakující pravidelné série zvuků (model M3): $r = \frac{1}{256}$; $p_r = \frac{1}{4}$; $\mu_g = 8$; $\sigma_g = 2$; $\mu_b = 8$; $\sigma_b = 2$; $h = 0$; $g = 0$; $b = \frac{1}{2}$. Stejně jako na obrázku 5.4 je celkový podíl chyb 5 %.

Model pro náhodné nárazy rezonujících předmětů (dále M4).

Dosud jsme počítali s tím, že pravděpodobnost změny bitů procházející zprávy je během dávky šumu stabilní. Takový přístup je ovšem nerealistický, většina šumů v nahrávce vypadá jako na obrázku 5.20: prvotní impuls nějakou dobu kmitá, přičemž je jeho amplituda tlumena.

Pokud bychom chtěli odvodit vzorec pro pravděpodobnost, s jakou model přetočí hodnotu bitu ve stavu B_i, museli bychom nejspíš provést percepční experimenty. Pro běžné lineární tlumení kmitání bychom naivně mohli očekávat vzorec 5.10 (kde b_1 a A jsou parametry).



Obrázek 5.20: Příklad reálného šumu nalezeného v nahrávce. Byl způsoben nejspíš položením nějakého předmětu na stůl.

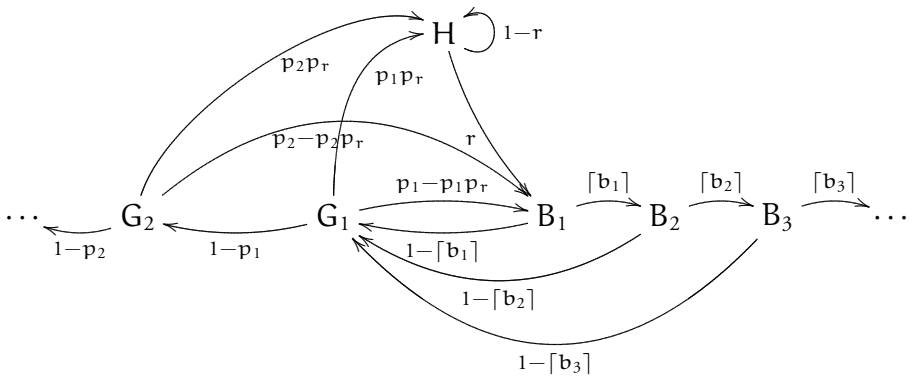
$$b_i = b_1 e^{-iA} \tag{5.10}$$

Ovšem celé lidské vnímání zvuků je relativní, a tak můžem předpokládat, že i chybovost způsobená hlukem bude v logaritmickém měřítku (vzorec 5.11 je vytvořen zlogaritmováním vzorce 5.10, což souvisí s vzorcem pro výpočet hladiny akustického tlaku, se kterým, předpokládejme, bude pravděpodobnost chyby lineárně korelovat):

$$b_i = \ln\left(c\, e^{iA}\right) = b_1 - iA \tag{5.11}$$

Pro model postavený na tomto vzorci je pravděpodobnost přechodu modelu do stavu G irelevantní: přechod B $\rightarrow$ G se prostě provede tehdy, když pravděpodobnost b_i dosáhne nuly nebo ji překročí. Jednoduchý algoritmus pro generování takového šumu zní: při přechodu G $\rightarrow$ B je podle nějakého klíče vygenerována pravděpodobnost b_1, která se následně po každém dalším kroku snižuje podle vzorce 5.11, přičemž když b_i spadne na/pod nulu, model přejde do stavu G. Můžeme si to opět znázornit na schématu (5.21).

Pravděpodobnost B_1 můžeme generovat buď náhodně z nějakého intervalu, nebo podle nějakého jiného realističtějšího rozdělení, třeba normálního. Vzhledem k nutnosti ověření a dalšího rozpracování je nutno tento model brát pouze jako návrh. Podobným způsobem by bylo možno se podívat i na další typické zašuměné kanály, u nichž rušivost je korelována s délkou trvání, například vlnobití, dočasné ohlušení po výstřelu apod.



Obrázek 5.21: Schéma generování šumu typu (M4): pravděpodobnost přechodu mezi stavy $G_i \to B_1 = p_i$, pravděpodobnost obráceného přechodu $B_i \to G_i$ je stoprocentní, pokud pravděpodobnost změny bitu v procházející zprávě ve stavu B_i (tuto pravděpodobnost jsme si označili jako b_i) je nulová a menší. Formálně toho v modelu dosahujeme tak, že od jednotkové pravděpodobnosti odečteme číslo $\lceil b_i \rceil$, tedy b_i zaokrouhlené vždy nahoru. Pravděpodobnost b_i se spočítá podle vzorce 5.11.

5.2 Koncept vkládání redundance na více úrovních

5.2.1 Paritní bit

V informatice bylo vyvinuto mnoho metod, jak výše uvedené zašuměné kanály překonat. V této podkapitole si ukážeme jednu jednoduchou strategii, která, jak se později ukáže, vede ke strukturám, které jsou podobné skutečnému textu. Začněme s jednoduchým paritním bitem, jak ho známe z kapitoly 1.3, pro připomenutí zde uvádím tabulku 5.1 (je totožná s tabulkou 1.3 A).

	Produktor		Recipient
Původní	1		1
zpráva:	0		0
	1	→	1
	1		1
	0		0
Redundance:	1		1

Tabulka 5.1: Přenesli jsme přes zašuměný kanál zprávu i paritní bit (označený rámečkem) a obé přišlo, shodou okolností, nepoškozeno. Paritní bit vypočítaný z doručené zprávy tedy odpovídá paritnímu bitu, který jsme společně se zprávou poslali.

Dejme tomu, že zprávu o pěti bitech a jeden paritní bit (obecně dohromady n

bitů) posíláme přes klasický shannonovský bezpaměťový symetrický kanál (jako na obrázku 5.1), kde $p_1 = p_0 = p = 0{,}05$. Pravděpodobnost, že zpráva dojde zcela v pořádku, je rovna $(1-p)^n = 0{,}95^6 \doteq 73{,}5\,\%$. Pravděpodobnost, že zpráva dojde s právě jednou chybou, a tedy že díky redundantnímu bitu recipient zjistí, že je nekonzistentní, je rovna $np(1-p)^{n-1} = 6 \times 0{,}05 \times 0{,}95^5 \doteq 19{,}3\,\%$. Analogicky pomocí binomického rozdělení můžeme vypočítat, že pravděpodobnost dvou chyb, které už náš paritní bit neodhalí, je rovna $6 \times 5 \times p^2(1-p)^{n-2} = 6 \times 5 \times 0{,}05^2 \times 0{,}95^4 \doteq 6,1\,\%$, tedy obecně pro k chyb:

$$b_{n,k} = \binom{n}{k} p^k (1-p)^{n-k} \tag{5.12}$$

Dosazením všech možností, kdy zmíněný kanál změní ve zprávě sudý počet bitů, dostaneme pravděpodobnost, že zpráva dojde špatně a bude neopravitelná (tabulka 5.2).

	Zpráva poškozena	Zpráva nepoškozena
Konzistentní	$b_{6,2} + b_{6,4} + b_{6,6} \doteq 3{,}0\,\%$	$b_{6,0} \doteq 73{,}5\,\%$
Nekonzistentní	$b_{6,1} + b_{6,3} + b_{6,5} - p(1-p)^5 \doteq 19{,}6\,\%$	$p(1-p)^5 \doteq 3{,}9\,\%$

Tabulka 5.2: Tabulka udává pravděpodobnosti, s jakými 5bitová zpráva redundovaná jedním paritním bitem, poté co prošla symetrickým bezpaměťovým zašuměným kanálem s pravděpodobností změny bitu $p = 0{,}05$, byla recipientem interpretována jako poškozená a jestli se tak stalo odůvodněně.

V případě, že zpráva projde přemazávacím kanálem (*erasure channel*, k němu viz obrázek 5.3), redundantní bit nám dovolí nejen určit, že je zpráva špatně, ale jeden bit můžeme dokonce dopočítat. Pravděpodobnost, že je poškozen právě jeden bit, je možno určit pomocí stejného vzorce jako v předchozím případě, a tak tabulka vypadá obdobně (5.3)

Zpráva dorazila zcela správně	$b_{5,0} \doteq 77{,}4\,\%$
Zprávu je možno dopočítat	$b_{5,1}(p-1) \doteq 19{,}3\,\%$
Zprávu není možno dopočítat	$b_{6,2} + b_{6,3} + b_{6,4} + b_{6,5} + b_{6,6} \doteq 3{,}3\,\%$

Tabulka 5.3: Tabulka udává pravděpodobnosti, s jakými bylo lze rekonstruovat 5bitovou zprávu redundovanou jedním paritním bitem, která prošla symetrickým bezpaměťovým *přemazávacím kanálem* s pravděpodobností přemazání bitu $p = \frac{5}{100}$.

Nyní si předveďme, jaká je pravděpodobnost, že obdobná a obdobně chráněná zpráva projde kanálem s dávkovým šumem (Gilbertův model). Jako příklad si vezměme stejný model dávkového šumu, jaký byl použit pro generování obrázku 5.6. Jeden

každý bit zprávy, která prochází tímto kanálem, má pravděpodobnost chyby 0,05, což odpovídá pravděpodobnosti p z předchozích odstavců. To samozřejmě není náhoda, zvolil jsem ji záměrně tak, abychom mohli porovnat účinnost paritního bitu na opravu chyb způsobených bezpaměťovým kanálem a paměťovým kanálem s obdobnou chybovostí.

Jaká je pravděpodobnost, že zpráva o pěti bitech dorazila zcela správně, a jaká, že dorazila s detekovatelným poškozením? Podobně jako v předchozím případě bychom mohli pravděpodobnosti odvodit algebraicky, pro složitější vztahy, jako je tento, je ovšem pohodlné využít metodu Monte Carlo — jednoduše implementujeme popsané modely šumu a metody vkládání redundance jako program a mnohokrát je použijeme (v našem případě 10^7krát).[9] Použití v tomto případě znamená zakódování zprávy, přidání šumu, dekódování a porovnání dekódované zprávy se zprávou původní.

Správnost	+	+	-	-
Konzistence	+	-	+	-
Shannonův	73,5	3,9	3,1	19,5
Gilbertův	86,5	1,1	5,4	6,9
M1	86,3	1,2	5,6	6,8
M2	86,2	1,3	5,6	6,9
M3	86,3	1,2	5,6	6,8

Tabulka 5.4: Tabulka udává pravděpodobnosti (v procentech), s jakými 5bitová zpráva redundovaná jedním paritním bitem, poté co prošla různými druhy zašuměných kanálů, byla poškozená (řádek Správnost) a jestli byla recipientem jako poškozená interpretována (řádek Konzistence, myšleno konzistence s paritním bitem). Znaménko plus značí pravdivostní hodnotu *pravda*, mínus *nepravda*, tedy například sloupec +- obsahuje frekvenci zpráv, které dorazily správně, ale nejsou konzistentní s paritním bitem (tzv. *false positives*). Zašuměné kanály jsou nastaveny tak, aby pravděpodobnost změny bitu byla $\mathrm{p} = 0{,}05$ (respektive mají parametry jako na obrázcích 5.6, 5.12, 5.15 a 5.19). Ve všech případech jsou 95procentní konfidenční intervaly nižší než desetina procenta.

Jak ilustruje tabulka 5.4, pro kanály s pamětí je tento způsob přidávání redundance neefektivní, neboť frekvence zpráv, které byly poškozeny, avšak poškození nebylo detekováno, se příliš neliší od frekvence zpráv, které byly poškozeny a poškození detekováno bylo, přičemž se dá předpokládat, že neodhalení přenosové chyby nese vysoké náklady.[10]

[9] Připomínám, že mé implementace všech zde použitých modelů kódování a modelů šumu najdete na adrese `http://milicka.cz/disertace.zip`, detaily v poznámce 2 na straně 49.

[10] Možná jste si všimli, že kanály s pamětí mají různou pravděpodobnost, že zpráva projde nepoškozena, přestože pravděpodobnost změny bitu je stále 5 %. V následujících kapitolách také uvidíme,

To platí téměř bezezbytku tehdy, když je délka zprávy rovna délce segmentu, za kterým posíláme redundantní bit (tedy pokud je každý n-tý bit redundantní a posíláme zprávu o n bitech), neboť je pravděpodobné, že pokud už je zpráva poškozená, pak je poškozená na více místech (dávkou) a pravděpodobnost, že byl poškozen sudý počet bitů, se blíží 0,5. To nám ukazuje graf na obrázku 5.22. Pro posílání krátkých zpráv je tedy jednoduché vkládání redundance v jednom rozměru neúčinné a pokud bychom u delších zpráv rozčleněných do více segmentů chtěli zjistit, který konkrétní segment je chybný, též nám mnoho nepomůže.

Z tohoto důvodu se s jednoduchým přidáváním paritního bitu na jedné úrovni nespokojíme.

5.2.2 Multidimenzionální redundance

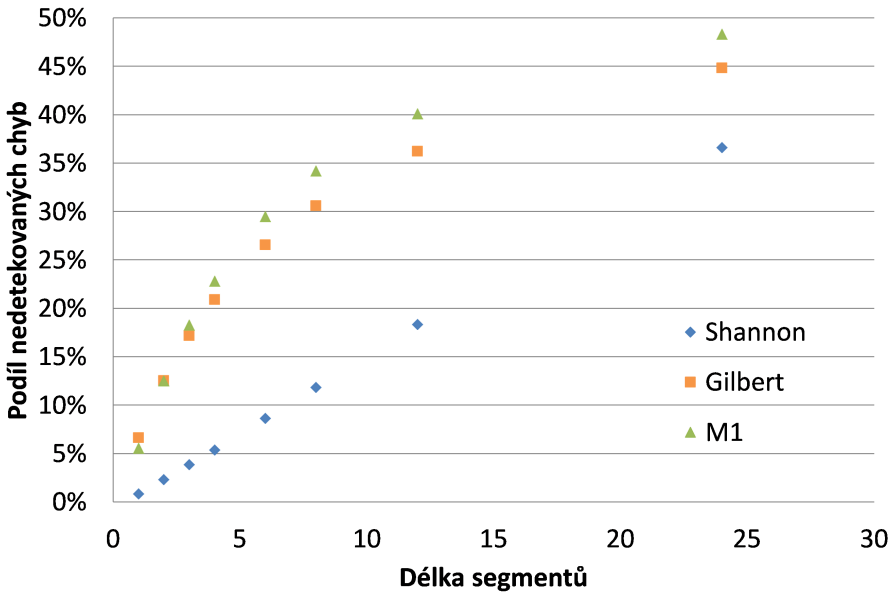
Pro delší zprávy můžeme přidávat paritní bit v pravidelných intervalech, jak ukazuje matice 5.5. Takto přidávané redundanci se říká horizontální (nebo též *longitudinal redundancy check*).

0	0	0	1	1	0	1	0	1
0	1	1	0	1	0	1	1	1
0	1	1	0	0	1	1	0	0
0	1	1	1	0	0	0	1	0
1	0	1	1	1	0	0	0	0
1	0	0	0	0	0	1	1	1
1	1	0	1	1	0	1	0	1
0	0	0	0	1	0	0	0	1

Tabulka 5.5: Matice znázorňující příklad použití horizontální redundance. Redundantní bit je určen pozicí (každý devátý bit je paritní vůči osmi předchozím), ale pro přehlednost je zobrazen v rámečku.

Jednoduché přidávání jednoho paritního bitu na zprávu nebo její úsek se ovšem špatně vyrovnává s dávkovým šumem (jak jsme si ukázali v předchozí podkapitole)

že zprávy po průchodu různými kanály je možné s různou úspěšností rekonstruovat. Teoreticky jsou kanály s pamětí lépe předvídatelné, respektive šum, který produkují, má nižší komplexitu — ve smyslu kapitoly 1.4 (shannonovské paradigma mluví o větší kapacitě kanálu). Když si kolmogorovovskou komplexitu popsaných šumů (tedy těch, které jsme použili pro generování tabulky 5.4) zkusíme aproximovat prostou kompresí Zipem (defaultní Zip ve Windows 7 64 b), pak shannonovský bezpaměťový kanál má komplexitu maximálně 0,55 b na 1 bit zprávy; Gilbertův dávkový šum 0,35 b; šum kanálu, který jsme si označili jako M1, má 0,35 b; M2 má 0,32 b a M3 má 0,30 b (platí pro parametry ve všech případech stejné jako na obrázcích 5.6, 5.12, 5.15 a 5.19). Přestože tento způsob měření má k ideálu velmi daleko, snižující se komplexita je přesně to, co bychom mohli očekávat.



Obrázek 5.22: Podíl chybných zpráv, které nebyly detekovány jako chybné. Ve všech případech má zpráva 24 b a byla rozdělena na segmenty zakončené paritním bitem. Jednotlivé značky udávají, jakou neúspěšnost jednotlivá řešení mají s ohledem na šum, kterým zpráva prošla. Jak vidíme, pokud je zpráva rozdělena na jeden segment o dvaceti čtyřech bitech, v případě Gilbertova a M1 modelu se pravděpodobnost selhání blíží náhodnému výběru a přidávání redundance tedy postrádá smysl. Zašuměné kanály mají stejné parametry jako ty, které byly použity k vytvoření tabulky 5.4.

a hlavně dokáže chyby pouze detekovat, nikoli opravovat. Přidejme nyní paritní bity i pro sloupce takto vzniklé matice (tabulka 5.6). Takto přidávané redundanci se říká vertikální (nebo též *transversal redundancy check*). Tím se dostáváme ke kódování, kterému se v literatuře říká vícerozměrová redundance (*multidimensional parity-check code*). V tomto případě jde o dvoudimenzionální kódování.

Dvourozměrná struktura

První, kdo se zabýval vícerozměrovým kódováním, byl zřejmě Elias (1954), pro formální popis struktur, které popisuji v této kapitole, a pro představu o jejich dnešním využití doporučuji Shea – Wong (2003) a další práce těchto autorů.[11]

[11] Multidimenzionální kódování se také někdy označuje termínem *product code* (násobný kód), který je ovšem homonymní s *kódem produktu*, kvůli čemuž je velmi špatně hledatelný na Internetu.

0	0	0	1	1	0	1	0	1
0	1	1	0	1	0	1	1	1
0	1	1	0	0	1	1	0	0
0	1	1	1	0	0	0	1	0
1	0	1	1	1	0	0	0	0
1	0	0	0	0	0	1	1	1
1	1	0	1	1	0	1	0	1
0	0	0	0	1	0	0	0	1
1	0	0	0	1	1	1	1	1

Tabulka 5.6: Matice znázorňující příklad použití horizontální redundance. Redundantní bity jsou určeny pozicí — každý devátý bit je paritní vůči osmi předchozím a posledních devět bitů je paritních vůči sloupcům, které jsou v matici nad nimi. Pro přehlednost jsou zobrazeny v rámečku. Redundantní bit vpravo dole je paritní vůči osmi bitům, které jsou samy redundantní, je tedy označen ve dvojitém rámečku (anglicky se takový bit označuje jako *parity-on-parity bit*). Všimněte si, že tento poslední bit je paritní vůči sloupci i vůči řádku.

Přidáním vertikální redundance sice zvýšíme podíl redundantních bitů v zakódované zprávě, zato ale bude příjemce moci opravovat chyby, aniž by musel odesilatele žádat, aby zprávu poslal znovu (jedná se tedy o samoopravný kód), podobně jako jsme si to již ukazovali na modulární redundanci v kapitole 1.4. Pokud zašuměný kanál ve zprávě přetočí hodnotu jednoho bitu, pak bude nekonzistentní jeden sloupec a zároveň jeden řádek, takže pomocí těchto souřadnic bude moci příjemce zprávy najít a opravit poškozený bit (což je ilustrováno v matici 5.7).

Samozřejmě počet detekovatelných chyb je mnohem větší. Aby chybně přenesená matice zůstala konzistentní, bylo by nutné, aby všechny řádky i sloupce obsahovaly sudý počet chyb, jako například v matici (5.8), kde jsou právě dvě chyby ve třetím a sedmém řádku a druhém a sedmém sloupci.

Opět použijeme metodu Monte Carlo a podíváme se, jaký je praktický rozdíl mezi těmito dvěma způsoby kódování na příkladu zprávy o šestnácti bitech. Aby byly kódy srovnatelné, použijeme dvourozměrný kód o velikosti 4×4 bity a jednorozměrný kód o déce řádku 2 bity, čímž docílíme přibližně stejného podílu redundantních bitů.[12] V tabulce 5.9 vidíme, jak si tato kódování poradí s jednotlivými modely šumů.

Vzhledem k tomu, že dvourozměrná matice má schopnost opravovat pouze jeden

[12] V prvním případě je to 9 redundantních bitů z 25, v druhém 8 bitů z 24 (každý třetí bit je redundantní). Podle Hammingovy notace má první kód parametry $(25, 16)$ a druhý $(3, 2)$, tuto notaci budu používat pro další případy (první číslo značí délku zprávy po zakódování, tedy včetně redundance, druhé číslo délku zprávy před zakódováním).

0	0	0	1	1	0	1	0	1
0	1	1	0	1	0	1	1	1
0	1	1	1	0	1	1	0	0
0	1	1	1	0	0	0	1	0
1	0	1	1	1	0	0	0	0
1	0	0	0	0	0	1	1	1
1	1	0	1	1	0	1	0	1
0	0	0	0	1	0	0	0	1
1	0	0	0	1	1	1	1	1

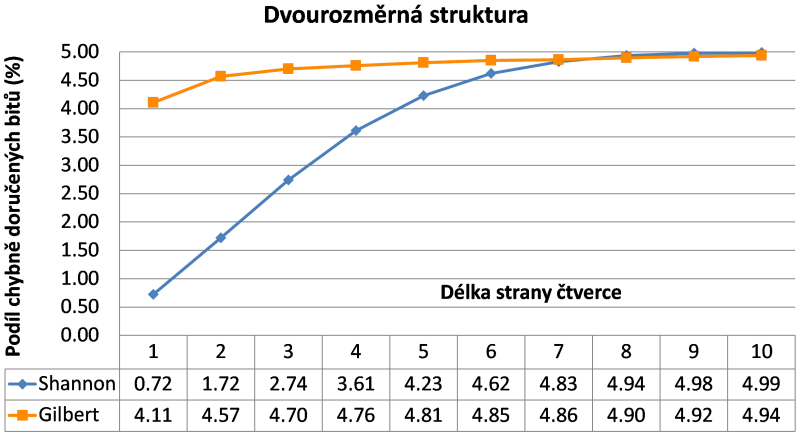
Tabulka 5.7: Data i redundance jsou stejné jako v tabulce 5.6, ovšem čtvrtý bit ve třetím řádku dorazil přes zašuměný kanál s převrácenou hodnotou, což se projeví tím, že třetí řádek a čtvrtý sloupec jsou nekonzistentní s paritními bity. To příjemci umožňuje vydedukovat, že byl poškozen právě tento bit, a opravit jej.

0	0	0	1	1	0	1	0	1
0	1	1	0	1	0	1	1	1
0	1	0	0	0	1	0	0	0
0	1	1	1	0	0	0	1	0
1	0	1	1	1	0	0	0	0
1	0	0	0	0	0	1	1	1
1	1	1	1	1	0	0	0	1
0	0	0	0	1	0	0	0	1
1	0	0	0	1	1	1	1	1

Tabulka 5.8: Data i redundance jsou stejné jako v tabulce 5.6, ovšem třetí a sedmý bit ve třetím řádku a třetí a sedmý bit na sedmém řádku dorazily přes zašuměný kanál s převrácenou hodnotou, což znamená, že ve všech sloupcích i řádcích je sudý počet chyb. Všechny řádky i sloupce jsou tak konzistentní se svými paritními bity. Příjemce tak nepozná, že zpráva dorazila poškozená.

chybný bit, nás nepřekvapí, že dvourozměrné kódování má vysokou úspěšnost pro Shannonův model šumu, ovšem s dávkovými šumy si neporadí, respektive úspěšnost se příliš neliší od jednorozměrného kódování. Nejčastěji totiž ve zprávě není chyba buď žádná, nebo je chyb více než jedna. V případě M3 má dokonce dvourozměrné kódování vyšší podíl *false positives* (zpráv, které byly interpretovány jako nepoškozené, avšak poškozené byly), což znamená větší množství nedorozumění — stavů, kdy recipient nerozumí, ale myslí si, že rozumí.

Nezajímá nás ovšem pouze sloupec -+ (zprávy, které jsou zdrojem nedorozumění), ale i sloupec ++, tedy zprávy, které dorazily v pořádku, neboť ukazuje na samoopravovací schopnost tohoto kódování. Přehledně schopnost opravovat chyby (respektive s výjimkou shannonovského kanálu spíše neschopnost) ukazují grafy 5.23 a 5.24, které vznikly tak, že čtvercem o hraně jeden bit (+ 1 bit redundantní) byla postupně zakódována zpráva o délce 10^8 b. Následně prošla zašuměným kanálem a byla dekódována a porovnána s původní zprávou. Totéž bylo provedeno se čtverci o délce strany 2–30.



	1	2	3	4	5	6	7	8	9	10
Shannon	0.72	1.72	2.74	3.61	4.23	4.62	4.83	4.94	4.98	4.99
Gilbert	4.11	4.57	4.70	4.76	4.81	4.85	4.86	4.90	4.92	4.94

Obrázek 5.23: Úspěšnost dvourozměrné struktury při přenosu dat. Čtverec 2×2 bity odstraní téměř dvě třetiny chyb z pětiprocentního shannonovského kanálu. U Gilbertova modelu je ovšem úspěšnost nevalná. Zašuměné kanály mají parametry stejné jako v tabulce 5.4.

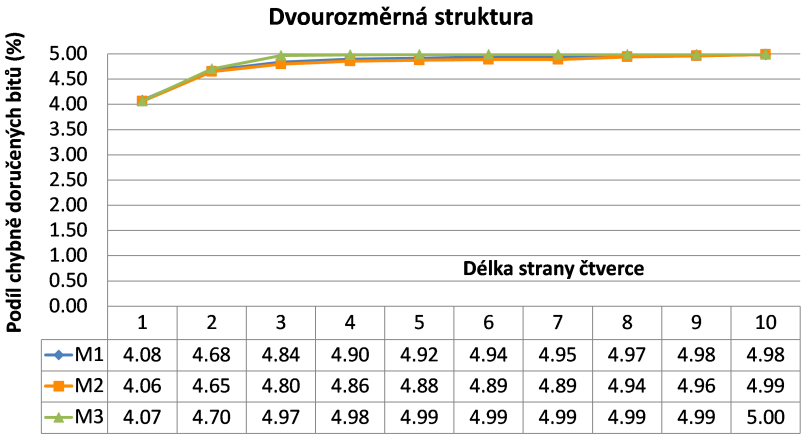
Pojďme se podívat, jak se situace změní, když přidáme další rozměr.

Trojrozměrná struktura

Protože tato disertace je tištěna pouze dvourozměrně, zkuste si onen třetí rozměr představit nad stránkou. Konkrétně krychli $4 \times 4 \times 4$. My zde (na obrázku 5.10) si ji

	Jednorozměrná				Dvourozměrná			
Správnost	+	+	-	-	+	+	-	-
Konzistence	+	-	+	-	+	-	+	-
Shannonův	29,2	14,8	2,0	54,0	64,2	3,1	1,7	31,0
Gilbertův	69,2	2,7	4,1	24,0	75,6	3,0	1,8	19,6
M1	67,2	2,0	4,7	26,1	71,0	3,8	2,6	22,7
M2	63,9	2,5	5,7	27,8	68,7	4,8	2,7	23,7
M3	76,1	0,7	2,0	21,2	77,4	1,2	2,0	19,4

Tabulka 5.9: Zpráva o 16 bitech je zakódována jednorozměrným kódem s parametry $(3, 2)$ a dvourozměrným kódem s parametry $(25, 16)$. Tabulka udává pravděpodobnosti (v procentech), s jakými zpráva, poté co prošla různými druhy zašuměných kanálů, byla poškozená (řádek Správnost) a jestli byla recipientem jako poškozená interpretována (řádek Konzistence). Zašuměné kanály mají parametry stejné jako na obrázku 5.4. Ve všech případech jsou 95% konfidenční intervaly nižší než desetina procenta.



Obrázek 5.24: Úspěšnost dvourozměrné struktury při přenosu dat. Podobně jako Gilbertův model, i další paměťové modely nejsou čtvercovou strukturou úspěšně opraveny. Zašuměné kanály mají parametry stejné jako v tabulce 5.4.

rozložíme po jednotlivých vrstvách a vložíme do ní stejnou zprávu jako v předchozím čtverci.

Bity pátého čtverce označeného jako R jsou paritní vůči bitům, které jsou pod nimi (představme si pátý čtverec jako horní vrstvu krychle). Tři rozměry nám umožňují opravu libovolného množství chyb na jednom řádku, což je zcela zásadní pro opravu dávkového šumu. Obecně platí, že n-rozměrná struktura dokáže plně opravit chyby v (n-2)-rozměrné struktuře, tedy že krychle dokáže opravit všechny chyby, které se objeví na jednom řádku (pokud jsou všechny ostatní řádky správně), čtyřrozměrná krychle pak všechny chyby na jedné ploše (pokud jsou ostatní plochy ve struktuře správně). Formálně a s odkazy na literaturu o tom pojednává Shea – Wong (2003, str. 8). Autoři uvádějí, že právě struktury tohoto typu byly inspirovány Gilbertovým dávkovým šumem a byly zavedeny právě za účelem jeho překonávání. Schopnost trojrozměrné struktury opravovat chyby ilustrují grafy 5.25 a 5.26. Nádavkem také získáváme větší detekci chyb, jak ukazuje tabulka 5.11, která ilustruje, jak si trojrozměrná struktura poradí s různými modely zašuměných kanálů.

Samozřejmě není nutné, aby struktura byla krychlová, místo krychle můžeme použít kvádr.

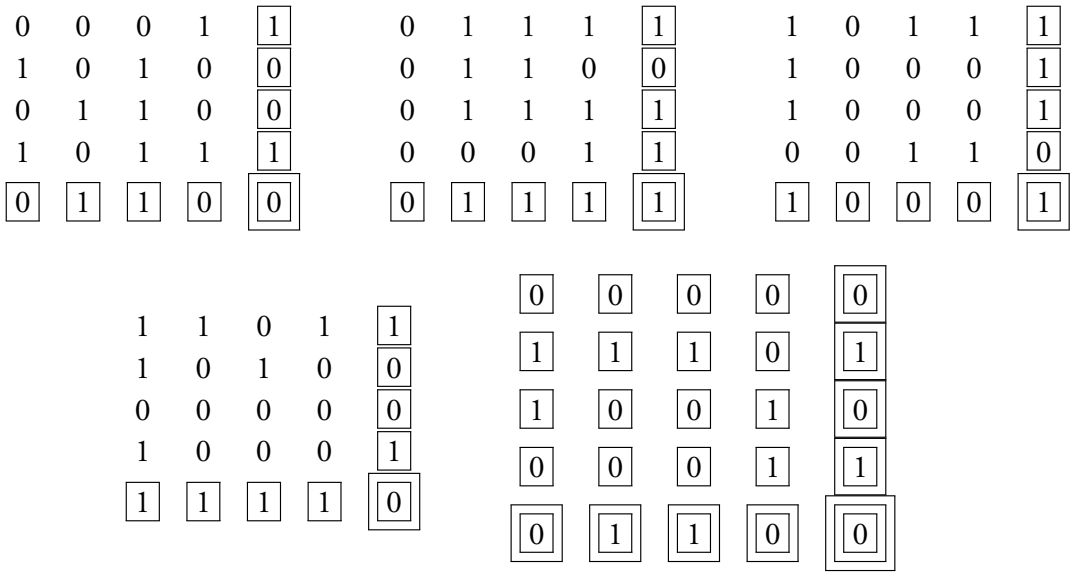
Čtyřrozměrná struktura

Nyní provedeme operaci, která bude klást na abstrakci čtenáře trochu vyšší nároky než dosud. Představte si, že zprávu rozdělíme do čtyřrozměrného hyperkvádru $4 \times 4 \times 2 \times 2$. Opět ho pro názornost rozložíme do plochy po jednotlivých vrstvách (tabulka 5.12). Počet redundantních bitů je odvozen od vzorce pro objem tělesa, do kterého posílanou zprávu vtěsnáme. Pro k-rozměrnou hyperkrychli o hraně n bitů, z toho jeden redundantní, se tedy počet redundantních bitů (R) spočítá podle následujícího vzorce (5.13).

$$R = n^k - (n-1)^k \tag{5.13}$$

Analogicky k tomuto vzorci (5.13) se počet redundantních bitů R v tabulce 5.12 spočítá jako $R = 5 \cdot 5 \cdot 3 \cdot 3 - (5-1) \cdot (5-1) \cdot (3-1) \cdot (3-1) = 161$. Rozhodnutím o počtu rozměrů a velikosti hrany tak regulujeme míru vkládané redundance. Analogicky k trojrozměrné struktuře umožňuje čtyřrozměrná struktura opravit libovolné dva bity nebo libovolnou plochu, tedy umožňuje překonat ještě větší dávky šumu než tři rozměry.

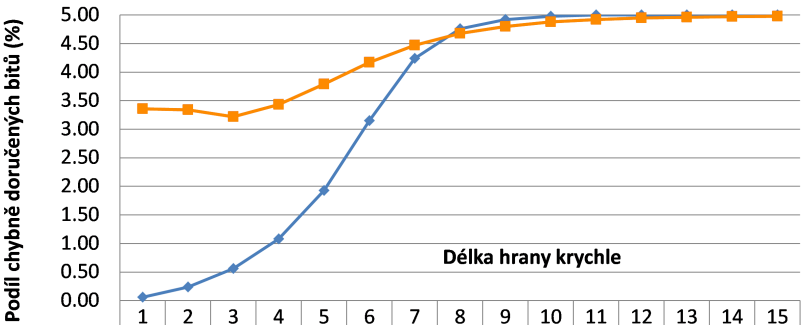
Podobně jako v předchozích případech, i zde si uvedeme porovnání s jednodušší strukturou (tabulka 5.13). Na první pohled nás zarazí, že vysoká schopnost opravování chybně přenesených zpráv (a to i pro paměťové modely) je vykoupena vysokou mírou chybných zpráv mylně považovaných za správné (*false positives*). Má implementace dekódování je ovšem jen jednou z mnoha možných; nebyl by problém vytvořit algoritmus, který opravuje chyby méně agresivně, a podíl *false positives* zmenšit na úkor oprav.



Tabulka 5.10: Data jsou stejná jako v tabulce 5.6, jen jsou přeskládaná ze čtverce do krychle o hraně délky pět bitů, z toho jeden paritní. Jako v předchozích případech paritní bit je pro přehlednost naznačen rámečkem, paritní bit, který počítá paritu z jiných paritních bitů, je naznačen dvojitým rámečkem, a trojitým rámečkem je značen bit, který je paritní vůči paritním bitům, které jsou paritní vůči jiným paritním bitům.

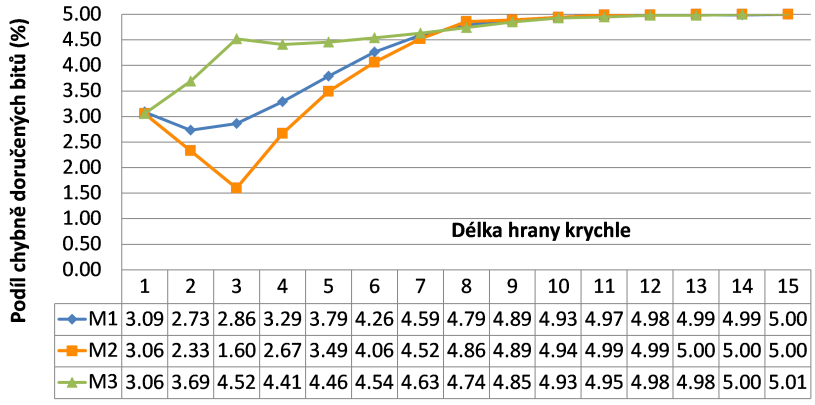
	Dvourozměrná				Trojrozměrná			
Správnost	+	+	-	-	+	+	-	-
Konzistence	+	-	+	-	+	-	+	-
Shannonův	59,6	8,6	1,5	30,2	92,9	1,2	1,2	4,6
Gilbertův	54,0	4,9	6,0	35,2	77,1	3,3	2,6	17,0
M1	44,7	5,3	9,3	40,8	74,4	3,8	3,6	18,2
M2	26,8	9,6	15,9	47,7	79,1	5,3	0,0	15,6
M3	67,1	1,7	3,0	28,3	70,8	1,3	2,0	25,9

Tabulka 5.11: Zpráva o 28 bitech je zakódována dvourozměrným kódem s parametry $(9, 4)$ a zpráva o 27 bitech trojrozměrným kódem $(64, 27)$. Tabulka udává pravděpodobnosti (v procentech), s jakými zpráva přišla poškozená (řádek Správnost) a jestli byla recipientem jako poškozená interpretována (řádek Konzistence). Zašuměné kanály mají parametry stejné jako v tabulce 5.4. Ve všech případech jsou 95% konfidenční intervaly nižší než desetina procenta.



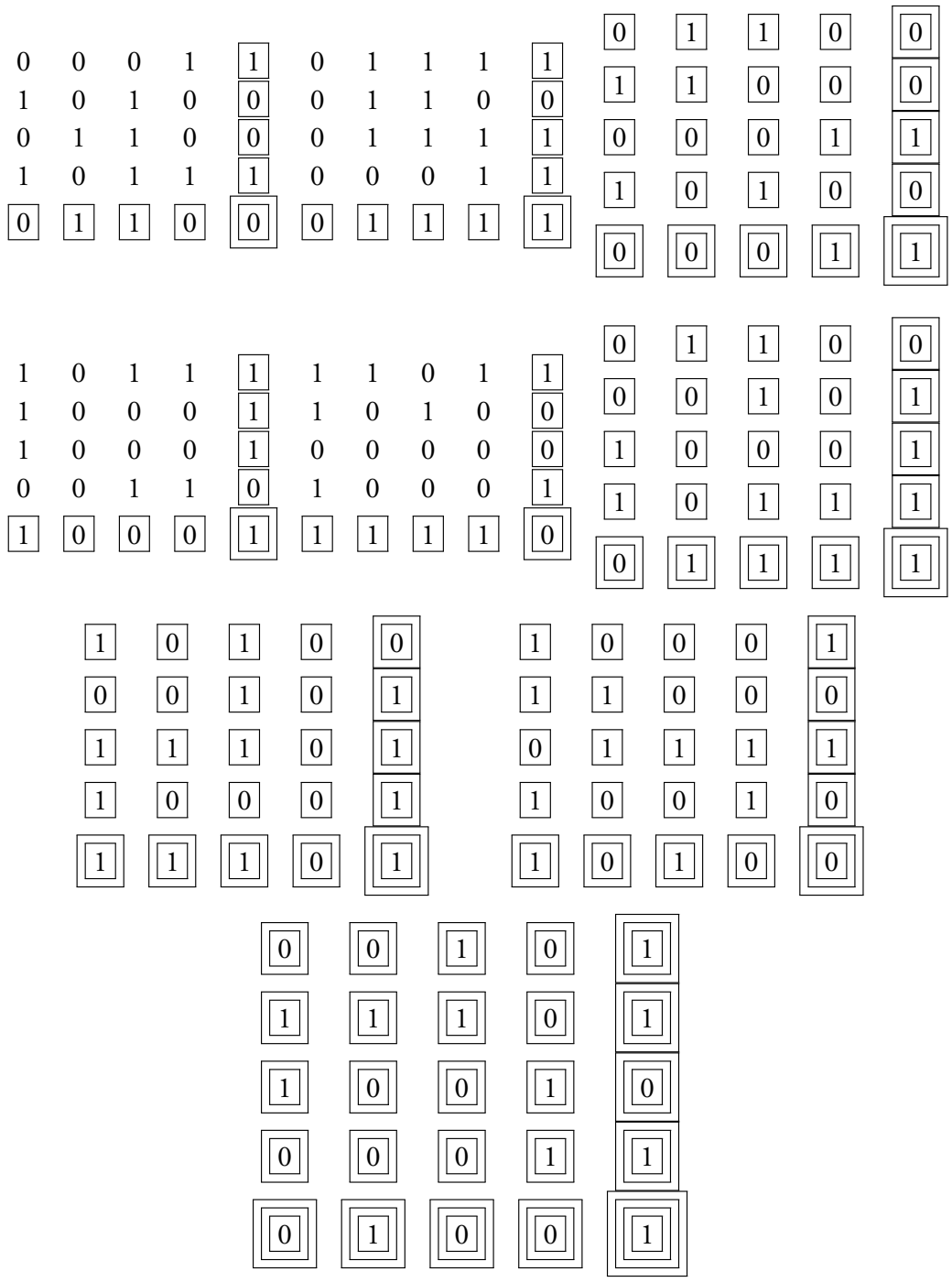
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Shannon	0.06	0.24	0.56	1.08	1.93	3.15	4.24	4.76	4.92	4.98	5.00	5.00	5.00	5.00	5.00
Gilbert	3.36	3.34	3.22	3.43	3.79	4.17	4.47	4.68	4.80	4.88	4.92	4.95	4.96	4.97	4.98

Obrázek 5.25: Úspěšnost trojrozměrné struktury při přenosu dat. Odstraňování chyb ze shannonovského kanálu je předvídatelné a v případě potřeby i jednoduše algebraicky odvoditelné. Chování struktury v konfrontaci s Gilbertovým modelem je méně intuitivní. Zašuměné kanály mají parametry stejné jako na obrázku 5.4.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M1	3.09	2.73	2.86	3.29	3.79	4.26	4.59	4.79	4.89	4.93	4.97	4.98	4.99	4.99	5.00
M2	3.06	2.33	1.60	2.67	3.49	4.06	4.52	4.86	4.89	4.94	4.99	4.99	5.00	5.00	5.00
M3	3.06	3.69	4.52	4.41	4.46	4.54	4.63	4.74	4.85	4.93	4.95	4.98	4.98	5.00	5.01

Obrázek 5.26: Úspěšnost trojrozměrné struktury při přenosu dat. Podobně jako Gilbertův model, i další paměťové modely se chovají neintuitivně. V tomto ohledu si povšimněme zejména modelu M2, který má maximální úspěšnost užití krychle o délce hrany tři bity. To zřejmě souvisí s parametry kanálu, neboť průměrná dávka má délku osm bitů a jedna plocha této krychle má obsah 16 bitů (jeden bit je redundantní, proto 4×4, a nikoli 3×3). Kanály mají parametry stejné jako na obrázku 5.4.



Tabulka 5.12: Data jsou stejná jako v tabulce 5.6, jen jsou přeskládaná místo do čtverce do čtyřrozměrného hyperkvádru o hranách délek $5 \times 5 \times 3 \times 3$ bitů, z toho jeden bit je vždy paritní. Jako v předchozích případech, paritní bit je pro přehlednost naznačen rámečkem, paritní bit počítající paritu z jiných paritních bitů pak vícenásobným rámečkem.

	Trojrozměrná				Čtyřrozměrná			
Správnost	+	+	-	-	+	+	-	-
Konzistence	+	-	+	-	+	-	+	-
Shannonův	51,6	6,6	6,5	35,4	87,4	1,6	3,6	7,3
Gilbertův	10,4	4,2	3,3	82,1	65,0	1,8	14,3	18,9
M1	7,1	3,9	3,7	85,2	64,1	2,2	15,8	17,9
M2	11,3	8,7	0,0	80,0	67,3	3,2	10,2	19,2
M3	16,3	1,0	1,7	81,0	18,4	0,9	1,2	79,5

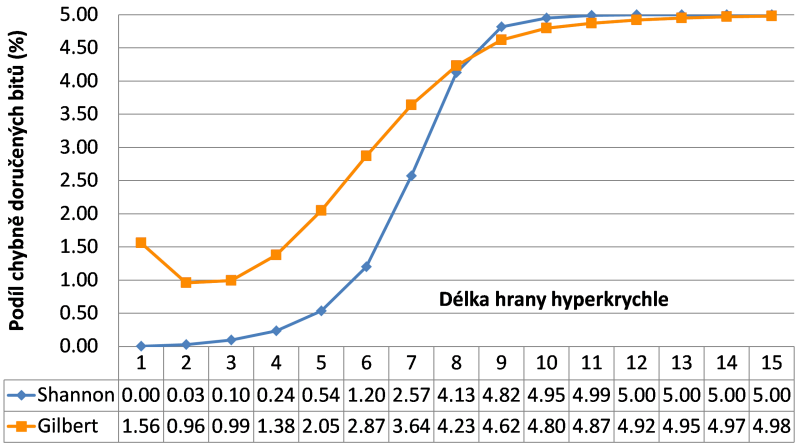
Tabulka 5.13: Zpráva o 243 bitech je zakódována trojrozměrným kódem s parametry $(64, 27)$ a zpráva o 256 bitech čtyřrozměrným kódem $(625, 256)$. Tabulka udává pravděpodobnosti (v procentech), s jakými zpráva, poté co prošla různými druhy zašuměných kanálů, byla poškozená (řádek Správnost) a jestli byla recipientem jako poškozená interpretována (řádek Konzistence, myšleno konzistence s paritním bitem). Zašuměné kanály jsou nastaveny tak, aby pravděpodobnost změny bitu $p = 0{,}05$ (respektive mají parametry jako na obrázcích 5.6, 5.12, 5.15 a 5.19). Ve všech případech jsou 95% konfidenční intervaly nižší než desetina procenta.

I nyní si uvedeme grafy zachycující schopnost čtyřrozměrného kódování opravovat chyby (5.27 a 5.28). Nejspíše i díky agresivnímu algoritmu je podíl opravených chyb výrazně vyšší než u trojrozměrné struktury.

Grafy 5.29, 5.30, 5.31, 5.32 a 5.33 rekapitulují jednotlivá kódování. Zkoušíme struktury jednu po druhé: vytvoříme čtverec o straně 1 bit a s jeho pomocí postupně zakódujeme zprávu o 10^8 bitech, následně totéž provedeme se čtvercem o straně délky dva bity atd. až do třiceti. Poté totéž provedeme s krychlí o hraně 1 až 30 bitů a hyperkrychlí. Aby byly výsledky porovnatelné, jsou v grafu vyneseny v závislosti nikoli na délce hrany, ale na tom, kolik bitů z odeslané zprávy zaujímá původní zpráva a kolik redundance.

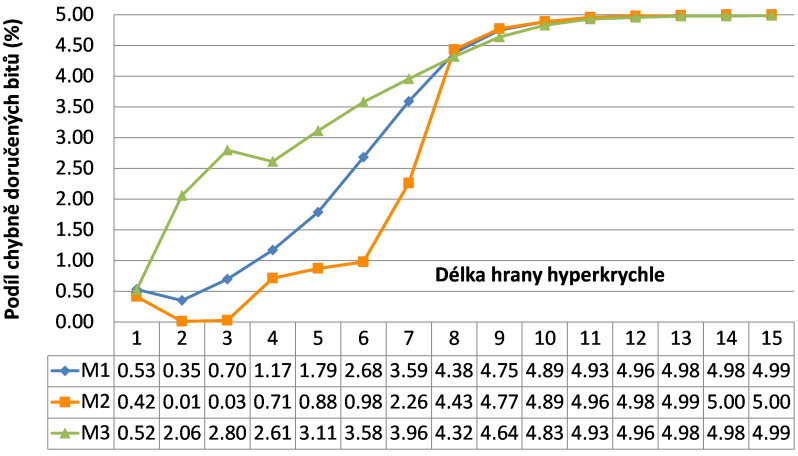
Vraťme se nyní ke čtyřrozměrné struktuře, která, jak se ukázalo, se dokáže se ctí vypořádat s paměťovými zašuměnými kanály. V tabulce 5.12 jsem čtyři rozměry rozdělil na čtverce, protože to tak bylo přehlednější, ovšem při posílání zprávy je možné ji poslat jako řadu za sebou následujících bitů. Zkusme si tedy onen čtyřrozměrný kvádr zobrazit v jednom rozměru (například v čase), zachovávajíce onu pověstnou saussurovskou „linearitu“ (tabulka 5.14).

Dělení této zprávy vám možná připomíná vícenásobné vnořené dělení v přirozeném jazyce, třeba na hlásky, morfémy, slova a věty. Můžete si například představit hlásky o pěti distinktivních rysech, z nichž jeden je redundantní, ze kterých se skládají morfémy o pěti hláskách, z nichž jedna je redundantní, z nich slova o třech morfémech (z nichž jeden morfém je redundantní) a z nich větu o třech slovech (z nichž



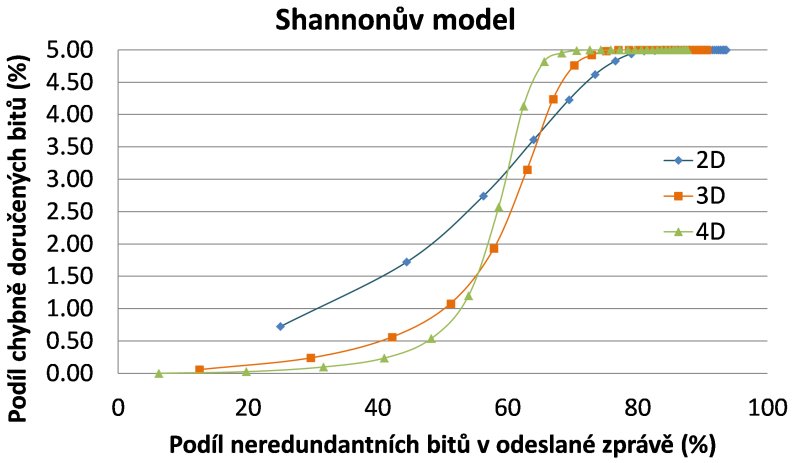
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Shannon	0.00	0.03	0.10	0.24	0.54	1.20	2.57	4.13	4.82	4.95	4.99	5.00	5.00	5.00	5.00
Gilbert	1.56	0.96	0.99	1.38	2.05	2.87	3.64	4.23	4.62	4.80	4.87	4.92	4.95	4.97	4.98

Obrázek 5.27: Úspěšnost čtyřrozměrné struktury při přenosu dat. 90 % chyb shannonovského kanálu je možno odstranit při použití hyperkrychle o hraně pět bitů. Ideální hyperkrychle pro odstraňování gilbertovského šumu má hranu tři bity. (Zašuměné kanály mají parametry stejné jako v tabulce 5.4.)

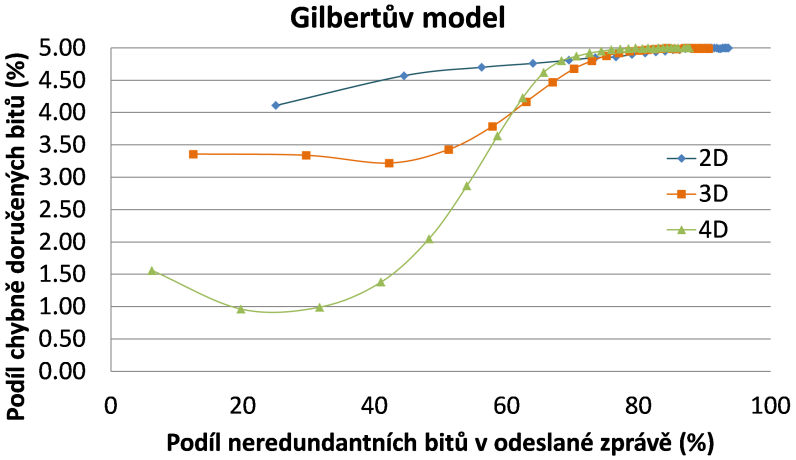


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M1	0.53	0.35	0.70	1.17	1.79	2.68	3.59	4.38	4.75	4.89	4.93	4.96	4.98	4.98	4.99
M2	0.42	0.01	0.03	0.71	0.88	0.98	2.26	4.43	4.77	4.89	4.96	4.98	4.99	5.00	5.00
M3	0.52	2.06	2.80	2.61	3.11	3.58	3.96	4.32	4.64	4.83	4.93	4.96	4.98	4.98	4.99

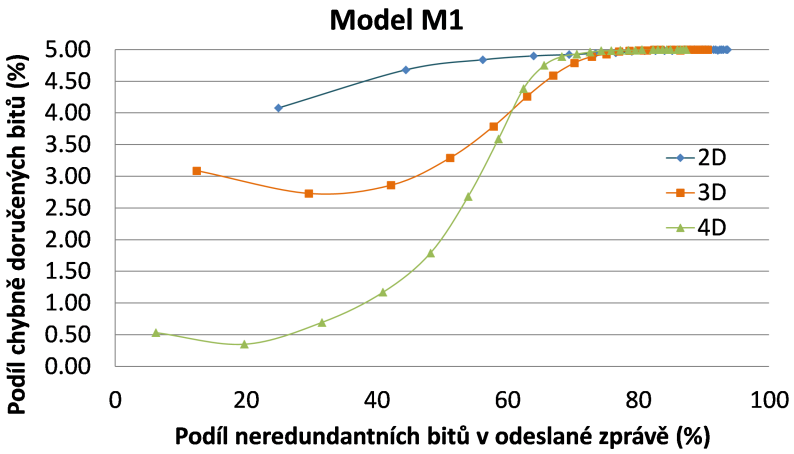
Obrázek 5.28: Úspěšnost čtyřrozměrné struktury při přenosu dat. Další neintuitivní chování. Pro model M2 má hyperkrychle o hraně dva a tři bity téměř stoprocentní úspěšnost, což zřejmě opět souvisí s parametry kanálu, neboť průměrná dávka má délku osm bitů a jedna plocha hyperkrychle o hraně dva bity (+ 1 redundantní) má obsah devět bitů, přičemž, jak bylo řečeno, čtyřrozměrná struktura umí bezezbytku opravovat právě jednu plochu. (Zašuměné kanály mají parametry stejné jako v tabulce 5.4.)



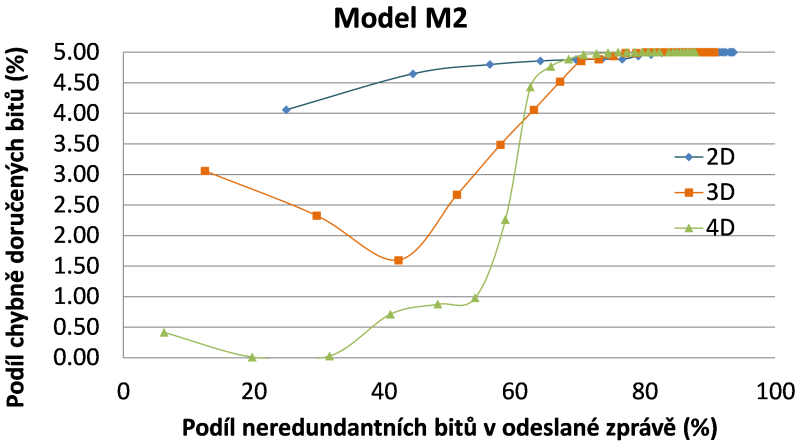
Obrázek 5.29: Úspěšnost dvou-, troj- a čtyřrozměrného kódování při překonávání shannonovského zašuměného kanálu (má stejné parametry jako na obrázku 5.4).



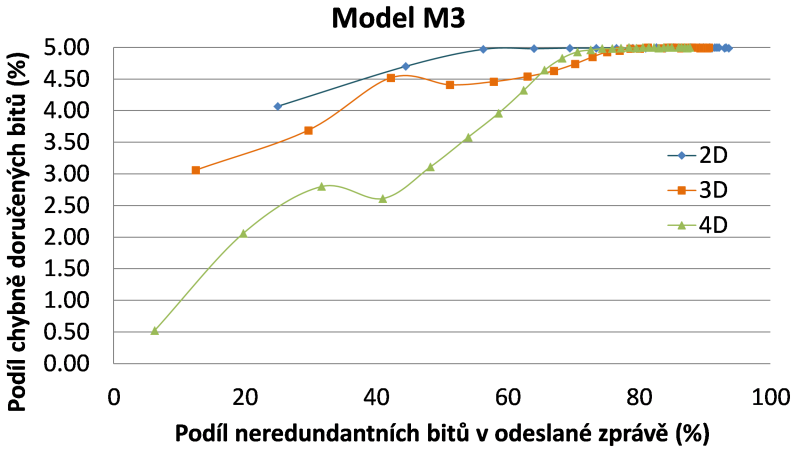
Obrázek 5.30: Úspěšnost dvou-, troj- a čtyřrozměrného kódování při překonávání Gilbertova zašuměného kanálu (má stejné parametry jako na obrázku 5.6).



Obrázek 5.31: Úspěšnost dvou-, troj- a čtyřrozměrného kódování při překonávání zašuměného kanálu podle modelu M1 (má stejné parametry jako na obrázku 5.12).



Obrázek 5.32: Úspěšnost dvou-, troj- a čtyřrozměrného kódování při překonávání zašuměného kanálu podle modelu M2 (má stejné parametry jako na obrázku 5.15).

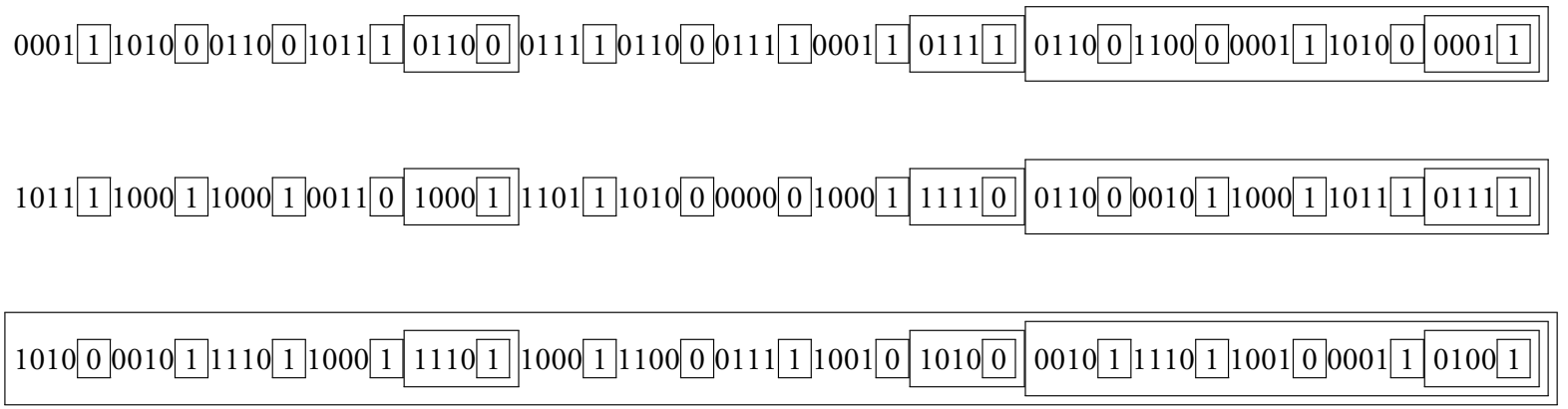


Obrázek 5.33: Úspěšnost dvou-, troj- a čtyřrozměrného kódování při překonávání zašuměného kanálu podle modelu M3 (má stejné parametry jako na obrázku 5.19).

jedno je opět redundantní). Když se vrátíme k zobrazení v hyperkvádru, tak jednotlivé řádky by znázorňovaly hlásky, plochy morfémy, kvádry slova a celý hyperkvádr by znázorňoval větu.

Dostali jsme se k takovému modelu přidávání redundance, který vyžaduje segmentaci na několika úrovních, jež se vizuálně podobá textu v přirozeném jazyce. Nicméně tvrdit, že ona podobnost je důkazem jeho smysluplnosti, by bylo pošetilé.[13] Navíc v přirozeném jazyce nic takového jako paritní bity nedopočítáváme, respektive to tak nevnímáme. Přiřazování paritního bitu je jenom jednou z mnoha možných funkcí, které lze k redundování použít. Abychom mohli pokračovat v tvorbě realistického modelu, musíme si tedy napřed v následující podkapitole ukázat, jaké funkce se v jazyce k redundování vlastně používají.

[13] Jak o podobné záležitosti poznamenal Stephen Wolfram, „bylo by to, jako kdybyste v nějakých datech rozpoznali Fibonacciho posloupnost, a usuzovali z toho, že jev, který za daty stojí, musí mít něco společného s králíky“.



Tabulka 5.14: Data i redundance jsou stejné jako v tabulce 5.6, jen jsou rozložené do jednoho rozměru.

5.3 Redundance v jazyce

Zatím jsme pracovali zejména s jednou funkcí pro vložení užitečné redundance: paritním bitem. Ovšem redundancí, která napomáhá úspěšnému přenosu komunikace, je vlastně každý redundantní prvek, který zvyšuje pravděpodobnost, že když dojde při přenosu k chybě, zpráva se stane pro recipienta nekonzistentní.

Redundantní jsou tedy jakékoli informace, které ze sebe vzájemně vyplývají nebo vyplývají z toho, co příjemce již ví ať už o sdělovaném předmětu, nebo o jazyce. Tedy pokud ovšem samotný fakt vyplývání není informací, kterou chce text sdělit.

Například kontext omezuje množinu slov, která můžete použít. Nebo alespoň zmenší pravděpodobnost užití některých slov, což bereme jako samozřejmost a přirozenou vlastnost jazyka, nicméně samozřejmé to není — ono omezení vyplývá z toho, že informace, které hodláme sdělit, musí být částečně redundantní vůči informacím, které už recipient zná.

Uveďme si kategorizaci nejčastějších strategií pro vkládání redundance, účelné mi připadá rozdělení na následující tři druhy:

1. Pravidla a omezení, která si produktor textu stanoví sám dopředu. Například před tímto seznamem jsem vás upozornil, že bude obsahovat tři prvky. Kdybych to nedodržel a uvedl jich méně, věděli byste, že je něco špatně, že text je nekonzistentní.

2. Pravidla a omezení, která jsou stanovena jazykem, čili nenesou význam a jsou součástí přenosového kódu. Například pokud byste v této kapitole viděli nadkritické množství neznámých slov a syntax, která podle vašich představ není validní, pak byste uznali, že textu nerozumíte.

3. Pleonastická vyjádření, tedy nevynucené opakování téže informace. Jde v podstatě o princip modulární redundance, která je popsána v kapitole 1.4. Pokud byste se teď podívali do kapitoly 1.4 a zjistili, že modulární redundance nespočívá v opakování téhož prvku vícekrát, pak byste právem získali pocit, že informace, kterou vám chci tímto textem sdělit, nedošla svého cíle nepoškozena.

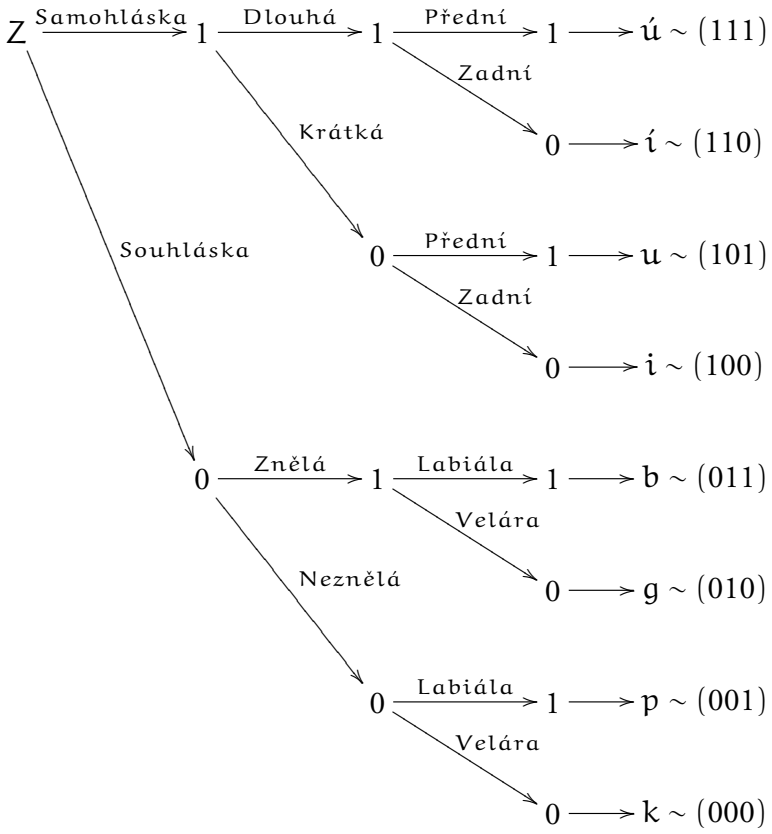
Toto rozdělení budu také uplatňovat v podkapitolách kapitoly 6. Co konkrétně si představuji pod redundancí tvořenou pomocí omezení a co pod redundancí pleonastickou, se dozvíte již v následujících podkapitolách.

5.3.1 Redundantní restrikce

Jak se dosahuje redundantnosti výpovědi pomocí restrikcí? Nejlépe si to vysvětlíme na příkladu. Představme si „jazyk“, který má osm písmen {u, ú, i, í, p, b, g, k}, která se

v textech vyskytují zhruba s rovnoměrnou pravděpodobností a jedno písmeno kóduje tři bity (neboť $8 = 2^3$).

Mohli bychom jednat s úplně libovolnými symboly, ale řadu symbolů jsem zvolil schválně tak, aby si lingvista mohl představit tříbitové kódování jako sled tří distinktivních rysů. Představme si, že symboly reprezentují takové hlásky, že *u*, *ú*, *i* a *í* jsou samohlásky, z toho *u* a *ú* zadní, *i*, *í* přední; *ú* a *í* dlouhé a *i* a *u* krátké. Tedy první z oněch tří bitů značí, že se jedná o samohlásku, druhý označuje pozici na předozadní škále a třetí bit délku. Znaky *p*, *b*, *g* a *k* jsou popsány obdobně: první distinktivní rys (bit) značí, že se jedná o souhlásku, což určuje, že další bity budou znamenat něco jiného než u samohlásek: druhý bit určuje, jestli je znělá nebo neznělá, a třetí, jestli se vyslovuje v hrdle nebo rty. Například *í* je zakódovatelné jako samohláska — dlouhá — zadní, čili 101, *g* se zakóduje jako 010. Pro úplnou názornost uvedeme rozhodovací strom na obrázku 5.34 a tabulku 5.15, které každému písmenu přiřazují jeho distinktivní rysy.



Obrázek 5.34: Rozhodovací strom, podle kterého se jednotlivým symbolům přiřadí bitová hodnota podle jejich distinktivních rysů.

Znak	Samohláska	Dlouhá	Přední	Znělá	Labiála
u	1	0	1	N/A	N/A
ú	1	1	1	N/A	N/A
i	1	0	0	N/A	N/A
í	1	1	0	N/A	N/A
p	0	N/A	N/A	0	1
b	0	N/A	N/A	1	1
g	0	N/A	N/A	1	0
k	0	N/A	N/A	0	0

Tabulka 5.15: Tabulka kódování jednotlivých symbolů pomocí tří bitů, jež reprezentují distinktivní rysy.

Například pro přenos slova *bugíb* musíme použít 15 bitů,[14] což souhlasí s tím, že všech kombinací, kterých může pětipísmenné slovo nabývat, je 8^5, tedy 32768, což je rovno 2^{15}.

Nyní uveďme do slovotvorby nějaká omezení, například že slovo musí začínat na písmeno z množiny $\mathrm{A} = \{\mathrm{p}, \mathrm{b}, \mathrm{g}, \mathrm{k}\}$ a že po každém písmenu z množiny A nesmí následovat písmeno z množiny A, ale buď konec zprávy, nebo písmeno z množiny $\mathrm{B} = \{\mathrm{i}, \mathrm{u}, \mathrm{ú}, \mathrm{i}, \mathrm{í}\}$; a naopak po písmeni z množiny B nesmí následovat písmeno z množiny B. Tedy jakési pravidelné střídání souhlásek a samohlásek. Slovo *bugíb* je tak validní, zatímco třeba *ggpubgíp* není s tímto pravidlem konzistentní. Pokud se při pouti přes zašuměný kanál slovo *bugíb* změní na *ugíb*, budeme vědět, že je poškozené.

Na první pozici tedy mohou být čtyři různá písmena, na druhé také čtyři, na třetí čtyři, na čtvrté čtyři, na páté čtyři. Celkový počet kombinací tedy může být $4 \times 4 \times 4 \times 4 \times 4 = 1024$. To je možno zakódovat deseti bity (neboť $1024 = 2^{10}$). Pokud při kódování použijeme původní kódování (tři bity na písmeno), pak pět bitů ze zprávy je redundantních.[15]

Pravidla v jazyce samozřejmě neplatí stoprocentně, i kdyby platila, vzhledem k naší neschopnosti je zcela dodržovat musí mít příjemce textu nastavenu určitou míru tolerance k chybám. Pravidla dokonce mohou být jen jakési tendence, takže použití nějaké konstrukce prostě jen u příjemce sníží subjektivně vnímanou pravděpodobnost, že zprávě rozumí tak, jak byla míněna.

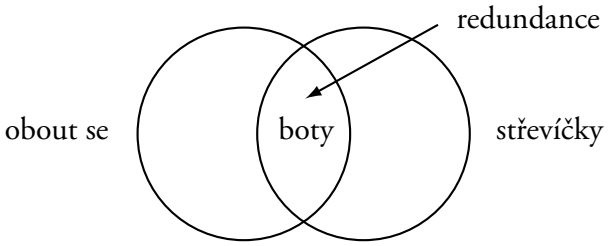
Zde se konečně dostáváme k vysvětlení toho, proč jedna výpověď ovlivňuje vhodnost nebo pravděpodobnost užití výpovědi jiné. Každá taková restrikce vnáší do textu redundanci. Situaci si můžeme osvětlit na příkladu následující věty:

[14] Konkrétně těchto 15 bitů: 011101010110011.

[15] Ovšemže bychom mohli redundanci při komunikaci systémově odstranit a racionalizovat přenosový kód tak, že se produktor a příjemce domluví na tom, že první bit z každého písmene vynechají a hodnotu písmene desambiguují podle jeho pozice. Slovo *bugíb* by se tak kódovalo jako 1101101011.

Obula jsem si střevíčky.

Přísudek obsahuje informaci o tom, že si mluvčí dala na nohy boty, a zároveň restringuje význam předmětu, který se na něj váže, na nějaký druh bot. Informace o botách je ve větě vyjádřena implicitně dvakrát (tuto situaci ilustruje obrázek 5.35).[16]



Obrázek 5.35: Informace o botách je implicitně obsažena jak ve sdělení, že se mluvčí obula, tak ve sdělení, že to, co si na sebe dala, byly střevíčky. První sdělení restringovalo možná pokračování věty, čímž se vytvořila redundance.

Jako jakousi tendenci k omezení je třeba vnímat i nevyrovnanou frekvenci inventáře hlásek, morfů a podobně. Inventář hlásek je dokonce (v „synchronním“ pohledu) konečný. Ale i pro morfémy a slova, která nemají velikost inventáře omezenu (naopak na neznámá slova narážíte každou chvíli i jako rodilý mluvčí), platí, že některé morfémy nebo slova jsou nepravděpodobné.

Funkce vymezující distribuci frekvence morfémů a slov v textech, jakož i různých jiných prvků v jazyce, má mocninný charakter (podle Zipfova modelu) a tuto nevyrovnanost lze klasifikovat jako redundanci. Podle Shannonova pojetí platí, že systémy, jejichž stavy nemají všechny stejnou pravděpodobnost výskytu, jsou redundantnější než ty, které ji stejnou mají (Mandelbrot, 1953). Jenže distribuce slov a Zipfův vztah jsou pro nás stále tajemné a kontroverzní.[17] Existuje totiž velké množství náhodných procesů, které vedou k této distribuci, zejména dynamické systémy s pozitivní zpětnou vazbou mají tendenci takovou distribuci prvků produkovat (Barabási, 2003, zejména kapitola Six Link a následující).

Co se jazyka týče, může jít o důsledek vlastností lidského mozku — priming se projevuje tím, že slovu nebo jinému jazykovému prvku, který jsme použili, stoupá pravděpodobnost opětovného užití. Po čase ona pravděpodobnost opět klesá. Simon

[16] Interpretace tohoto omezení jakožto redundantní informace v lingvistice již zdomácněla, ostatně je na ní založeno MI-Score, dnes asi nejoblíbenější metrika kolokability. *Mutual information* je, kromě svého klasického shannonovského smyslu, také definována v termínech kolmogorovovské komplexity, k tomuto tématu doporučuji opět Li – Vitányi (2013, str. 198).

[17] Vysvětlení toho, proč místo pojmu Zipfův zákon rozlišuji mezi Zipfovým vztahem a Zipfovým modelem, najdete v poznámce 14 na straně 117.

(1955), aniž by ovšem mluvil o primingu, z tohoto procesu odvodil funkci, která velmi dobře modeluje distribuci slov, a Rapoport (1982) na podobném principu založil svůj model Zipfova vztahu. Další kontroverze se týká vlivu nejazykových skutečností, mnoho z nich se totiž ve světě distribuuje tímto způsobem, aniž by to jakkoli souviselo s jazykem. Anekdotická je připomínka Chomského, že věta *I live in New York* se bude v korpusu vyskytovat častěji než *I live in Dayton (Ohio)*. Přitom právě distribuce velikosti měst Zipfovu modelu zhruba odpovídá. Podle Stefanowitsche (2005) velikost populace těchto měst s jejich frekvencí v korpusu dobře koresponduje, což je názornou ilustrací toho, že distribuce prvků v jazyce může být odrazem těchto distribucí.

Toto tvrzení můžeme zobecnit a přiznat si, že v této chvíli nemáme příliš jasno v tom, které struktury v jazyce jsou zděděny z pozorování vnějšího světa, ze schopnosti našich smyslů, ze způsobu, jakým mozek zpracovává a ukládá informace, a které jsou skutečně dány jazykem, tedy metodou mezilidské komunikace. Navíc některé metody, které náš mozek při komunikaci používá, se nejspíš vyvinuly v „době předjazykové“, například rozlišování hlásek je zřejmě jen speciální aplikací kognitivních systémů obecně používaných pro segmentaci a clusterování (Oudeyer, 2006). Také paměť není ničím jiným než časovým zašuměným kanálem, který informace musí překonat, aby se nám vybavila; strategie pro komunikaci mezi lidmi může být odvozena od strategie, již užíváme pro komunikaci se svým pozdějším já, neboť ta je evolučně starší.

Redundance vzniklá vlivem nějakých s jazykem nesouvisejících procesů ovšem může být exaptována a využita pro účely komunikace — příjemce textu ji může využít pro odhalování a korekci chyb stejně jako redundanci, kterou produktor vložil za tímto účelem záměrně.

5.3.2 Pleonastická redundance

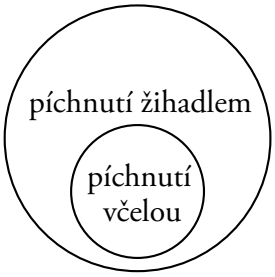
Pleonastická redundance, oproti restrikční redundanci popsané v předchozí kapitole, není vnucená — mluvčí ji tedy nevkládá proto, aby dostál požadavkům jazykového kódu, ale proto, že z nějakého důvodu chce. Je asi intuitivně pochopitelná, přesto si ji pro jistotu také ukážeme na příkladu. V následujícím souvětí informace, že mluvčího včela píchla žihadlem, vyplývá ze sdělení, že ho ona včela vůbec píchla (obr. 5.36).

Představ si, píchla mě včela, normálně žihadlem, potvora jedna.

Vyplývání může být i oboustranné, tedy ekvivalence, čili přímé opakování:

Zatracená včela, zatracená včela.

Podobně jako u restrikční redundance, ono vyplývání nemusí být úplné ani výlučné, proto obvykle nemůžeme říct, který ze dvou prvků je redundantní. Dva prvky



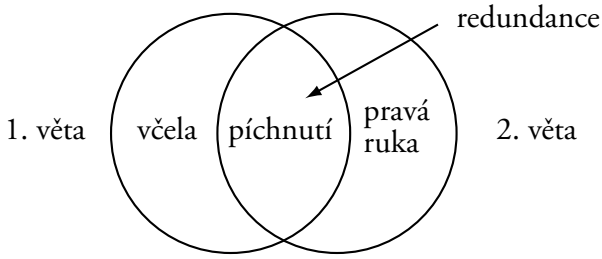
Obrázek 5.36: Píchnutí včelou je obvykle provedeno žihadlem. Informace o píchnutí žihadlem je tedy uvedena v souvětí dvakrát.

textu, které jsou vzájemně redundantní, totiž obvykle oba obsahují i nějakou neredundantní informaci.

Například si uveďme dvě věty:

Píchla mě včela, dostal jsem žihadlo do pravé ruky.

První věta obsahuje informaci o píchnutí mluvčího a o tom, že ho provedla včela. Z druhé věty se dozvídáme, že mluvčí byl píchnut a že byl píchnut do pravé ruky. Informace o píchnutí je tedy v jedné z těchto vět redundantní, nemůžeme ovšem určit, ve které ze slov je ona informace redundantní (tuto situaci ilustruje obrázek 5.37).



Obrázek 5.37: Informace o píchnutí je v jedné z těchto vět redundantní, nemůžeme ovšem říci, ve které z nich.

Už v kapitole 1.2 jsem zmínil, že co ze zprávy je redundance a co komplexita, záleží na recipientovi. Dokonce i přímé opakování identické kopie prvku může přinést novou informaci. Samotný akt redundování totiž může mít pro příjemce informační hodnotu: pokud je nějaká informace redundovaná intenzivněji než jiné, může recipient předpokládat, že je tato informace důležitější a že na jejím přenosu produktorovi více záleželo. Popřípadě může recipient vydedukovat, že si produktor myslí, že na ní recipientovi více záleží, nebo by alespoň více záležet mělo. To, co produktor zamýšlí jako redundanci, tedy může být recipientem přijato jako komplexita.

Ostatně strategie použitelné pro redundanci mohou být nejen jako komplexita pochopeny, jejich použití takto může být i zamýšleno produktorem (například jako

emfatizace). Pro utvoření představy, jaké funkce mohou prvky typické pro redundanci nabývat kromě samotné redundance, doporučuji článek (Wit – Gillette, 1999, od strany 13).

To, jestli příjemce interpretuje informaci jako komplexitu, nebo jako redundanci, hlavně závisí na jeho znalostech jazykového kódu a světa vůbec. Například manželce s dítětem adresujete větu:

Dejte si pozor na toho čmeláka, ať vás nekousne.

Pro dítě může být informace, že čmeláci koušou, zcela nová a zásadní, zatímco pro manželku představuje redundanci. Ve skutečnosti si produktor nikdy nemůže být jist znalostmi a dekódovacími schopnostmi recipienta, který často ani není dopředu konkrétně znám, rovněž vlastnosti zašuměného kanálu může jen odhadovat, pročež míra redundance je vždy suboptimální a recipient musí být schopen se s tím vyrovnat. Například vy sami, pokud vám tato kapitola přijde nesrozumitelná, nejspíš dohledáváte informace z jiných zdrojů, naopak pokud podle vás obsahuje pouze dobře známé informace, pak ji čtete jen rychle a zběžně.

5.4 Realistický model vkládání redundance

V době, kdy vznikal jazyk, nebylo důležité, jestli je kódovací metoda optimální, ale jestli je možné ji implementovat pomocí stávajícího lidského kognitivního aparátu.

K dispozici byl mozek — výkonný nástroj pro rychlou ztrátovou kompresi dat a s tím související kategorizaci (viz 3. odstavec kapitoly 2.2), což je úvaha, kterou přede mnou učinil již Oudeyer (2006, str. 62), když se pokoušel najít způsob, jakým se jazyk dopracoval k segmentaci a diskretizaci hlásek a distinktivních rysů, tvrdě, že je to jazyk, který se přizpůsobuje obecným učebním mechanismům, nikoli obráceně.

Neustálé přiřazování kategorií jednotlivým vjemům a následné hledání vztahů mezi těmito kategoriemi bylo a je denním chlebem nejen člověka, ale i dalších zvířat, ať už těmito kategoriemi je potenciální predátor, nebo kořist. Přitom některé vztahy byly častější než jiné a některé byly prakticky nemožné. Kumulace nezvyklých vztahů znamenala nekonzistenci, a tak působila i jako přirozená kontrola schopností kognitivního aparátu (pokud zvíře vidí něco, co kategorizuje jako čápa a to něco bude mít barvu, kterou kategorizuje jako červenou, nejspíš je něco v nepořádku).

Tyto předjazykové schopnosti pak posloužily i při tvorbě protokolu pro přenos informací mezi lidmi — jazyka.

Enkódování obecně vypadá takto: informační proud je rozdělen na krátké segmenty (řádově jednotky bitů / distinktivních rysů). Do každého segmentu jsou doplněny další bity (distinktivní rysy) tak, aby se dal zařadit do nějaké povolené kategorie („existující“ hlásky). Následně se již jedná s touto kategorií a není nutné jednat s jednotlivými bity (nicméně je to možné). Z těchto segmentů se skládají vyšší segmenty

(v případě češtiny, arabštiny a dalších běžných jazyků jsou další kategorií morfémy), přičemž opět pouze některé kombinace a posloupnosti kategorií jsou povolené. I tyto vyšší segmenty je možné rozřadit do kategorií, přičemž kategorizací může být paralelně několik (například v případě češtiny kategorizace na předpony, kořen, přípony a koncovky a zároveň podle významu) a opět platí, že jen některé kombinace a posloupnosti kategorií jsou povolené, takže při skládání vyššího celku je mnoho posloupností „zakázaných“ nebo velmi málo častých[18] — takže je třeba upravit posloupnost segmentů nebo segmenty přeměnit tak, aby se změnila jejich kategorie, nebo přidat jiné redundantní segmenty. Tímto způsobem můžeme pokračovat dál kategorizací slov na slovní druhy a kategorizací podle významu (třeba podle holonymních a hyperonymních vztahů, následuje výběr povolené syntaktické struktury, popřípadě dalších struktur, které hrají roli. Tak můžeme pokračovat dále ke stále vyšším celkům.

Samozřejmě netvrdím, že enkódování probíhá reálně v mozku právě tímto způsobem (od nižších celků po vyšší), právě naopak tuto otázku ponechávám nezodpovězenu. Osobně předpokládám, že tvorba textu probíhá značně chaoticky. Pokud bychom se podívali na rovinu morfémy — slova — věta, dá se očekávat, že v průběhu tvorby věty mluvčí několikrát změní předpokládané syntaktické struktury a že těmto strukturám bude průběžně přizpůsobovat morfematické složení použitých slov (zejména morfémy, jež určují slovní druh a koncovky), jejich pořadí a podobně. Pro nás je důležitá vzniklá struktura, která je značně redundantní, a to na mnoha vzájemně nezávislých úrovních. Jak k této struktuře mozek konkrétně dospěje, nás nyní nezajímá, důležité je, že texty takto skutečně vypadají.

Proč je výhodné, aby se restrikční struktury vytvářely z kategorií, do kterých segmenty řadíme, a nikoli ze samotných segmentů? Počet běžných struktur (ostýchám se napsat „povolených struktur“, ale mnozí mluvčí to tak vnímají) musí být relativně nízký (aby bylo snadné určit, která struktura je „nepovolená“), ovšem z dynamiky jazyka vyplývá, že počet různých segmentů nemůže být omezen. Používání relativně omezeného počtu kategorií znamená, že onen omezený počet povolených struktur vytváříme z typizovaných prvků, jejichž soubor může být konzervativní navzdory rychlému vývoji jazyka. Uživatel jazyka tak není překvapen, že se objeví nový druh bot, a ve chvíli, kdy se poprvé dozví, že existují například *martensky*, může být schopen určit, jestli věta „obula jsem si ty nové martensky“ je vnitřně konzistentní, nebo ne.

Samozřejmě existují funkce, které dokáží sérii prvků o nekonečně mnoha stavech doplnit redundantními prvky tak, aby společně tvořily strukturu, která bude při porušení nekonzistentní,[19] nicméně ještě jednou zopakuji: v klíčový okamžik při vývoji

[18] Například slovo Obama sice můžeme interpretovat jako kořen *Ob* a hovorovou koncovku 7. pádu *-ama*, nicméně kombinace kategorie *jména řeky* a kategorie *plurálu* je neobvyklá, takže pravděpodobnost výskytu tohoto slova v tomto významu není velká.

[19] Například *exkluzivní OR* pro fuzzy množiny, které by se užívalo podobně jako se běžné XOR používá pro dopočítávání paritních bitů.

jazyka bylo zásadní, která funkce se k podobnému typu úloh již využívá a je dostatečně efektivně implementovaná; nejspíš tehdy byly k dispozici schopnosti bleskového zařazení vjemu do kategorie (*ejhle, zajíc… nebo králík, to je teď jedno*) a ohodnocení vztahů mezi těmito kategoriemi a jejich pravděpodobností (*… a běží směrem ke mně; to znamená, že je něco v nepořádku; copak ho asi honí, že se toho bojí víc než mě?*).

Jak tuto strukturu dekódujeme? Opět připomínám, že i tento postup je zde popsán pouze obecně a konkrétní způsob dekódování není tématem této studie. Na každé úrovni jsou segmenty interpretovány, podle interpretace jsou zařazeny do kategorií a následně příjemce zjišťuje, jestli struktury z těchto kategorií jsou povolené, respektive obvyklé, přičemž je možné vzniklou strukturu na vyšší úrovni taktéž kategorizovat a taktéž zkontrolovat, jak dobře zapadá do vyšších struktur. Ona pravidla, jakož i schopnost jejich dodržování u produktora, nejsou stoprocentní, takže proces je hledáním nejpravděpodobnější interpretace a na této pravděpodobnosti (myšleno subjektivní pravděpodobnosti, nikoli frekventistické) závisí pocit porozumění.

Je výhodné, pokud existuje několik na sobě nezávislých kategorizací a několik paralelních pravidel, která je nutné dodržovat. Díky tomu je možné přesněji lokalizovat chybu.

Tato kapitola byla dosud poměrně abstraktní, vraťme se nyní ke konkrétnímu rozboru jazykových struktur, abychom zjistili, že celá tato spekulace není v rozporu s představou hlavního lingvistického proudu ani s tím, jak jazyk vnímají běžní mluvčí.

Máme větu:

Pozítří si obuju martensky.

Vezmeme-li slova jedno po druhém, u každého jsme schopni určit, jestli je to „povolené slovo“ nebo ne, a to i tehdy, pokud jsme ho nikdy neslyšeli — například slovo martensky se skládá z možného kořenového morfému *martens* a přípony *-ky* ve správném pořadí a bez porušení jakýchkoli derivačních pravidel. Zároveň jsme schopní tolerovat docela velké množství variability nejen v prvcích, z jakých se struktury skládají, ale i v samotných strukturách, například nás nevyvede z míry ani jedna z variant *obuju*/*obuji*. Ovšem třeba už varianta *obujám* by nás upozornila na to, že je něco v nepořádku.

Nyní se přesuneme o úroveň výš a slova rozřadíme do různých kategorií, které nás napadnou:

pozítří → Udání času (v budoucnosti).

si

obuju → Sloveso v budoucím čase, modifikované pomocí *si*, vyžaduje doplnění podstatným jménem s akuzativní koncovkou, jehož význam lze kategorizovat jako druh bot.

martensky → Substantivum s koncovkou, která se dá interpretovat jako akuzativní. Hyperonymum *martensek* jsou *boty*.

Tyto kategorie spolu utvářejí tři na sobě nezávislé struktury (každou z nich značím jinou barvou):

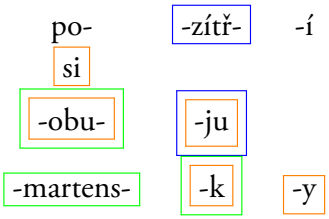
pozítří → Udání času (v budoucnosti).

si

obuju → Sloveso v budoucím čase, modifikované pomocí *si* vyžaduje doplnění podstatným jménem s akuzativní koncovkou, jehož význam lze kategorizovat jako druh bot.

martensky → Substantivum s koncovkou, která se dá interpretovat jako akuzativní. Hyperonymum *martensek* jsou *boty*.

A zde se dostáváme ke zlatému hřebu této kapitoly: tento dvojí způsob zachovávání konzistence (na úrovni slov a na úrovni vět) je redundantní ve dvou rozměrech. Struktury na úrovni slov spolu pojí morfémy, které bezprostředně sousedí, a v rámci terminologie, kterou jsme si zavedli v kapitole 5.2.2, se jedná o horizontální redundanci. Struktury, které spolu pojí kategorie napříč jednotlivými slovy v rámci jedné věty, jsou poněkud nesystematickou implementací redundance vertikální (transverzální). Následující tabulka toto tvrzení ilustruje na jednotlivých morfémech (opět morfémy podílející se na určité struktuře označujeme rámečkem stejné barvy):



Takové rozdělení do struktur je velmi hrubé a poněkud zavádějící, to je ostatně důvod, proč raději postulujeme rozřazování do kategorií, ovšem není možné zapřít, že hlavní podíl na zařazení do zmíněných kategorií nemají všechny morfémy stejný.

Arabština je v tomto češtině velmi podobná, vzhledem k tomu, že se také jedná o jazyk s komplexní morfologií. Na arabském příkladu si ovšem ukážeme, jak v tomtéž duchu fungují i jiné než syntaktické struktury. Mějme větu:

Bi-rūḥ bi-dam nafdīka yā Ṣaddām!

بِرُوح بِدَم نَفدِيكَ يَا صَدَّام!

Svou duší, svou krví obětujeme se Ti, Saddáme![20]

Opět můžeme pozorovat dodržování struktur na úrovni slov (slova jsou utvořena z existujících morfémů ve validních kombinacích) a na úrovni věty (existující slova jsou složena podle validních syntaktických pravidel). Ovšem kromě struktur popsaných na českém příkladě můžeme pozorovat další, paralelní strukturu: věta je totiž typizovanou formulí a jen některé její prvky jsou variabilní. Místo jména Ṣaddām může být jméno jiného diktátora, vůdce nebo státní entity, pokud to bude cokoli jiného, bude formule interpretována jako parodie. *Duch* a *krev* jsou také zamrzlé a jediná variabilita, kterou najdeme, se týká určenosti (užívá se i *bi-'r-rūḥ bi-'d-dam* a *bi-rūḥ wa-dam*). Variace mají tendenci dodržovat stejné metrum a rým, přestože se o poesii nejedná. Takovéto formule s variabilními složkami nejsou v jazyce okrajovým jevem, právě naopak, setkáváme se s nimi každý den při pozdravu a nejen při něm.

Jako jednoduchý příklad využití popsaného kódování při opravě chyb si představme větu, kterou jsme viděli s několika překlepy: *Zamůlovala jsem se do svých kozuček.* Pokud jste rodilými mluvčími češtiny, tak vás nejspíš napadlo několik způsobů, jak větu pozměnit tak, aby byla konzistentní. Napřed se ale zkusme zamyslet nad tím, jak jste vlastně poznali, že je tyto operace třeba provést. Na úrovni hlásek je věta naprosto v pořádku. Na morfematické úrovni již nalezneme jednu nekonzistenci: morf *-můl-*, který není zrovna běžný, ostatní morfy jsou v pořádku — *kozuček* můžeme analyzovat jako validní morfy *-koz-* (kořenový) *-uč-* (také kořenový, například *uč-it*), přípona *-ek*. Druhá chyba se projeví teprve na slovní úrovni, neboť pokud jsme si morf *-koz-* i morf *-uč-* zařadili do kategorie kořenových morfů, tak slovo *koz-uč-ek* porušuje pravidlo, že v jednom slově nesmí být dva kořenové morfy těsně vedle sebe (které platí pro drtivou většinu českých slov, s čestnými výjimkami). Díky segmentaci se nám podařilo chyby přesně lokalizovat.

Nyní přichází složitější část: oprava chyb. Pokusíme se udělat co nejmenší změny v morfému *-můl-*, aby nám vznikl validní morfém, který by konzistentně zapadal do struktury slova (*za-?-ova-l-a*). Už změnou jednoho distinktivního rysu získáme jeden takový morf (*-mul-*), a pokud povolíme změnu na úrovni hlásek, najdeme jich několik (*-hůl-*, *-kůl-*, *-mol-*, *-mal-*, *-mil-*, *-můr-*…). Už na úrovni slov jsme schopní

[20] Uvedené heslo je v různých variantách velmi oblíbené po celém arabském světě, neboť se dobře skanduje na demonstracích. Tato varianta vyjadřuje oddanost Ṣaddāmu Ḥusaynovi, ovšem místo něj si můžeme představit jakéhokoli jiného diktátora, ideového vůdce nebo státní entitu, obzvláště pokud záchovává rým a metrum (tedy má dvě slabiky a končí na *-ām* nebo alespoň *-ān* nebo *-an*), jako například *Salmān*, *Yaman*, *waṭan* (vlast), ovšem najdeme užití i se jmény diktátorů, kteří do poetické struktury zoufale nezapadají (Baššār, nebo dokonce Qaḍḍāfī…) . Pozoruhodné je, že krev a duch jsou v jednotném čísle, jedná se tedy o obětování jakéhosi kolektivního ducha a společné krve; přitom tradiční verze této formule, dochovaná v různých variantách v historických pojednáních a ve sbírkách ḥadīṯů, se naopak pojí s individuálními dušemi, majetky a podobně (například al-Wāqidího *Kitāb al-maġāzī* z 9. století n. l. udává variantu: *Layta ʾinnā nafdīka bi-'l-ʾanfusi wa-'l-ʾawlādi wa-'l-ʾamwāl*, čili *věru obětujeme se ti svými dušemi a dětmi a majetky*).

rozhodnout, které řešení je pravděpodobnější, neboť slova jako *zamulovat* nebo *zamůrovat* jsou sice kombinatoricky možná, nicméně se s nimi člověk nesetkává každý den. Ovšem jak slovo *zamalovat*, tak *zamilovat* jsou obě normální česká často používaná slova. Díky vnořenému segmentování můžeme rozhodnout, které je to pravé, jen je třeba postoupit o úroveň výš: na větné úrovni je vazba *zamalovat se do někoho/něčeho* používaná velmi zřídka, na rozdíl od vazby *zamilovat se do někoho/něčeho*. První chybu prozatím považujme za vyřešenou (píši prozatím, neboť při odstraňování druhé chyby by se mohlo ukázat, že toto řešení je nepravděpodobné).

U slova *kozuček* je situace složitější, neboť se nám chybu nepodařilo izolovat tak přesně, změny tedy mohou být v rámci celého slova, takže připadají v úvahu slova jako *kouček*, *kozáček*, *kozaček*, *koziček*… Opět ovšem platí, že slovo musí zapadat do struktur nad ním, takže můžeme vytřídit slovo *kozáček*, které není v genitivu vyžadovaném předložkou *do*. Ostatní uvedená slova však v užším výběru zůstanou a musíme mezi nimi rozhodnout opět podle širšího kontextu. Ideální je, pokud se poblíž najde věta, která je pleonastická s uvedenou větou (například může následovat věta: „Občas si je obuju jen tak, i když nikam nejdu," kde slovo *obout* vyžaduje nějakou kategorii bot). Pokud takovou větu nenajdeme, přenosový protokol selhává a my se musíme rozhodovat podle smyslu (respektive redundance dané věty vůči nám známým informacím). Třeba pokud následuje věta *Občas si je hladívám*, pak vám nezbude, než zvažovat pravděpodobnost jednotlivých možností na základě nejazykových vědomostí — jestli je mluvčí dost divná na to, aby si hladila kozačky, jestli by se koučky nechaly pohladit a koho tím vlastně myslí, jestli kozičky nejsou moc plaché a kde je vůbec chová a jestli kozičky nemyslí ve významu květenství černého bezu a jestli není hlazení těchto květenství divnější než hlazení kozaček. Přitom si nikdy nemůžete být jisti, že znáte všechny relevantní informace, například že kozačky mluvčí mají kožešinový lem a mluvčí ráda hladí chlupatá zvířata a jejich kožíšky.

Tím rozhodně nechci říct, že by opravování chyby a desambiguace podle smyslu byla málo častá, naopak, zejména v případě, kdy se nemůžeme spolehnout na přesnost přenosového kódu (například při komunikaci s cizincem) nebo ho sami známe málo (při učení se nového jazyka), na pravděpodobnost smyslu prakticky spoléháme. Což mi potvrdíte, pokud jste se někdy pokoušeli luštit například moderní poesii z jazyka, který jste nedostatečně ovládali — předávanou informací může být prakticky cokoli a není možné se spolehnout na to, že bude nějak korespondovat se světem, jak ho znáte, pročež čtení experimentální poesie je mnohem obtížnější než řekněme společenského románu.

Otázka, jak odstraňování šumu při dekódování probíhá ve skutečnosti, je asi mnohem závažnější než v případě enkódování, neboť tato procedura je už z podstaty komputačně náročná, zvláště v případě mnohočetného poškození na malém úseku. Předpokládám, že zásadní úlohu hraje priming, respektive to, že příjemce může očekávat,

že jeho asociace budou podobné asociacím produktora textu, čili může si vytvořit jakési hypotézy ohledně pokračování sdělované zprávy a skutečný text používat pro potvrzování/vyvracení těchto hypotéz. Navíc běžní lidé disponují docela velkou schopností vcítění se (což je asi také předjazyková schopnost).

Nejsme agnostičtí vůči zašuměnému kanálu (na rozdíl od pokusů v předchozí kapitole), a to jak při enkódování, tak dekódování. Při dekódování dokážeme odhadnout, co dělá kanál — například na šišlání nebo jinou řečovou vadu či mírně odlišný jazykový kód si zvykneme. Také při enkódování se vlastnostem kanálu přizpůsobujeme: tím, že porušujeme pravidla, čímž snižujeme redundanci, pokud není potřeba (o tom v kapitole 6), naopak zvyšování redundance pomocí pravidel od určité meze už nejde, což se řeší vkládáním prostých pleonasmů, ať už jednoduchým opakováním informací, nebo vkládáním úseků textu, které obsahují ztrátově komprimované informace uvedené i jinde v textu (tato strategie se typicky používá v úvodních a závěrečných segmentech na vyšších hladinách, například na úrovni kapitol nebo delšího vyprávění).

Logicky nic nebrání vzniku systému, ve kterém segmenty nejsou vůbec formalizovaně ohraničené (ostatně tak tomu bylo v systémech představených v kapitole 5.2.2 — lze toho docílit jednoduše tak, že vlastní informaci nasegmentujeme do stejně dlouhých úseků), nicméně často potřebujeme překonat vymazávací kanál (pro připomenutí viz obrázek 5.2) — k čemuž se hodí různé markery konců segmentů, ať již jednoznačné, nebo fuzzy. Formalizovaně ohraničená segmentace je výhodná také v dialogických systémech, kdy se příjemce může zeptat přímo na konkrétní segment, kterému neporozuměl.

Na tyto markery a na to, co z toho plyne, se podrobněji podíváme v kapitole 7. Nyní nás ovšem čeká kapitola rozebírající konkrétní způsoby vkládání redundance na jednotlivých úrovních.

Kapitola 6

Příklady vkládání redundance na různých úrovních

V jazyce existuje mnoho strategií, jak redundanci vložit a druhů redundance, které vloženy být mohou. V kapitole 5.3 jsme si ukázali, že jako redundance mohou sloužit různá omezení a pleonasmy.

Informace, které nám redundance pomáhá přenést, můžeme rozdělit do čtyř kategorií:[1]

1. Symboly pro kategorie nebo instance kategorií, jak je máme konceptualizovány.
2. Významové interakce mezi nimi.
3. Deiktické vztahy mezi symboly.
4. Markery pro hranice segmentů (aby byli příjemci schopni alespoň přibližně určit hranice slov, vět a podobně, jako je to vyznačeno na obrázku 4.1).

[1] Toto dělení začne dávat větší smysl, když ho spojíme s představou jakési anarchistické syntaxe, s níž jsem vás seznámil už na obrázcích 3.1–4.3. Představte si, že syntaktické vazby značí významové interakce mezi prvky, přičemž není nutně určeno, který závisí na kterém, ani nejsou žádná umělá ideologicky daná omezení těchto vazeb. Takže graf znázorňující tyto vazby nemusí být strom (a obvykle ani nebude) a vazby mohou jít přes hranici věty či jakkoli velkého celku, neboť segmentace je chápána až jako sekundární. Dále do grafu mohou být zaznamenávány vztahy deiktické a segmentace na slova, věty atd. nebo jakékoli jiné vyšší segmenty, které anotátor uzná za vhodné. Jak segmentace na nejnižší prvky, tak vazby mezi nimi nejsou určené žádným tisícistránkovým anotačním manuálem, ale volně si je určuje sám anotátor podle vlastní představy o jazyce, v ideálním případě může každé vazbě přiřadit nějakou subjektivně vnímanou váhu. Je samozřejmé, že anotátorů pak musí být pro každý text několik, třeba i několik stovek, a právě rozdíly a shody mezi nimi mohou být to nejzajímavější. Takto pojatý výzkum syntaxe navíc plně vyhovuje „fyzikalistickému“ přístupu k jazyku, o kterém jsme pojednali v kapitole 1.6.

K této vizi jsem došel při neúspěšném pokusu využít PDT k lingvistickým účelům, k testování hypotézy týkající se syntaxe, neboť jsem zjistil, že skrze data nezkoumám jazyk ani představu mluvčích o něm, ale manuál, podle kterého byla data oanotována, a schopnost anotátorů ho dodržovat.

Všechny tyto informace mohou být vyjádřeny buď segmenty, které jsme si definovali v kapitole 3.1.2 (vlastně morfémy), nebo předem danými schématy (slovosled a podobně).

Následující seznam způsobů, které se v jazyce k redundování používají na jednotlivých úrovních, je nutně neúplný a má sloužit k tomu, aby si čtenář udělal základní představu, jak si konkretizovat modely, o nichž jsem pojednával v předchozí kapitole.

6.1 Úroveň distinktivních rysů

Samotný hláskový inventář obvykle bývá systémově redundantní, tedy hlásku od libovolné jiné hlásky dělí více než jeden distinktivní rys, například české krátké a dlouhé *i* (tak jak je vyslovováno v Čechách) se liší jak v délce, tak na škále vyšší — nižší a přední — zadní.

Taková redundance může být při produkci textu regulována nedbalou výslovností. Ostatně nedbalou výslovností může být regulována systémová redundance i na vyšších úrovních, při běžné komunikaci není problém polknout různě velkou skupinu hlásek, někdy celý morf nebo i slovo. K těmto redukcím doporučuji Machač – Zíková (2013), popřípadě Johnson (2004), která přímo pracuje s teorií informace.

Na otázku „a chcete ty hranolky s kečupem, nebo s tatarkou?“ je očekávána odpověď s komplexitou o velikosti jednoho bitu. Nicméně variabilita jazykového kódu nám nedovoluje měnit velikost morfémů, nemůžeme tedy poslat cokoli menšího než několik desítek bitů („s tatarkou!“), což není právě efektivní, tedy alespoň zredukujeme několik distinktivních rysů a vyslovíme něco jako sə:kó͡u.

Distinktivní rysy jsou tedy v reálné komunikaci do značné míry nezávislé na hranici hlásek, třeba taková nazalita v češtině se nemusí držet hlásky, ke které patří (Machač – Zíková, 2013, str. 19), také emfatičnost v arabštině může být suprasegmentální (Davis, 2006–2009, str. 637), nicméně pokud potřebujeme protlačit informaci přes předpokládaný velmi zašuměný kanál, můžeme využít pečlivé artikulace a toho, že některé kombinace distinktivních rysů v rámci jedné hlásky nejsou povoleny, třeba v inventáři arabštiny nenajdeme neznělou bilabiálu.

6.2 Úroveň hlásek

Mluvčí může regulovat míru redundance už na úrovni hlásek. Třeba tím, že uplatní jednoduchou modulární redundanci, jak jsme si ji ukázali v kapitole 1.4, ať už tak, že zvedne hlas, nebo prodlouží interval realizace (v komiksovém zápisu: *nneeeesssssllyyšš-šššššťťťťťťťťmmm*).

Redundanci také způsobují omezení, která různé jazyky kladou na hláskové kombinace. Můžeme to ilustrovat na tom, že arabské slovo (ve spisovné arabštině) nesmí začínat na souhláskový klastr. Kromě toho, že toto pravidlo pomáhá segmentaci na

slova při percepci mluvené řeči, zvyšuje i redundanci, neboť pokud slyšíte souhláskový klastr na začátku promluvy, máte solidní jistotu, že jste opravdový začátek přeslechli nebo že se nejedná o spisovnou arabštinu.

Podobně vokalická harmonie v turečtině může pomoci segmentovat na slova, ale v mnoha případech je redundantní.

V arabštině je inventář kořenů omezen pravidlem, že se vedle sebe nesmí vyskytovat dvě podobné hlásky (tzv. *obligatory contour principle*, více viz Rosenthall (2006–2009)), což v šablonách, kde se tyto dvě hlásky dostávají do sousedství, zvyšuje zřetelnost a zabraňuje jejich splynutí, ale v šablonách, kde jsou tyto hlásky odděleny samohláskou, je redundantní.

6.3 Úroveň morfémů

Redundance na úrovni morfémů je dána už tím, že se jedná o nejmenší jednotky, které reprezentují nějaký významový celek (kategorii, systém, ostatně takto jsme si je definovali), a tím musí být jejich informační velikost stabilní, navzdory proměnlivému kontextu, který implikuje různou komplexitu. Například pokud se mě nad šachovým stolkem zeptáte, jestli chci černé, nebo bílé, pak moje odpověď „bílé“ má komplexitu jeden bit, zatímco když se mě zeptáte nad čtyřmi lahvemi vína, jestli chci bílé, růžové, červené nebo sekt, pak ta samá odpověď bude mít komplexitu dva bity. Avšak délka obou odpovědí je stejná a lišit se ani nemůže, neboť morfém, aby mohl fungovat v komunikaci, musí zůstat stejný. Jak bylo řečeno, jeden způsob, jak množství této přirozeně se vyskytující redundance regulovat, je redukce distinktivních rysů, což ovšem třeba v psaném textu není vůbec možné; další možností je pak (pokud to jde) nahradit morfém morfémem v daném kontextu synonymním, který má jinou informační délku.

Podobně jako hlásky, i morfém je možné kombinovat s určitými omezeními, která mohou být v celkovém kontextu redundantní, například arabská slova mohou mít jenom jeden kořenový morfém (výjimky ovšem existují, viz Bielický, 2007, kapitola 3.3), afixy mají pevnou pozici — českou příponu *-ší* není možné použít jako předponu, přestože žádná homonymní předpona neexistuje.

Morfologická úroveň je nejnižší úrovní, kde má smysl mluvit o pleonasmech, přičemž je zajímavé, že pleonastické morfémy jsou (přinejmenším v češtině a arabštině) obvykle obligatorní, zatímco pleonasmus na slovní úrovni je volnější a mluvčí s ním může nakládat takřka bez omezení. Například česká i arabská osobní zájmena v roli podmětu, která jsou dublována morfologickými markery u sloves, mohou být vynechána,[2] zatímco vynechání oněch morfologických markerů by bylo vnímáno jako agramatické.

[2]Například věta *Já už jdu* může být bez velkých významových změn nahrazena větou *Už jdu*. Podobně alternuje konstrukce *obchod se smíšeným zbožím* s konstrukcí *obchod smíšeným zbožím*.

Pravidlo, že ze dvou redundantních prvků je obligatorní ten, který je na nižší úrovni, ovšem neplatí za všech okolností. Například v arabštině může být vynechán redundantní marker ženského rodu, jak dokládá citát ze slovníku Tahḏīb I, str. 170, od al-'Azharīho, citováno podle (Versteegh, 1997, str. 22):

> 'Abû 'Ubayd said: "Imra'a 'âshiq 'a woman in love', without the feminine ending -a, and likewise rajul 'âshiq 'a man in love'". I say: The Arabs delete the feminine ending from the feminine attribute in many words, e.g. [in the expression] "you regard her as stupid, since she is bâkhis 'deficient'". They also say imra'a bâligh "a nubile woman" when she has reached puberty, and they call a female slave khâdim "servant". In these words the masculine form is the same.

Dodávám, že něco podobného má i čeština (třeba slovem *miláček* můžeme označovat i osobu ženského pohlaví). Vynechání redundantního ženského rodu je v arabštině u některých slov naopak povinné — u pojmenování osob, které musí být už z definice ženského pohlaví, například *ḥāmil* (těhotná), *murḍi'* (kojící) nebo *bikr* (žena, jež ještě nerodila).

Morfémy mohou být vzájemně redundantní s omezeními na vyšší úrovni, například ustálený slovosled arabské věty, který v moderní spisovné arabštině (v drtivé většině vět) jednoznačně určuje, co je podmět a co předmět, je dublován soustavou koncovek, které mají stejnou úlohu (a které jsou v běžné řeči vynechávány).[3]

Podobně v češtině předložky musí stát bezprostředně před jmény, ke kterým se váží, ovšem stejně jsou na ona jména navěšeny markery přináležitosti s těmito předložkami (předložkové pády).

Jazyky se samozřejmě liší v množství a typu povinné redundance. Například v arabštině musí být marker určenosti (*al-*) navěšen jak na podstatném jméně, tak na přídavném jméně, které ho rozvíjí (*wa-ḫaraǧat al-bint al-ǧamīla* = a vyšla krásná dívka), zatímco v germánských jazycích stačí uvést slovo mající funkci tohoto markeru pouze jednou *(and the beautiful girl went out)*. V češtině a jiných indoevropských jazycích je marker množství připojen jak ke jménu — podmětu, tak ke slovesu — přísudku (*dívky běžely*), zatímco v arabštině je množství označeno pouze na jméně (*ǧarat al-banāt*, platí pouze pro slovosled VSO).

[3] Corriente (1971) tvrdí, že koncovky podstatných jmen (*i'rāb*) nejsou redundantní v 10 % ve staroarabské poesii a v 0 % v Koránu a taktéž v zanedbatelném procentu případů v pozdějších textech (str. 36). Situaci srovnává s ruštinou, kde je redundantních okolo 70 % koncovek (str. 46). Dávám k dispozici 95% binomické konfidenční intervaly, které jsou pro ruštinu 22,1 % – 36 %, pro arabskou poesii 7,8 % – 13,2 % a pro Korán 0 % – 1,8 %, které Corriente neuvádí, ale bylo možno je dopočítat.

6.4 Úroveň slov

Od této úrovně dál není úplně možné mezijazykové srovnání, neboť to, co mluvčí jazyka označují za slovo, je značně variabilní. Pro analytickou angličtinu je tato úroveň podobná úrovni morfematické, zatímco v syntetických jazycích se podobá spíše té větné.

Podobně jako na morfematické úrovni, i zde funguje pleonasmus, ostatně učebnicové příklady pleonasmu jako *dárek zdarma* nebo *modrý blankyt* jsou právě na úrovni slov.

Také zde může být pleonasmus přidáván systematicky a povinně, například ve větě *Vešel do domu* je sloveso *vejít* s povinnou předložkou *do* vzájemně pleonastické. A také zde se jazyky liší: v angličtině je podobné vyjádření (*he entered into the house*) vnímáno jako stylistická neobratnost a v arabštině by byla tato konstrukce (*daẖala fi ’l-bayt*) vnímána jako agramatická. Obligatornost této vazby v češtině má racionální vysvětlení, sloveso *vejít* se často používá v kontextu, ve kterém je možné ho nahradit slovesem *vyjít*,[4] přičemž požadovaná informace se často týká právě distinkce směru dovnitř/ven, která je u této dvojice vyjádřená pouze jednou hláskou, jejíž realizace ještě k tomu variuje mezi dialekty, takže když pražský mluvčí řekne *vyšla,* je to na východní Moravě interpretováno jako *vešla*.

I zde samozřejmě funguje omezení slovosledu. Zajímavý způsob je pak vytvoření jakéhosi schématu (syntaktického, zvukového, významového…), které následně dodržujeme i v následující větě. Vezměme si například následující arabské přísloví:

شَيئَانِ لَا تُصَدِّقهُمَا أَبَدًا: دُمُوعُ النِسَاء وَقُلُوبُ الرجَال

وَشَيئَانِ لَا تُكَذِّبهُمَا أَبَدًا: دُمُوعُ الرجَال وَقُلُوبُ النِسَاء

> Šaj’āni lā tuṣaddiqhumā ’abadan: dumūʿu ’n-nisā’ wa-qulūbu ’r-riğāl;
>
> wa-šaj’āni lā tukaḏḏibhumā ’abadan: dumūʿu ’r-riğāl wa-qulūbu ’n-nisā’.
>
> Dvěma věcem nikdy nevěř: slzám žen a srdcím mužů;
>
> a dvě věci nikdy nezpochybňuj: slzy mužů a srdce žen.

V arabském originále vidíme, že první část vět (před dvojtečkou) se liší jen v kořeni jednoho slova, jinak je vše zcela stejné, včetně šablony slovesa a jeho rekce (na rozdíl od českého překladu).[5] Taktéž šablona slova pro *slzy* (*dumūʿ*) odpovídá šabloně pro

[4] Příkladem ze života budiž situace, kdy vám manželka telefonuje, že „už vyšla ven“ místo pouhého „už jsem vyšla“, které by nemuselo projít přes zašuměný kanál GSM.

[5] Stručné uvedení do arabské morfologie naleznete v poznámce 8 na straně 36.

srdce (*qulūb*), podobně šablona pro *ženy* (*an-nisā'*) odpovídá šabloně pro *muže* (*arriǧāl*). Kromě toho je první a druhý řádek spojen chiasticky, jak je patrné i z českého překladu.

Podobné šablony nemusí být ustálené jen v rámci jednoho textu, ale mohou tvořit jakési očekávané formule uplatnitelné kdekoli jinde. Například struktura výše uvedeného přísloví je využita i v jiném arabském přísloví:

شَيئَان لَا تَثِق بِهِمَا: شَمس الشِتَاء وَقَلبُ النِسَاء.

Šaj'āni lā taṯiq bihimā: šamsu 'š-šitā' wa-qalbu 'n-nisā';

Dvěma věcem nikdy nevěř: zimnímu slunci a srdci ženy.

Kubát (2010, str. 24) nazývá podobná schémata strukturními formulemi a dokládá, že je na nich založená celá staroarabská poesie, tedy že se nejedná o nějaký výstřelek. Tato figura byla používána proto, že zjednodušovala básnickou improvizaci, a navíc měla funkci estetickou. Estetice redundance bude věnována celá podkapitola 6.9.

Ještě jeden druh redundance najdeme v uvedených příslovích: je jím jinak poměrně vzácné udání počtu prvků, které chceme vyjmenovat; kdyby pak mluvčí zmínil jen jeden, příjemce bude vědět, že je seznam neúplný.

Podobnou konstrukci najdeme i v jiných příslovích, například ve Starém zákoně (Př 30 podle Bible Kralické):

> 18. Tři tyto věci skryty jsou přede mnou, nýbrž čtyry, kterýchž neznám:
>
> 19. Cesty orlice v povětří, cesty hada na skále, cesty lodí u prostřed moře, a cesty muže při panně.

Tento druh redundance se obvykle přidává tam, kde se předpokládá ústní šíření přes mnoho lidí a uložení v paměti. Asi nejslavnějším příkladem je Desatero. Dá se očekávat, že tato strategie je účinná pro překonávání vymazávacího dávkového kanálu (obr. 5.2).

6.5 Úroveň vět a souvětí

Podobně jako slova, i věty mohou být pleonastické. Trochu škodolibě si pojďme rozebrat slavný citát o pleonasmech z učebnice slohu z počátku století (Strunk, 1918, kapitola 2, pravidlo 18, číslování vět doplněno JM):

> [1] Vigorous writing is concise. [2] A sentence should contain no unnecessary words, [3] a paragraph no unnecessary sentences, [4] for the same reason [5] that a drawing should have no unnecessary lines [6] and a machine no unnecessary parts.[6] [7] This requires not that the writer

[6] Ironií je, že srovnáním jazyka a stroje o 64 let později Campbell (1982, str. 73) vysvětloval naopak nutnost redundance v jazyce.

> make all his sentences short, [8] or that he avoid all detail [9] and treat his subjects only in outline, [10] but that every word tell.

Význam první věty se částečně kryje s významem věty druhé a třetí. Věty 2 a 3 sdílejí obdobnou strukturu. Taktéž věty 5 a 6 mají vzájemně podobnou strukturu. Věty 8 a 9 jsou vzájemně redundantní, stejně tak věty 2 a 10. Poznamenejme, že i přes vysokou celkovou míru redundance odstavec není vnímán čtenářem jako neestetický a dokud není podrobně rozebrán, tak si čtenář pleonasmů nejspíš ani nevšimne. Pleonasmus tedy sám o sobě není neestetický, teprve až jeho neefektivní užití.

Podobně jako v předchozí podkapitole, i zde můžeme citovat z biblických přísloví konstrukci, kdy je udán počet souvětí, která budou následovat (Př 30 podle Bible Kralické):

> 24. Čtyry tyto věci jsou malé na zemi, a však jsou moudřejší nad mudrce:
> 25. Mravenci, lid nesilný, kteříž však připravují v létě pokrm svůj;
> 26. Králíkové, lid nesilný, kteříž však stavějí v skále dům svůj;
> 27. Krále nemají kobylky, a však vycházejí po houfích všecky;
> 28. Pavouk rukama dělá, a bývá na palácích královských.

6.6 Úroveň odstavců a kapitol

Odstavec je segment, který je typický pro psaný text, v mluvené řeči mu odpovídají nejspíše jednotlivé repliky v dialogu, v monologu též najdeme dělení na úseky delší než věta, při moderní prezentaci komentář jednoho slidu, v poesii je ekvivalentem strofa, sloka.

Kapitola je útvar též typický pro psaný text, nicméně s podobnou úrovní se setkáme i v mluvené řeči (například jednotlivé příběhy Tisíce a jedné noci). Typickými redundantními kapitolami, díky kterým čtenář pozná, jestli porozuměl zbytku textu nebo nikoli, jsou v odborné literatuře tradičně úvod a závěr, v krásné literatuře pak žádné typizované řešení neexistuje. Úvod a závěr mají také jasně omezenou pozici v rámci uceleného textu. Podobně najdeme i redundantní úvodní a závěrečné odstavce v jednotlivých kapitolách. V této práci se často setkáte s další obvyklou strategií, totiž že poslední odstavec kapitoly je redundantní vůči kapitole následující (například v kapitole 3.2).

Poziční redundance funguje i na této úrovni, je mnoho žánrů, které mají pevně danou strukturu — typické motivy, restrikce jejich sledu a podobně. Příkladem může

být antická tragedie a arabská qaṣīda,[7] ovšem je třeba si přiznat, že literární teoretici tyto struktury přeceňují a obzvlášť původní literární památky tohoto žánru strukturu, která se jim přisuzuje, příliš nedodržovaly.

6.7 Úroveň ucelených textů

Intertextualita je přirozenou součástí jak krásné literatury, tak našich běžných promluv, pročež i „ucelené texty“ mohou být redundantní. Existují i ucelené texty, které jsou dokonale pleonastické vůči jinému ucelenému textu (školní referáty a zápisky do čtenářského deníku atd.), také se ovšem setkáme s pleonasmem částečným, jako jsou komentáře, výklady, popisy téže události různýma očima (u nás asi nejznámějším příkladem jsou evangelia a Čapkův Hordubal).

6.8 Redundance specifická pro psaný text

Psaný text přirozeně vyžaduje jinou úroveň redundance i jiné strategie jejího vkládání než text mluvený, nicméně vzhledem k obvyklé korelaci obou kódů rozdíl není nijak zásadní, zvláště u slabičných a hláskových písemných systémů, které jsou odvislé od zvukové stránky jazyka.

Na druhou stranu, v kulturách s dlouhou tradicí písemného záznamu se vyvinul specifický literární styl, který se od mluveného více či méně liší, a rozdíly konvenují rozdílům v možnostech produkce i zpracování textu. Kromě toho znakové písemné systémy (egyptské hieroglyfy, klínopis, čínské znaky) vkládají do textu redundantní determinativy, které umožňují lepší desambiguaci.

Text psaný hláskovým písmem vyžaduje zřejmě menší množství redundance než text mluvený, což se ukazuje například na tom, že semitské písemné kódy si už tisíce let vystačí bez krátkých samohlásek, což je docela výrazná ztrátová komprese.

Abychom byli důslední, je třeba zmínit, že i hláskové písemné kódy dávají mnohdy přednost jednoznačnému zápisu, například v angličtině nebo francouzštině se historické rozdíly mezi slovy uchovávají, i když v mluveném jazyce splynou do homofonie (což ovšem může souviset spíš s obecně konzervativním pravopisem). V arabském písmu najdeme zcela jednoznačný marker gramatického ženského rodu u jmen (tzv. *tā’ marbūṭa*). Také syntaktická závislost na slovese se u jmen v písmu označuje (tzv. *ochranný alif*), zatímco v mluveném kódu se koncovka, která má stejnou úlohu, běžně vypouští.

[7] Útvar charakteristický pro klasickou arabskou poesii, který se (zjednodušeně řečeno) podle arabských literárních teoretiků skládá ze tří částí — nasību (básník vzpomíná na svou milou, lká nad opuštěným tábořištěm…), raḥīlu (básník cestuje a popisuje přírodní scenérii a své jízdní zvíře) a faḫru (básník chválí toho, jemuž je báseň určena).

Zajímavý způsob přidávání redundance, který se vyvinul speciálně pro potřeby delších textů zapsaných na nespojitém materiálu (knihy), je číslování stran, které nám pomáhá zjistit, jestli nějaká stránka nechybí. Dřív než se strany začaly číslovat, používala se jak v Evropě, tak na Blízkém východě jiná strategie, spočívající v uvedení prvního slova na stránce na okraji stránky předchozí a posledního slova na stránce na okraji stránky následující.

6.9 Redundance specifická pro poesii

Každé omezení, které poetická forma klade na autora, vnáší do textu redundanci. Metrum, rýmy, paralelismus membrorum, zvukový paralelismus, chiasmus, refrén, aliterace, složitější struktury na úrovni strof nebo celých básní, jako je třeba forma akrostichu, jsou efektivním způsobem vkládání redundance na více úrovních.

Pro poesii je typická další paralelní segmentace (stopy, verše, sloky…), která sice obvykle respektuje segmentaci slovní a větnou, ne však nutně nebo důsledně. Tato segmentace tvoří další vrstvu pro vkládání redundance, jakousi druhou vrstvu přenosového protokolu, která je do určité míry nezávislá na přenášených informacích.

Během dlouhých stovek tisíc let mezi tím, kdy se objevil jazyk a kdy písmo, sloužila lidská paměť a ústní tradice jako zašuměný kanál pro přenos informací na velké časoprostorové vzdálenosti. To je dost času na to, aby se vyvinuly specifické mechanismy, jak informaci obalit redundancí tak, aby tímto kanálem prošla co nejméně poškozená. Jedním z těchto mechanismů byla poesie. Ze studia orálních společností, které zachytila historie,[8] si můžeme utvořit představu o tom, jak velkou roli básník ve společnosti zastával. Ať už to byl básník kmenový, nebo potulný, oba zajišťovali kontinuitu základních představ o světě, etických norem a historie, která určovala mezikmenové vztahy, přičemž schopnost vyznat se v nich a správně určit, kdo je nepřítel, kdo spojenec a kdo co komu dluží, byla záležitostí života a smrti.

Poetické formy oplývající chytře vloženou redundancí považujeme obvykle za estetické, což je přirozený důsledek toho, že efektivně plní svou úlohu při přenosu informací — můžeme předpokládat, že komu se líbila poesie, měl větší šanci na přežití. Podobně došlo k estetizaci zpěvu ptáků, šumění listí, zurčení potoka a vůně ovoce a květů, které znamenají, že jsme v prostředí příhodném pro dlouhodobé setrvání, naopak nepříjemný je zvuk komára a pach hniloby. Tato estetizace je možná napevno zakódovaná v lidském genomu, čemuž by napovídalo, že vzájemně podobné poetické formy se používají v naprosto rozdílných kulturách.

Z tohoto pohledu by milostná poesie byla vlastně až důsledkem estetizace poetické formy, která proběhla primárně kvůli spolehlivějšímu šíření informací. Na opěvovaný protějšek působila jednak estetika formy, napevno zadrátovaná v mozku, jednak

[8] Jako třeba společnost předislámské Arábie, ale i v podstatě moderní ruský nebo srbský venkov, kterémuž tématu se věnuje například Foley (2002).

inteligence manifestovaná schopností tuto formu dodržet a potenciální společenské postavení touto schopností dosažené, nebo alespoň dosažitelné. To mohlo dále podpořit pohlavní výběr kombinace alel, které skládání textů opatřených metrem, rýmem a podobně umožňovaly.

Samozřejmě se nesnažím vyvrátit literárněvědnou představu, že samotná forma může nést význam, obzvlášť je-li porušena. Zejména s rozšířením písma se úloha poetické formy jakožto přenosového protokolu se zvýšenou redundancí vytrácí, nicméně její estetické působení si můžeme užívat dodnes.

Tuto kapitolu chápejte jako spekulativní literárněvědný odpočinek.

6.10 Estetika komplexity a redundance

Předchozí kapitola nás přivádí k otázce estetiky obecně. Je příznačné, že méně efektivní druhy redundance jsou považovány za neestetické, typicky modulární redundance bezprostředně se opakujících prvků (pleonasmus v původním slova smyslu, například *notoricky známá drogová narkomafie*), která je náchylná k chybám v dávkově zašuměném kanále (*burst-noise channel*). Za nevhodný je pleonasmus označován obzvláště tehdy, když je jeho užití způsobeno neznalostí (*LED dioda*, *SEO optimalizace*). Naopak jiné druhy redundance jsou vítány, zejména ty, které jsou tvořeny dodržováním restrikcí (gramatická a pravopisná „bezchybnost“; metrum, rým a jiné druhy paralelismu).

Zajímavé je podívat se z pozice teorie informace na estetiku vůbec. Za esteticky přitažlivou je pokládána vysoká míra komprimovatelnosti, a to ve všech druzích umění, což důkladněji zkoumal Jürgen Schmidhuber (2007, 1997b). Ten toto pozorování vysvětluje tak, že nás mozek odměňuje nikoli za přísun čerstvých informací, ale za přísun kolmogorovovské komplexity, přičemž nízká míra redundance je způsob, jak si dopřát většího množství komplexity, čímž toho o světě víme víc, a můžeme tedy snáze předvídat. Jakožto odborník na strojové učení Schmidhuber (2007, str. 2) připomíná, že v drtivé většině našeho života nemá naše učení žádnou externí odměnu a že se tedy typově jedná o *unsupervised learning* (učení bez dohledu, bez učitele). To, že náš organismus odměňuje mozek dopaminem za snižování redundance již přijatých zpráv (uvolňování místa pro nové zprávy) a za přijímání komplexity nové, obzvláště takové, která pomáhá snižovat redundanci zpráv již přijatých, není náhoda — takový mechanismus mu totiž pomáhá přežít a rozmnožit se. Tohoto mechanismu „zneužívají“ umělci, kteří tvoří díla, jež jsou s určitým kognitivním úsilím dobře komprimovatelná, přičemž po komprimaci je mozek náležitě odměněn. Příkladem budiž krápníková klenba (*muqarnasy*, jak je známe zejména z perské architektury), která se na první pohled zdá velmi složitá a až po nějaké době v ní člověk s úžasem začne nacházet jednoduchý systém.

Ovšemže vysoká komprimovatelnost a přínos informace pro komprimaci už uložených informací není jediným zdrojem estetična.

Dostáváme se opět k tomu, že teorie informace je výborným prostředkem funkcionálního vysvětlení. Tentokrát nám pomáhá pochopit, proč se nám líbí snadno komprimovatelné vzory a proč matematici milují *elegantní* důkazy, tedy takové, jejichž popis zabírá málo místa ve formálním jazyce, který používají. Vraťme se nyní ke kapitole 2.2. Schmidhuberův přístup nám ukazuje, proč jsou některé teorie považovány za esteticky krásné, respektive proč není nemístné používat ve vědě pojem estetičnosti a elegance — jedná se o metaforu informační úspornosti a nízké míry redundance.

Tady vidíme, proč má smysl teorii informace znát a mít ji na paměti, má totiž široký záběr: může sloužit jako explanatorní mechanismus ve fyzice a může být také explanatorním mechanismem pro chování fyziků. A samozřejmě i lingvistů.

Kapitola 7

Hranice segmentů

V mluvené řeči jsou hranice segmentů obvykle nějak naznačeny: přízvukem, melodií, pauzou a podobně. Také v dnešních psaných textech hranice segmentů nějakým způsobem označujeme. Z pohledu našeho modelu to dává smysl: podívejme se na strukturu z kapitoly 5.2.2, matici 5.6, a představme si, že jednotlivé řádky nejsou stejně dlouhé a že nám nic nenaznačuje, kde mají hranice — redundance nám může pomoci najít tyto hranice (všechny celky musí být konzistentní), ovšem jednak je tato operace komputačně velmi náročná, jednak tento postup drasticky omezí potenciál určování chyb a dopředné korekce (zjistit chybu můžeme pouze tehdy, když je kód natolik chybný, že se nám vůbec nepodaří najít žádný způsob, jak zprávu rozčlenit na konzistentní segmenty).

Nicméně vzpomeňme si na pozdně antickou a středověkou latinu a řečtinu, které hranice segmentů graficky nenaznačovaly (tzv. *scriptura continua*, viz kapitolu 4.2, str. 43). Je tedy zřejmé, že segmenty jsou od sebe odděleny i jinak, zpráva musí obsahovat i nějakou informaci o hranicích segmentů. Tato kapitola bude o tom, jak taková informace obvykle vypadá a co z toho vyplývá. Všechna měření budou provedena na českém a arabském korpusu a všechny příklady budou z arabštiny a češtiny, takže je asi nemístné usuzovat, že naše závěry mají univerzální platnost, přesto si však myslím, že zákonitosti zde popsané mají obecný charakter a jako takové by měly být i v budoucnu dále testovány.

7.1 Hranice segmentů — orámování

Přímočarým způsobem, jak naznačit hranice segmentů, je vkládání subsegmentů či subsubsegmentů,[1] které jsou pro tento účel vyhrazeny, respektive je to část jejich významu.

[1] Jako subsegmenty označujeme segmenty, které jsou o úroveň níž; subsubsegmenty jsou o dvě úrovně níž. Tedy pro větnou úroveň jsou subsegmenty slova a subsubsegmenty jsou morfy.

Pro větnou úroveň je typické užití některých spojek a spojovacích partikulí, příkladem mohou být subsegmenty *který* a *že* a jejich arabské ekvivalenty *ʾallaḏī* a *ʾan*, dále slova, jimž se v arabské tadici říká *ʾinna a její sestry* a jež s různými sémantickými modifikacemi často uvozují věty. V arabštině jsou i typizované subsubsegmenty *wa-* a *fa-*, přičemž *fa-* funguje jako hranice spíše pro celá souvětí, takže ho můžeme označit za hraniční subsubsubsegment (morf je o tři úrovně níž než souvětí, *fa-* je tedy subsubsubsegmentem souvětí, které delimituje; kromě samotné delimitace nese další význam, například logické vyvozování, v klasické arabštině může značit změnu mluvčího).

I na vyšších úrovních najdeme subsegmenty, které mají za úkol určovat hranice a strukturovat: arabská korespondence využívá například fráze *fīmā baʿdu* nebo *wa-baʿdu*, které mají za úkol oddělit úvod dopisu od jeho obsahu. Pro klasický arabský text je typická kompozice sestávají z výčtu témat, o kterých bude pojednáno, a jejich následném oddělování pomocí fráze *ʾammā* — *fa-* (pro celky, řekněme na úrovni souvětí) nebo fráze *fīmā yataʿallaqu bi-* pro vyšší celky, řekněme na úrovni odstavců nebo krátkých kapitol (ekvivalentem může být české *co se týče*, které ovšem není užíváno tak soustavně).

Tím se dostáváme na úroveň kapitol, kde v klasické arabské literatuře najdeme *basmalu*[2] a podobné v náboženství ukotvené fráze vyjadřující například, že *Bůh je v této věci znalejší*; v případě vypravování pak fráze jako *slyšel jsem* nebo *doneslo se mi*. Pohádky pak tradičně začínají frází *kāna mā kāna* či *kāna yā mā kāna*, což má podivuhodně doslovný ekvivalent (ovšem postrádající arabskou dvojznačnost) v českém *bylo nebylo*.

Na úrovni slov v arabštině a češtině žádné specializované hraniční morfémy nenalezneme, nicméně setkáme se zde s rigorózním morfosledem — některé morfy se vyskytují typicky na začátku/konci slova, takže kromě svého významu ještě mohou nést informaci o jeho hranicích. V arabštině dokonce existuje třída hlásek, které se v těchto hraničních morfémech mohou vyskytovat (tzv. *az-zawāʾidu*).[3] O těchto hláskách a jejich českých ekvivalentech podrobněji pojednávám v příloze B.

7.2 Jádro segmentu

V češtině i v arabštině se uplatňuje ještě jedna metoda jak rozlišit jednotlivé segmenty, respektive určit jejich počet: jeden ze subsegmentů může tvořit jakési jádro, respektive

[2] Tedy vyjádření *bi-ʾsmi ʾl-lāhi ʾr-raḥmāni ʾr-raḥīm*, což se do češtiny obvykle překládá jako *ve jménu Boha milosrdného, slitovného* — formule zahajující většinu *sūr* Koránu a vkládaná dodnes na začátek důležitějších textů, a to nejen náboženských.

[3] Arabské afixy se mohou skládat pouze ze souhlásek *y, t, s, m, n, w, ʾ, b, f, l, k*, ostatní se mohou vyskytovat pouze v kořenech. Zmíněná řada hlásek dohromady dává mnemotechnickou větu *yatasammanū bi-fulkin* (يتسمنوا بفلك — doslova „oni se stávají tučnějšími na lodi“), kterou jsem se před deseti lety naučil z podivuhodné Sedláčkovy gramatiky (Sedláček, 1898, str. 8) a dnes ji poprvé používám.

spadat do kategorie, která se v segmentu může objevit typicky pouze jednou. Na úrovni morfémů je takovým jádrem kořen (v arabštině společně s šablonou), na úrovni vět sloveso.

Tyto subsegmenty v segmentech nemusí být obsaženy zcela nutně a ve výjimečných případech jich může být více než jeden, takže segmentace podle nich nevede ke stoprocentnímu výsledku, nicméně v kombinaci s ostatními prvky mohou pomoci.

Typické je, že tyto subsegmenty jsou delší než zbytek (respektive obsahují více subsubsegmentů). Například české kořenové morfy v našem vzorku obsahují průměrně 3,1 hlásek, zatímco ostatní průměrně 1,4 hlásek;[4] arabská slovesa, která můžeme považovat za jádro arabské věty, neboť jen stěží může být ve větě více než jedno, má průměrně 3,5 morfů, zatímco ostatní slova průměrně 2,3 morfů.[5]

7.3 Co vyplývá z takového vkládání delimitační informace

Shrňme si poslední dvě podkapitoly: do segmentů je kromě vlastní přenášené informace a redundance vložena informace, díky které je jednodušší určit hranici segmentu. Tato informace může být obsažena ve vyhrazených subsegmentech (česká spojka *načež*) i subsubsegmentech (arabská spojovací partikule *fa-* se stejným významem).

Velikost této vložené informace (můžeme jí říkat třeba delimitační) je principiálně nezávislá na velikosti segmetu, který ohraničuje. To ovšem neznamená, že musí být stále stejně velká — délka segmentů variuje kolem určité hodnoty, počet delimitačních subsegmentů a subsubsegmentů také, ovšem ony variance nejsou vzájemně korelovány.[6]

Jak bylo řečeno, delimitační subsegmenty a subsubsegmenty mají často i jinou funkci než delimitační, takže je obtížné je od ostatních odlišit. Nezávislost delimitační informace na velikosti vlastní informace a redundance má však zcela konkrétní důsledky pro podobu textu.

Začněme zcela intuitivním modelem, který zachycuje text tak, jak ho vnímáme jakožto mluvčí. Segment se skládá z běžných subsegmentů, jejichž délka je nezávislá na délce segmentu. Například průměrný počet morfů ve slově je nezávislý na tom, kolik slov má věta, jejíž je slovo součástí. Délka segmentu L na úrovni n (tedy L_n)

[4] Měřeno na novele *Krysař* od Viktora Dyka, segmentaci provedla Zuzana Komrsková.

[5] Měřeno na části knihy *Kalīla wa-Dimna* , jejímž autorem je *ʿAbdallāh ibn al-Muqaffaʿ*, segmentace vlastní. O tomto a o v předchozí poznámce zmíněném textu a způsobu segmentace bude detailněji pojednáno v kapitole 7.4.

[6] Vhodnou, snadno představitelnou metaforou pro delimitační segmenty jsou třeba rámy obrazů — nemám tuto informaci exaktněji ověřenu, ale při procházení větší galerie, řekněme Louvru, získáte dojem, že tloušťka rámu nezávisí na velikosti obrazu. Tedy že malé obrazy mohou mít rámy úzké i široké, takže miniatury mohou mít rámy širší než největší plátna, ale najdou se i drobné kresby se subtilním rámem nebo úplně bez něj.

tedy neovlivňuje průměrnou délku svých subsegmentů (t. j. segmentů na nižší úrovni, značím ji $\bar{L}_{n-1}$). Funkce, která tuto průměrnou délku vyjadřuje, je proto konstantní:

$$\bar{L}_{n-1} = a \tag{7.1}$$

Tento model rozvineme o představu popsanou v předchozích podkapitolách: každý segment obsahuje určité množství delimitačních subsubsegmentů, jejichž délka je taktéž nezávislá na délce segmentu. Například určité stabilní procento arabských vět obsahuje delimitační partikuli *fa-* a podobně, přičemž tato pravděpodobnost je nezávislá na tom, kolik má která věta slov. Průměrný počet těchto subsubsegmentů si označme parametrem b. Získáme tak přehlednou rovnici:

$$\bar{L}_{n-1} = a + \frac{b}{L_n} \tag{7.2}$$

Dále se segment skládá z delimitačních subsegmentů. Průměrný počet delimitačních subsegmentů je též nezávislý na délce segmentu, ovšem s výjimkou segmentu o jednom subsegmentu. Těžko se může vyskytovat česká věta skládající se z jediného slova, které je zároveň slovem delimitačním, například slova *načež*, neřkuli jednoslovná arabská věta obsahující pouze slovo *ʾinna* (respektive takových vět je minimální množství), nebo slovo skládající se pouze z předpony. Můžeme očekávat, že tyto delimitační subsegmenty budou mít, vzhledem ke své pomocné funkci, jinou průměrnou délku než běžné segmenty. Pravděpodobnost jejich výskytu si označme p a rozdíl mezi jejich průměrnou délkou a průměrnou délkou běžného segmentu si označme jako d. Dostaneme se tak k mírně složitějšímu vzorci pro $L_n > 1$:

$$\bar{L}_{n-1} = a + \frac{b}{L_n} + \frac{pd}{L_n} \tag{7.3}$$

Přičemž pro snadnější manipulaci můžeme parametry p a d nahradit jedním parametrem c, takže c interpretujeme jako $c = pd$.

$$\bar{L}_{n-1} = a + \frac{b}{L_n} + \frac{c}{L_n} \tag{7.4}$$

Tento vzorec, připomínám, platí pouze pro $L_n > 1$; pro $L_n = 1$ stále platí rovnice 7.2, z níž dosazením dostaneme $L_{n-1} = a + b$. Při ověřování modelu na datech budeme z technických důvodů používat vzorec vzniknuvší spojením oněch dvou vzorců, rovnici 7.5.[7]

$$\bar{L}_{n-1} = a + \frac{b}{L_n} + \frac{c \ \min(1, L_n - 1)}{L_n} \tag{7.5}$$

[7] V této podobě model naleznete v článku Milička (2014), kde je jeho vysvětlení poněkud abstraktnější, nicméně možná názornější a detailnější.

Funkce $\min(1, L_n - 1)$ má jedinou úlohu, a totiž zajistit, aby při segmentu o délce jeden subsegment byl parametr c vynásoben nulou. Totéž by mohlo být vyjádřeno různými způsoby, například:

$$\tilde{L}_{n-1} = a + \frac{b}{L_n} + \frac{c\ \lceil 1 - \frac{1}{L_n} \rceil}{L_n} \tag{7.6}$$

Model průměrné délky subsegmentů, jak je vyjádřen v posledních dvou rovnicích, je pro běžného mluvčího značně neintuitivní. Říká, že když třeba roztřídíme slova podle počtu morfů, pak délka oněch morfů bude záviset na délce oněch slov. Tedy například pokud vybereme z textu všechna slova o pěti morfech a spočítáme průměrnou délku oněch morfů a následně vybereme všechna slova o dvou morfech a též spočítáme průměrnou délku, že se ty dvě délky budou lišit a že tato odlišnost je systematická. Ba co víc, že onen systém, který platí pro délku morfů ve slovech, je stejný jako ten, podle kterého se řídí délka slov ve větách a délka vět v souvětích. Jako mluvčí ovšem na vědomé úrovni nic takového nepozorujeme. Na místě je tedy otázka: platí tento model skutečně pro reálný text?

7.4 Ověření na datech

Model z předchozí kapitoly ověříme na jednom českém textu a na části textu arabského. Budeme se tedy muset spokojit s poměrně malým korpusem, nicméně vzhledem k náročnosti segmentace textů na morfy a věty, které musí být prováděno ručně, to bude i tak ojedinělý počin.

7.4.1 Český text

Český text, Dykova novela *Krysař*, byl očištěn od formátovacích znaků a následně segmentován na pěti úrovních:

Na úrovni hlásek. Zde jsem vycházel z písemné podoby textu, takže bylo třeba pouze odstranit spřežky (*ch* → *X*) a rozvinout kondenzované tvary ([*bpv*]*ě* → [*bpv*]*je*; *mě* → *mňe*), následně byl přepis ručně zkontrolován.

Na úrovni morfů. Morfematickou analýzu provedla Zuzana Komrsková, přičemž se řídila (Grepl et al., 1994; Hrbáček, 1984) a svou intuicí (ovšem ovlivněnou bohemistickým vzděláním).

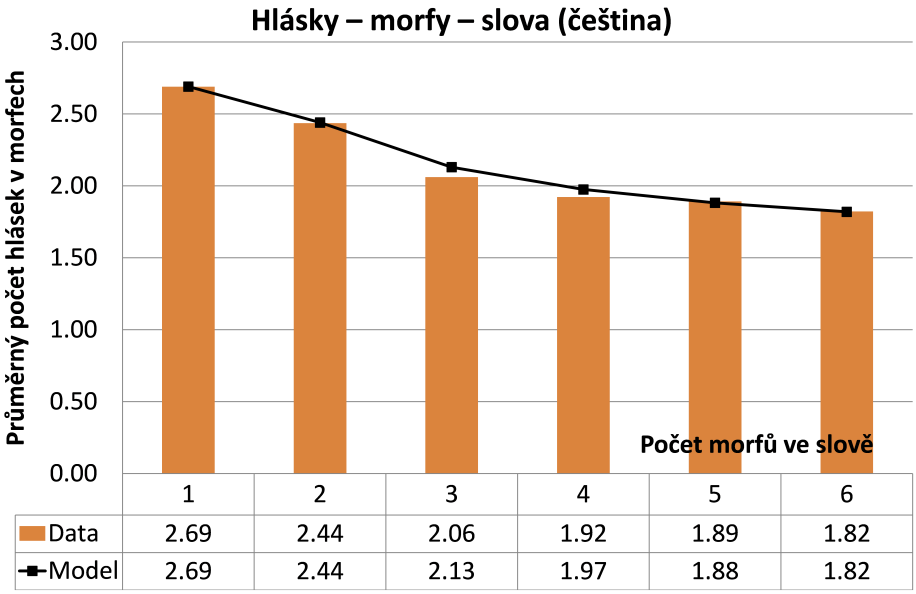
Na úrovni slov. Primárně byly brány v potaz mezery mezi slovy a další delimitační znaky v psaném textu. Takovým segmentům se často říká „ortoslovo“. Opět byla důležitá ruční kontrola.

Na úrovni vět. Provedl jsem sám ručně, přičemž jsem kladl důraz na to, aby byly správně označeny vložené vedlejší věty.

Na úrovni souvětí. Asi jediná úroveň, kde bylo možno provádět segmentaci jednoduše a systematicky na základě grafické podoby textu, nicméně i zde proběhla ruční kontrola.

Na takto segmentovaném textu, jehož ukázku a podrobnější popis najdete v příloze C.1, byl následně změřen vztah mezi délkou segmentů a délkou jejich subsegmentů pomocí k tomu naprogramovaného nástroje.[8]

Výsledky na úrovni hlásky — morfy — slova vidíte na grafu 7.1. Náš model nejlépe sedí s parametry $a = 1{,}51$; $b = 1{,}18$; $c = 0{,}68$. Tato funkce s nejmenší průměrnou absolutní odchylkou (0,022, což odpovídá $R^2 = 0{,}99$) byla nafitována pomocí software Eureqa od Nutonian (Schmidt – Lipson, 2009).

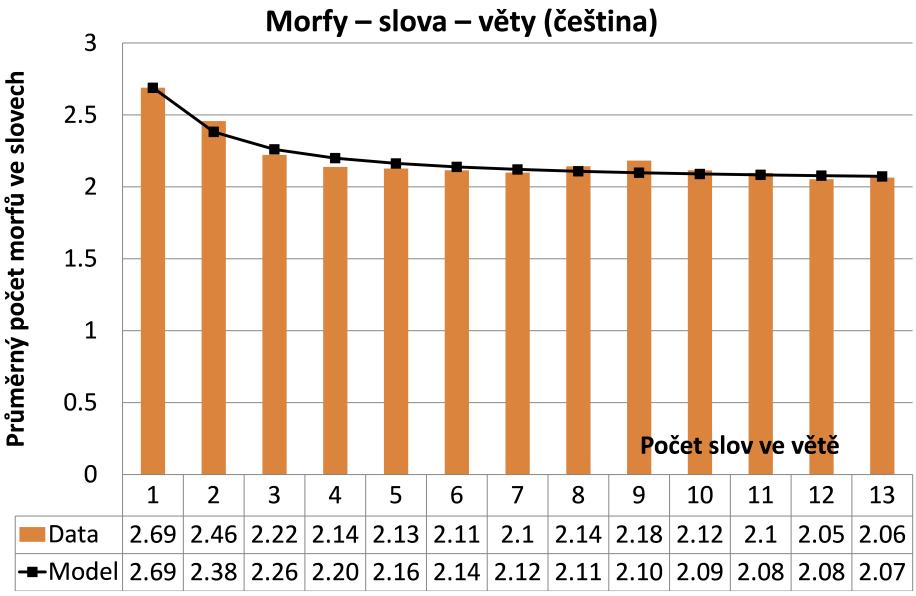


	1	2	3	4	5	6
Data	2.69	2.44	2.06	1.92	1.89	1.82
Model	2.69	2.44	2.13	1.97	1.88	1.82

Obrázek 7.1: Úroveň hlásek — morfů — slov.

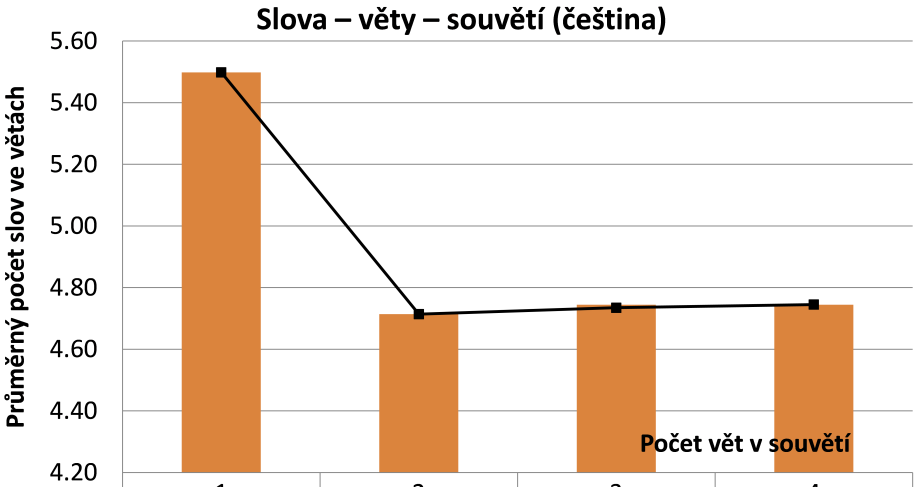
Výsledky na úrovni morfy — slova — věty jsou zobrazeny v grafu 7.2. Byl použit stejný model, tentokrát s parametry $a = 2{,}02$; $b = 0{,}67$; $c = 0{,}06$. Nízká hodnota parametru c znamená, že delimitační slovo je průměrně skoro stejně velké jako slova ostatní. Minimalizována byla opět průměrná absolutní odchylka (0,037, což odpovídá $R^2 = 0{,}89$).

[8] Tento nástroj najdete, podobně jako celý segmentovaný text a všechna ostatní data týkající se této disertace, na adrese `http://milicka.cz/disertace.zip`.



	1	2	3	4	5	6	7	8	9	10	11	12	13
Data	2.69	2.46	2.22	2.14	2.13	2.11	2.1	2.14	2.18	2.12	2.1	2.05	2.06
Model	2.69	2.38	2.26	2.20	2.16	2.14	2.12	2.11	2.10	2.09	2.08	2.08	2.07

Obrázek 7.2: Úroveň morfů, slov a věty.



	1	2	3	4
Data	5.50	4.71	4.74	4.74
Model	5.50	4.71	4.73	4.75

Obrázek 7.3: Slova — věty — souvětí.

Výsledky na úrovni slova — věty — souvětí najdete na grafu 7.3. Parametry jsou $a = 4{,}78$; $b = 0{,}72$; $c = -0{,}85$. Záporný parametr c znamená, že delimitační věta je průměrně kratší než věty ostatní. Průměrná absolutní odchylka má hodnotu 0,0025; $R^2 = 1{,}0$ (k tomuto podezřele dobrému výsledku je nutno dodat, že fitovat model o třech parametrech na datovou řadu o čtyřech hodnotách není žádné velké hrdinství).

7.4.2 Arabský text

I pro arabštinu byl vybrán jednoduchý narativní text, a to část knihy *Kalīla wa-Dimna*, jež je arabskou verzí Paňčatantry a jejímž autorem je *ʿAbdallāh ibn al-Muqaffaʿ*, vlastním jménem *Rūzbih*.[9] Vzhledem k náročnosti segmentace jsem oanotoval pouze přibližně tisícislovný úsek textu. Podobně jako v případě češtiny, i arabská segmentace proběhla na pěti úrovních:

Na úrovni hlásek. Vycházel jsem z redakčně vokalizovaného textu,[10] nicméně počítačově zpracovaná předloha byla nevokalizována, bylo tedy nutné hlásky doplnit, přičemž jsem rovnou ručně převáděl text do fonologického přepisu.

Na úrovni morfů. Morfematickou analýzu jsem provedl podle následujícího klíče:[11] za morf jsem považoval kořen, kmen, slovesné koncovky, jmenné koncovky, klitika (jako *wa-*, *fa-*, *li-*, *bi-*), člen, zájmenné přípony, ženský rod u jmen a některé další morfémy jako například předpony *ma-* a *mu-* pro tvoření jmen. Slovesné a jmenné koncovky, vzhledem k jejich syntetické povaze, jsem navzdory množství významů, které nesou, nerozděloval (takže například slovo *fa-taḍribūna* je segmentováno jako čtyři morfy). Toto rozdělení je arbitrární a existuje množství alternativních způsobů.[12]

Na úrovni slov. Primárně byla brána v potaz grafická podoba, tedy opět mezery mezi slovy a další delimitační znaky. Také zde jsme se nespokojili s pouze automatickým dělením a proběhla ruční kontrola.

Na úrovni vět. Opět jsem kladl důraz na to, aby byly správně označeny vložené vedlejší věty, vzhledem k povaze arabského zacházení s interpunkcí (a vzhledem

[9] U nás vyšla v překladu profesora Oliveriuse jako Kalíla a Dimna (Dar Ibn Rushd, 2004).

[10] Konkrétně *Kitāb Kalīla wa Dimna*, nakladatelství *Maktabat Lubnān*, třetí vydání, Bejrūt 1991. Redaktor, jenž byl autorem vokalizace, sice není uveden, nicméně vzhledem k místní tradici nejspíše půjde o reprint vydání od Louise Scheikha.

[11] Opět připomínám, že stručné uvedení do arabské morfologie naleznete v poznámce 8 na straně 36.

[12] Například bych mohl toto slovo rozdělit na pět morfů: kořen *ḍrb*, kmen $R_1R_2iR_3$, předponu *ta-*, koncovku *-ūna* a klitikon *fa-*. Proti takovému rozdělení ovšem hovoří to, že význam předpony *ta-* je zcela určen teprve po přidání koncovky (ve tvaru *ta-* <kořen + kmen> *-u* značí tvar 1. osoby ženského rodu, zatímco *ta-* <kořen + kmen> *-ūna* je tvarem 2. osoby mužského rodu).

k tomu, že není původní a byla dodána do textu až dodatečně) jsem segmentaci prováděl ručně podle vlastního uvážení, primárně podle syntaktických struktur, přiznávám, že značně ovlivněn českou tradicí dělení na věty (tedy například věty začínající na *ʾinna*, po němž následuje sloveso ve finitním tvaru, jsem chápal jako jednu větu, nikoli jako věty dvě, jak je zvykem v arabské tradici).

Na úrovni souvětí. Segmentaci jsem sice provedl (na základě editorem ex post redakčně vložené interpunkce), ovšem vzhledem k malému počtu vět ve vzorku jsem se rozhodl ji nepoužít.

Přiznávám, že má segmentace je příliš arabistická/orientalistická a že ideální by bylo zadat segmentaci rodilému mluvčímu, ideálně mnoha rodilým mluvčím, k čemuž jsem ovšem neměl prostředky. I tak se ovšem můžeme podívat na výsledky, přičemž můžete provést modifikace mé segmentace (příklad a detailní popis najdete v příloze C.2) nebo se alespoň podívat na konkrétní dataset.[13]

Výsledky na úrovni hlásky — morfy — slova najdete v grafu 7.4, přičemž model se podařilo nafitovat při minimalizované průměrné absolutní odchylce 0,046 (odpovídá $R^2 = 0{,}98$). Hodnoty parametrů jsou $a = 1{,}14$; $b = 2{,}38$; $c = 0{,}82$ (průměrná odchylka 0,046; $R^2 = 0{,}98$). Podle tohoto modelu tak delimitační hlásky zabírají víc místa než v češtině, což nás přivádí k hypotéze, že hranice arabských slov budou snáze určitelné než hranice slov českých. Tato hypotéza přímo vybízí k testování pomocí nějakého percepčního experimentu.

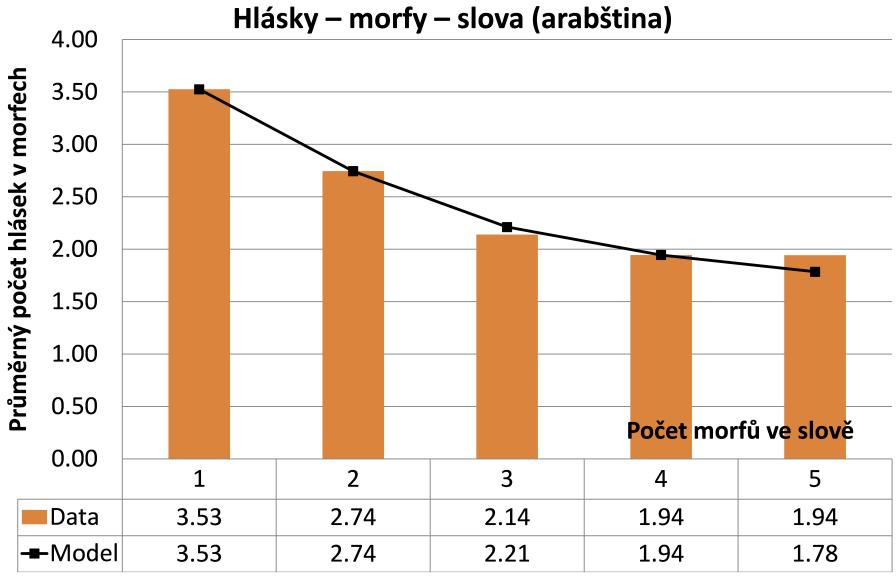
Výsledky na úrovni morfy — slova — věty vypadají lépe (graf 7.5) a korespondují s tím, co jsme viděli na českém textu: parametry jsou $a = 2{,}73$; $b = 0{,}94$; $c = -0{,}14$ a nafitovat se je podařilo s průměrnou absolutní odchylkou 0,012, což odpovídá determinačnímu koeficientu $R^2 = 1$.

7.5 Menzerathův vztah

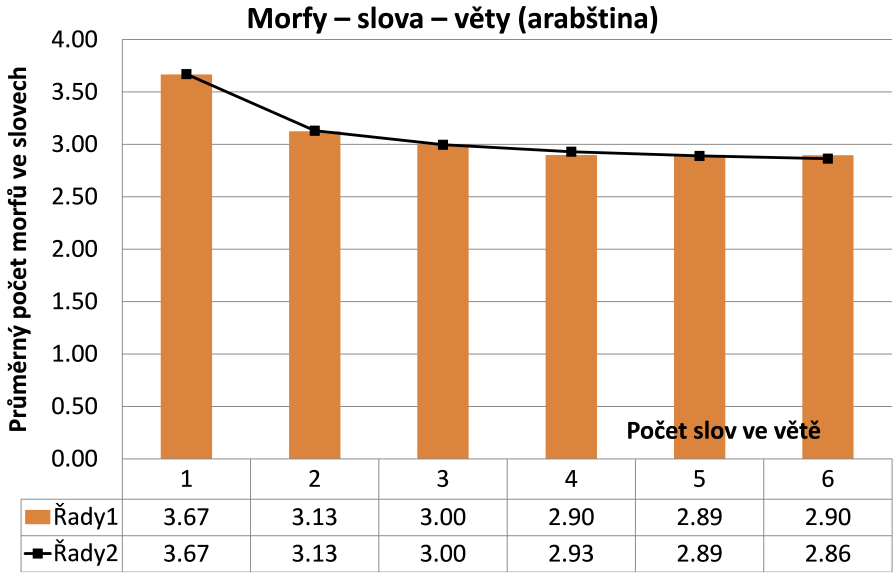
Jako první si podivného vztahu mezi délkou segmentů a délkou jejich subsegmentů všiml Paul Menzerath (1928, 1954). Sám Menzerath by tento vztah modeloval nejspíše rovnicí 7.2, jak vyplývá z (Menzerath, 1954, str. 108), ovšem pro tuto rovnici žádnou explanaci nehledal.

Opravdová fascinace Menzerathovým vztahem nastala až po roce 1980, kdy Gabriel Altmann (1978, 1980) zveřejnil své modely a kdy skupina kvantitativních lingvistů kolem časopisu Glottometrika postupně zjišťovala, že je všudypřítomný nejen v lingvistice, ale i v ostatních systémech pracujících s informacemi, například v genetice (většina literatury na toto téma je zachycena v (Altmann, 2014)). Altmannova zásluha na rozšíření studia tohoto problému je natolik uznávaná, že se dnes mluví

[13] Opět připomínám, že vše najdete na adrese `http://milicka.cz/disertace.zip`.



Obrázek 7.4: Hlásky — morfy — slova.



Obrázek 7.5: Morfy — slova — věty.

o Menzerath-Altmannově zákoně (*Menzerath-Altmann Law*), přičemž se tak označuje jak několik různých funkcí, tak samotný vztah mezi délkou segmentů a subsegmentů a několik dalších jevů.[14]

Ostatně jedním z motivů k této práci byla otázka měření dat pro Menzerathův vztah — jak správně segmentovat? Tato otázka přímo vyvolává otázku další: co vlastně jsou ony segmenty a proč vůbec vznikají? V této kapitole beru jako samozřejmost, že segmentace je chápána podle sémantických kritérií, tedy například slova jsou segmentována na morfy, jejichž délka je měřena počtem hlásek, nicméně původní Menzerathův vztah byl zamýšlen jako vztah mezi délkou slabik v hláskách a délkou hlásek v milisekundách, což je spíše technické hledisko. Pro Menzerathův vztah na těchto nesémantických úrovních nejspíš bude třeba hledat jiné vysvětlení (popřípadě je chápat jako vedlejší produkt Menzerathova vztahu na úrovních definovaných sémanticky).

Je docela možné, že můj model funguje i na jiný druh segmentace, než jak ho představuji, nebo dokonce na segmentace úplně jiných jevů, než je text, nicméně pak je třeba počítat s tím, že i explanace může být zcela odlišná, respektive že je možné model na data nafitovat čistě náhodou. Ostatně monotónně klesající funkce, jak ji vidíme na grafu 7.1, mohla být vyprodukována mnoha různými stochastickými jevy. A taky je možné ji popsat mnoha fundamentálně odlišnými rovnicemi.

Zde je jen několik málo příkladů funkcí, které bylo možné nafitovat na úroveň morfů — slov — vět na arabském korpusu se srovnatelnou úspěšností jako Altmannovy a mé modely. Kromě množství polynomických funkcí, které se počtem parametrů blížily počtu datových bodů datasetu, nebo jej dokonce překračovaly, se genetickému algoritmu programu Eureqa (Schmidt – Lipson, 2009) podařilo najít například tyto vcelku úhledné rovnice:

$$\tilde{L}_{n-1} = a + \frac{b \max(L_n, c)}{L_n} \tag{7.7}$$

$$\tilde{L}_{n-1} = a + \frac{b}{e^{L_n}} \tag{7.8}$$

$$\tilde{L}_{n-1} = a + \frac{b}{L_n \sqrt{L_n}} \tag{7.9}$$

[14] Chci důsledně rozlišovat mezi vztahem a jeho modelem, a proto budu mluvit o *Altmannově modelu Menzerathova vztahu*. Nedostatečné rozlišování mezi vztahem a jeho modelem je ovšem v kvantitativní lingvistice běžné, takže se ztotožňuje např. *Zipfův vztah* (vlastně *rank-frequency relation*) a *Zipfův model* (t.j. $f_n = f_1/n$). Paradoxně se jako Zipfův zákon označují i nezipfiánské varianty Zipfova modelu, které by ovšem Zipf sám označil jako „introducing the devil into the formula" (Wyllys, 1981, str. 58), takže místo termínu *Mandelbrotův model Zipfova vztahu* se používá termín *Zipf-Mandelbrotův zákon*. Použití slova *zákon* modelům propůjčuje domněle vyšší statut, podle vzoru Newtonových zákonů a podobně.

$$\tilde{L}_{n-1} = a + \frac{b}{{L_n}^2} \tag{7.10}$$

$$\tilde{L}_{n-1} = a + {L_n}^{-b} \tag{7.11}$$

$$\tilde{L}_{n-1} = a + \frac{b}{{L_n}^2} + c\, e^{L_n} \tag{7.12}$$

$$\tilde{L}_{n-1} = a + \frac{b}{{L_n}^{L_n}} - c\, {L_n}^2\, {L_n}^{L_n} \tag{7.13}$$

Proto vidím jako naprosto nezbytné postupovat při odvozování vzorce směrem od principů ke vzorci a od něj k naměřeným datům, jako jsem to učinil v předcházejících podkapitolách;[15] ovšem i při zachování tohoto pravidla existuje mnoho způsobů, jak model měnit. Například si dokážu představit návrh, že velikost části redundance přidávané k segmentu není závislá na jeho délce lineárně, ale logaritmicky (v souladu

[15]V tomto ohledu stojí za bližší pohled odvození Altmannova modelu. Altmannův model je ekvivalentní úpravou diferenciální rovnice $\frac{d\tilde{L}_{n-1}}{dL_n} = -a\tilde{L}_{n-1}$, kterou můžeme do běžné řeči přeložit jako tvrzení, že míra změny velikosti průměrného subsegmentu je lineárně závislá na jeho velikosti, tedy například že čím menší je průměrný subsegment v segmentech o dvou subsegmentech, tím menší bude rozdíl mezi ním a průměrným subsegmentem v segmentech a třech subsegmentech. Přičemž ovšem vůbec není jasné, proč by onen vztah měl platit a z jakého principu vychází či jaký stochastický jev zachycuje. Toho si je samozřejmě Gabriel Altmann vědom — jednak nemá problém v následujícím článku vztah změnit za podobným způsobem derivovanou rovnici lépe odpovídající empirickým datům (Altmann, 1980, str. 2–3), jednak píše doslova „[i]t remains theoretically not fully validated hypothesis [...]. We merely suspect that it is somehow connected with the principle of least effort or with some not yet known principle of balance recompensating lenghtening on one hand with shortening on the other“ (Altmann, 1980, str. 5), jednak v souladu se svým programem vyjádřeným v (Altmann, 1978, str. 6) chápe svůj model nikoli jako příspěvek k explanaci Menzerathova vztahu, ale jako testovatelné východisko, ze kterého je možné Menzerathův vztah používat jako explanační princip pro jiné jazykové jevy (Altmann, 1980, str. 8–9). Explanační a interpretační vakuum bylo jeho současníky nejspíše pociťováno, neboť o pár let později Köhler (1984) vydal článek s příznačně pokorným názvem *Zur Interpretation des Menzerathschen Gesetzes*, v němž popisuje explanatorní princip, kdy postuluje *Strukturinformation*. Tou míní informaci, kterou si kognitivní aparát musí vytvářet pro zpracování segmentů a která závisí na počtu subsegmentů, a tak tlačí na to, aby v segmentech s více subsegmenty byly subsegmenty kratší, neboť se tak lépe vejdou do pracovní paměti; načež předpokládá, že změna velikosti kapacity potřebné ke zpracování je nepřímo úměrně závislá na počtu segmentů, čímž se dostává ke kýžené diferenciální rovnici (str. 180, tamtéž), aniž ovšem udává přesný postup, který právě k tomuto vztahu vede. Přesto byl během posledních třiceti let Altmannův model málokdy kritizován (výjimkou je (Meyer, 2007)) a minimálně variován, těžko říci, jestli za to mohla nedostupnost původní publikace, magie spojená s infinitezimálním počtem pro část výzkumníků, nepotřebnost explanace pro jinou část či všeobecná potřeba nějakého stabilního a obecného lingvistického zákona pro všechny zúčastněné.

s kódováním typu Hammingova kódu, viz str. 17). Výsledný vzorec by pak vypadal třeba následovně:[16]

$$\tilde{L}_{n-1} = a + \frac{b + c\ \lceil 1 - \frac{1}{L_n} \rceil + d \log_2 L_n}{L_n} \tag{7.14}$$

Je třeba ovšem počítat s tím, že každá metoda segmentace je arbitrární a že metody segmentace mohou mít obrovský vliv na to, jak modely sedí. To je jeden z důvodů, proč není možné přímo a jednoduše porovnávat úspěšnost modelů podle toho, jak si vedou v modelování dat; ještě důležitější je, že neexistuje konsensus ohledně metriky úspěšnosti modelu — viz Mačutek – Wimmer (2013). Proto si myslím, že je zcela zásadní, aby model každého vztahu byl chápán jako otevřený problém.

[16]Velmi hrubý návrh, ovšem nafitovat jde vcelku solidně (což ovšem není překvapení vzhledem k tomu, že má 4 parametry). Průměrná absolutní odchylka pro český text na úrovni hlásky — morfy — slova je 0,008($R^2 = 1$), pro úroveň morfy — slova — věty 0,025($R^2 = 0,95$) a slova — věty — souvětí 0,09($R^2 = 0,90$). Pro arabský pak na úrovni hlásky — morfy — slova 0,02($R^2 = 1$) a pro úroveň morfy — slova — věty 0,035($R^2 = 0,95$).

Závěr

قَالَ الخَلِيلُ بنُ أَحمَدٍ
الرِجَالُ أَربَعَةٌ،
رَجُلٌ يَدرِي وَيَدرِي أَنَّهُ يَدرِي فَذٰلِكَ عَالِمٌ فَاتَّبِعُوهُ،
وَرَجُلٌ يَدرِي وَلَا يَدرِي أَنَّهُ يَدرِي فَذٰلِكَ نَائِمٌ فَأَيكَذُوهُ،
وَرَجُلٌ لَا يَدرِي وَيَدرِي أَنَّهُ لَا يَدرِي فَذٰلِكَ مُستَرشِدٌ فَأَرشِدُوهُ،
وَرَجُلٌ لَا يَدرِي وَلَا يَدرِي أَنَّهُ لَا يَدرِي فَذٰلِكَ جَاهِلٌ فَارفُضُوهُ.

Řekl al-Ḫalīl ibn Aḥmad:

„Jsou čtyři druhy mužů.

Muž, jenž ví a ví, že ví, to je mudrc, následujte ho.
Muž, jenž ví a neví, že ví, ten spí, probuďte ho.
Muž, jenž neví a ví, že neví, to je student, učte ho.
Muž, jenž neví a neví, že neví, to je ignorant, zavrhněte ho.“

Citát pochází z al-Ġazālīho Oživení věd náboženských[17] a vzhledem k jeho stáří mu snad odpustíme jistý sexistický nádech — takové jsou samozřejmě i ženy. Pevně věřím, že pozorní čtenáři i čtenářky této disertace se přesunuli do kategorie těch, kdo ví, proč ví, že ví, když ví, a ví, že ví, a proto pro ně bude tato kapitola, v souladu s výše napsaným i s úctyhodnou tradicí, z velké části redundantní.

Je více způsobů, jak číst tuto práci. Můžete ji chápat jako snahu o ucelený pohled na lingvistiku jakožto vědu a na předmět jejího bádání prizmatem teorie komunikace; jako ukázku toho, že její obecné principy mohou být dobrým nástrojem pro explanaci lingvistických jevů. Je málo lingvistických prací, které přímo explicitně zmiňují

[17] Abū Ḥāmid Muḥammad ibn Muḥammad al-Ġazālī: *ʾIḥyāʾ ʿulūmi ʾd-dīn (rubʿu ʾl-ʿibādāt, kitābu ʾl-ʿilm, al-bāb as-sādis)* — Oživení věd náboženských (díl první, kniha první, kapitola šestá). Tento výrok se traduje v různých variantách a poprvé byl nejspíš zapsán v 8. století našeho letopočtu ibn al-Ḫaṭṭābem v knize *Ǧamharatu ʾašʿāri ʾl-ʿArab*.

kolmogorovovskou komplexitu a redundanci a stále není běžné, aby se s teorií komunikace operovalo jako s hlavním explanačním principem, přestože implicitně je v lingvistice obsažen snad už před Shannonem a Kolmogorovem a přestože na shannonovské teorii informace je z velké části založeno současné počítačové zpracování přirozeného jazyka. Tato práce může sloužit i jako lehký úvod do tématu pro lingvisty, kteří se s kolmogorovovskou teorií komunikace nesetkali; snažil jsem se, aby výklad byl přístupný i pro ty, kteří si v matematice příliš nevěří. V této úloze by práci mohly dopomoci pečlivě zvolené reference — žádná položka se na seznam literatury nedostala bez zvažování jejího přínosu pro čtenáře a u zásadních témat jsem se snažil zahrnout jak původní autory, kteří se jimi zabývali poprvé, tak současnou literaturu, která přehledně a čitelně shrnuje dosavadní vývoj.

Druhý způsob, jak práci chápat, je však důležitější: mou primární snahou bylo podat vysvětlení toho, proč je mluvená a psaná řeč segmentovaná do slov, vět a souvětí, poukázat na nesamozřejmost takového dělení a na důsledky, které má. Mé vysvětlení je založené na teorii komunikace, obecných pravidlech, kterým z principu podléhá každá metoda přenosu informací ať už mezi lidmi, nebo jinými entitami. Ono vysvětlení je jenom jedním z mnoha možných, rozhodně nechci říct, že segmentace nemá i jiný důvod, může jich být spousta a nejspíš se budou vzájemně podporovat. Dokonce ani neříkám, že vkládání redundance na více úrovních je vůbec nejdůležitější faktor pro existenci segmentace, segmentace může mít mnoho funkcí.

Mým cílem bylo vzbudit vůbec debatu o tomto fenoménu, neboť jsem nenašel, že by si někdo tuto otázku vůbec položil, nikoli podat dokonalé a konečné vysvětlení. Text má být disertací v etymologickém slova smyslu: *cestou* k vědění, nikoli jejím cílem. Má inspirovat k úvahám a diskusi, nikoli podávat konečný argument. Nicméně hypotézy o povaze šumů, s nimiž se jazyk musí vypořádat, jakož i model vkládání redundance, jsou formulovány natvrdo a jsou plně testovatelné — tedy je možné pomocí nich predikovat naše schopnosti a chování a tyto predikce testovat. Nedisponoval jsem zdroji k provedení percepčních testů, nicméně jsou v plánu (hypotézy zaujaly několik kolegů) a na jejich výsledky jsem skutečně zvědavý.

* * *

Pokud bych měl shrnout hlavní myšlenku celé disertace: lingvistika chápe slova, věty a souvětí jako jakési celky, ovšem není jasné, v čem ona ucelenost spočívá. Podle této studie ona ucelenost znamená, že na konci těchto segmentů jsme schopni rozhodnout, jestli je text konzistentní a jestli mu rozumíme, a to lépe než uprostřed nich. Tato vlastnost není samozřejmá, ale je dána tím, že je do nich specifickým způsobem vložena redundance.

Seznam použité literatury

ALTMANN, G. Bibliography: Menzeraths Law. *Glottotheory*. 2014, 5, 2, s. 121–123. doi: 10.1515/glot-2014-0008.

ALTMANN, G. Towards a Theory of Language. *Glottometrika*. 1978, 1, s. 1–25.

ALTMANN, G. Prolegomena to Menzerath's law. *Glottometrika*. 1980, 2, s. 1–10.

ASHBY, W. R. Principles of the self-organizing system. In FOERSTER, H. – G. W. ZOPF, J. (Ed.) *Principles of Self- Organization: Transactions of the University of Illinois Symposium*, s. 255–278. Pergamon Press: London, 1962.

BAKR, J. *Nuṣūṣ fī 'n-naḥwi 'l-'arabīji 1, 2 (Texty o arabské gramatice)*. s.d.

BANE, M. Quantifying and measuring morphological complexity. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, Sommerville, USA, 2008. ISBN 978-1-57473-423-2.

BAR-HILLEL, Y. An Examination of Information Theory. *Philosophy of Science*. 1955, 22, 2, s. 86–105. ISSN 00318248.

BARABÁSI, A.-L. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, reissue edition, April 2003. ISBN 0452284392.

BATESON, G. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Bungay Suffolk: Chandler Publishing, 1972. ISBN 9780226039053.

BIČOVSKÝ, J. Why languages never change. In DOČKALOVÁ, L. (Ed.) *Indoevropská fonologie a morfologie: Sborník z Brněnského indoevropeistického kolokvia*. Brno: Masarykova univerzita, 2010a. ISBN 978-80-210-5301-4.

BIČOVSKÝ, J. *Intuice srovnávací metody*. Disertace, Filozofická fakulta, Univerzita Karlova v Praze, 2010b. Dostupné z: <https://is.cuni.cz/webapps/zzp/download/140001070/>.

BIELICKÝ, V. *Způsoby rozšiřování arabské slovní zásoby s důrazem na publicistický styl*. Diplomní práce, Filozofická fakulta, Univerzita Karlova v Praze, 2007. Dostupné z: <https://is.cuni.cz/webapps/zzp/download/120064562/>.

BOERSMA, P. – WEENINK, D. Praat: doing phonetics by computer [Computer program]. Version 5.4.15, 2015. Dostupné z: <http://www.praat.org/>.

CAMPBELL, J. *Grammatical man: information, entropy, language, and life*. Colección Popular. Simon and Schuster, 1982. ISBN 9780671440619.

CARSTAIRS-MCCARTHY, A. *The Evolution of Morphology*. Oxford Uhniversity Press, 2010. ISBN 978-0-19-929978-2.

CORRIENTE, F. C. On the Functional Yield of Some Synthetic Devices in Arabic and Semitic Morphology. *The Jewish Quarterly Review*. 1971, 62, 1, s. 20–50. ISSN 00216682.

COULMAS, F. *The Blackwell Encyclopedia of Writing Systems*. Wiley-Blackwell, 1999. ISBN 063121481X.

DAVIS, S. *The Encyclopedia of Arabic Language and Linguistics*, 3, Velarization, s. 636–638. Brill Academic, 2006–2009. ISBN 9004149732.

HOOP, R. – KORPEL, M. C. – PORTER, S. E. (Ed.). *The Impact of Unit Delimitation on Exegesis. 7 / Pericope. Scripture as Written and Read in Antiquity*. Brill, 2009. ISBN 9004171622.

ELIAS, P. Error-free Coding. Technical Report 285, Research Laboratory of Electronics, Massachusetts Institute of Technology, September 1954.

ENFIELD, N. J. *Natural causes of language. Frames, biases, and cultural transmission*. Berlin: Language Science Press, 2014. ISBN 978-3-944675-50-3.

FALTÝNEK, D. *Sémiotické primitivy v konstrukci gramatik*. Disertační práce, Filozofická fakulta, Univerzita Palackého v Olomouci, 2011. Dostupné z: <http://theses.cz/id/u25w0s/>.

FEYERABEND, P. *Against method: outline of an anarchistic theory of knowledge*. Humanities Press, 1975. ISBN ISBN 0-391-00381-X.

FIEDLEROVÁ, A. Z osmého sešitu Staročeského slovníku. *Naše řeč*. 1978, 61, 4, s. 212–216. ISSN 0027-8203.

FOLEY, J. M. *How to Read an Oral Poem*. Urbana: UIP, 2002. ISBN 0-252-02770-1.

GEERAERTS, D. *Theories of Lexical Semantics*. Oxford University Press, 2010. ISBN 0198700318.

GILBERT, E. N. Capacity of a Burst-Noise Channel. *Bell System Technical Journal*. 1960, 39, 5, s. 1253–1265. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1960.tb03959.x.

GREPL, M. et al. *Příruční mluvnice češtiny*. Karolinum, 1994. ISBN 80-7106-134-4.

GRUENDLER, B. *The Development of the Arabic Scripts. From the Nabatean Era to the First Islamic Century According to Dated Texts*. Harvard Semitic Studies. Eisenbrauns, 1993. ISBN 1555407102.

HAJIČ, J. – HAJIČOVÁ, E. Syntactic tagging in the Prague Dependency Treebank. In MARCINKEVICIENE, R. – VOLZ, N. (Ed.) *Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe"*, s. 55–68, Kaunas, Lithuania, 1997. TELRI. ISBN 9986-501-09-1.

HALLIDAY, M. A. K. – HASAN, R. *Cohesion in English (English Language Series)*. English Language Series No. 9. Longman, 1976. ISBN 0582550319,9780582550315.

HASPELMATH, M. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*. 2007, 11, 1, s. 119–132. doi: 10.1515/LINGTY.2007.011.

HASPELMATH, M. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*. 2011, 45, 1, s. 31–80. doi: 10.1515/flin.2011.002.

HEMPEL, C. G. *Symposium on Sociological Theory*, The Logic of Functional Analysis, s. 271–287. New York: Harper and Row, 1959.

HRBÁČEK, J. *Úvod do studia českého jazyka*. Univerzita Karlova, 1984. ISBN 80-7066-869-5.

HURFORD, J. R. *The Origins of Meaning*. Oxford University Press, 2007. ISBN 9780199207855.

HURFORD, J. R. *The Origins of Grammar*. Oxford University Press, 2012. ISBN 978-0-19-920787-9.

ITKONEN, E. *What is Language? A study in the Philosophy of Linguistics*. Publications in General Linguistics, 8. Turku: University of Turku, 2003. ISBN 951-29-2617-2.

JOHNSON, K. Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, s. 29–54, 2004.

JUOLA, P. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*. 1998, 5, 3, s. 206–213. doi: 10.1080/09296179808590128.

KEIDAN, A. Word boundaries in Pāṇini and Avesta: a linguistic view. In *Materialy čtenij pamjati I. M. Tronskogo*, s. 145–152. Sankt–Peterburg Nestor–Istorija, 2007. ISBN 978-594047-219-3.

KÖHLER, R. System theoretical linguistics. *Theoretical Linguistics*. 1987, 14, 2–3, s. 241–258. doi: 10.1515/thli.1987.14.2-3.241.

KÖHLER, R. *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, Synergetic Linguistics, s. 760–775. New York: Walter de Gruyter, 2005. ISBN 978-3-11-019414-2.

KÖHLER, R. Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika*. 1984, 6, s. 177–183.

KOLMOGOROV, A. Three Approaches to the Quantitative Definition of Information. *Problems of Information Transmission*. 1965, 1, 1, s. 1–7. ISSN 0032-9460.

KORPEL, M. C. A. – OESCH, J. M. – PORTER, S. E. (Ed.). *Method in Unit Delimitation*. 6 / *Pericope. Scripture as Written and Read in Antiquity*. Brill, 2007. ISBN 9789004165670.

KUBÁT, P. *Nepolapitelná slova*. Diplomní práce, Filozofická fakulta, Univerzita Karlova v Praze, 2010. Dostupné z: <https://is.cuni.cz/webapps/zzp/download/120013751/>.

KUHN, T. S. *Struktura vědeckých revolucí*. OIKOYMENH, 1997. ISBN 80-86005-54-2.

KÜPPERS, B.-O. – HAHN, U. – ARTMANN, S. (Ed.). *Evolution of Semantic Systems*. Springer-Verlag Berlin Heidelberg, 2013. ISBN 978-3-642-34996-6.

LEVITIN, D. J. Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*. 1994, 56, 4, s. 414–423. doi: 10.3758/BF03206733.

LI, M. – VITÁNYI, P. *An Introduction to Kolmogorov Complexity and Its Applications*. Monographs in Computer Science. Springer New York, 2013. ISBN 9781475738605.

LI, M. – VITÁNYI, P. M. *Handbook of Theoretical Computer Science. Volume A: Algorithms and Complexity*, Kolmogorov Complexity and its Applications, s. 188–254. Elsevier; MIT Press, 1990. ISBN 0444880712.

MACHAČ, P. – ZÍKOVÁ, M. *Redukční procesy v řeči z hlediska fonetických rysů*, Studie k moderní mluvnici češtiny 5, K české fonetice a pravopisu, s. 45–68. Univerzita Palackého v Olomouci, 2013. ISBN 978-80-244-3526-8.

MACKAY, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2005. Dostupné z: <www.inference.phy.cam.ac.uk/mackay/itila/>.

MAČUTEK, J. – WIMMER, G. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*. 2013, 20, 3, s. 227–240. doi: 10.1080/09296174.2013.799912.

MANDELBROT, B. An informational theory of the statistical structure of language. *Communication theory*. 1953, 84, s. 486–502. ISSN 1468-2885.

MARTÍN, F. M. D. P. The Mirage of Morphological Complexity. In *Paper presented at the Workshop on Quantitative Measures in Morphology and Morphological Development*, UC San Diego, January 2011. Center for Human Development.

MENZERATH, P. Über einige phonetische Probleme. In *Actes du premier Congres international de linguistes*. Sijthoff Leiden, 1928.

MENZERATH, P. *Die Architektonik des deutschen Wortschatzes*. 3. F. Dümmler, 1954.

MEYER, P. Two semi-mathematical asides on Menzerath-Altmann's law. *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday*. 2007, 62, s. 449.

MILIČKA, J. Menzerath's Law: The Whole is Greater than the Sum of its Parts. *Journal of Quantitative Linguistics*. 2014, 21, 2, s. 85–99. doi: 10.1080/09296174.2014.882187.

MILIČKA, J. Rank-frequency Relation and Type-token Relation: Two Sides of the Same Coin. In IVAN OBRADOVIĆ, E. K. – KÖHLER, R. (Ed.) *Methods and Applications of Quantitative Linguistics — Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)*, s. 163–171. Academic Mind, 2013. ISBN 978-86-7466-465-0.

MILIČKA, J. Synergetic linguistics: Do we need better explanatory mechanism? *Glottotheory*. 2015, 6, 2. ISSN 2196-6907.

MITZENMACHER, M. A survey of results for deletion channels and related synchronization channels. *Probability Surveys*. 2009, 6, s. 1–33. ISSN 1549-5787.

MRŠTÍKOVÁ, K. Ruční anotace morfologické segmentace češtiny [online], 2014. Dostupné z: <http://is.muni.cz/th/383264/ff_b/>.

OCELÁK, R. "Categorical Perception" and Linguistic Categorization of Color. *Review of Philosophy and Psychology*. 2015, s. 1–16. ISSN 1878-5158. doi: 10.1007/s13164-015-0237-4.

OUDEYER, P.-Y. *Self-Organization in the Evolution of Speech*. 6 / *Studies in the Evolution of Language*. Oxford University Press, January 2006. doi: 10.1093/acprof:oso/9780199289158.001.0001.

PLATÓN. *Faidros*. Knihovna antické tradice. OIKOYMENH, 2000. Přeložil Novotný, F. ISBN 80-7298-015-7.

POPPER, S. K. R. *The Logic of Scientific Discovery*. Taylor & Francis, 2005. ISBN 0-203-99462-0.

POPPER, S. K. R. *Objective Knowledge*. Oxford university press, 1979. ISBN 0198243707.

PRUNET, J.-F. – BÉLAND, R. – IDRISSI, A. The Mental Representation of Semitic Words. *Linguistic Inquiry*. 2000, 31. doi: 10.2307/4179126.

RAPOPORT, A. Zipf's Law Re-visited. In GUITER, H. – ARAPOV, M. V. (Ed.) *Studies on Zipf's Law*. Bochum: Brockmeyer, 1982. s. 1–28. ISBN 3-88339-244-8.

RIALLAND, A. Phonological and phonetic aspects of whistled languages. *Phonology*. 8 2005, 22, s. 237–271. ISSN 1469-8188. doi: 10.1017/S0952675705000552.

ROBINSON, F. Technology and Religious Change: Islam and the Impact of Print. *Modern Asian Studies*. 2 1993, 27, s. 229–251. ISSN 1469-8099. doi: 10.1017/S0026749X00016127.

ROSENBLUETH, A. – WIENER, N. The Role of Models in Science. *Philosophy of Science*. 1945, 12, 4, s. 316–321. ISSN 0031-8248.

ROSENTHALL, S. *The Encyclopedia of Arabic Language and Linguistics*, 3, Obligatory Contour Principle, s. 461–463. Brill Academic, 2006–2009. ISBN 9004149732.

SAENGER, P. H. *Space Between Words: The Origins of Silent Reading*. Stanford University Press, 1997. ISBN 0-8047-2653-1.

SCHMIDHUBER, J. A Computer Scientist's View of Life, the Universe, and Everything. In FREKSA, C. (Ed.) *Foundations of Computer Science: Potential — Theory — Cognition*. Berlin: Springer, 1997a. s. 201–208. ISBN 978-3-540-69640-7.

SCHMIDHUBER, J. Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity & Creativity. 2007. doi: abs/0709.0674. Dostupné z: <http://arxiv.org/abs/0709.0674>.

SCHMIDHUBER, J. Low-Complexity Art. *Leonardo*. 1997b, 30. doi: 10.2307/1576418.

SCHMIDT, M. – LIPSON, H. Distilling Free-Form Natural Laws from Experimental Data. *Science*. 2009, 324, 5923, s. 81–85. doi: 10.1126/science.1165893.

SCHWENK, K. Why Snakes Have Forked Tongues. *Science*. 1994, 263, 5153, s. 1573–1577. doi: 10.1126/science.263.5153.1573.

SEDLÁČEK, J. *Al-Kitâbu. Mluvnice arabského jazyka*. Knihkupectví Fr. Řivnáče, 1898.

SHANNON, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal*. 1948, 27, 3, s. 379–423. ISSN 1538-7305.

SHEA, J. M. – WONG, T. F. *Multidimensional Codes*, 3, s. 1538–1551. Wiley Encyclopedia of Telecommunications, 2003. ISBN 978-0-471-36972-1.

SIMON, H. A. On a Class of Skew Distribution Functions. *Biometrika*. 1955, 42, 3/4, s. 425–440. ISSN 0006-3444.

SLOMAN, A. What's information, for an organism or intelligent machine? How can a machine or organism mean? In *Information and computation. Essays on scientific and philosophical understanding of foundations of information and computation*. London: Hackensack, NJ: World Scientific, 2011. s. 394–438. ISBN 978-981-4295-48-2.

STEFANOWITSCH, A. New York, Dayton (Ohio), and the Raw Frequency Fallacy. *Corpus Linguistics and Linguistic Theory*. 2005, 1, 2, s. 295–301. doi: 10.1515/cllt.2005.1.2.295.

STRUNK, W. *The Elements of Style*. Ithaca, N. Y.: Priv. print., 1918. ISBN 1-58734-060-7.

VERSTEEGH, K. *Landmarks In Linguistic Thought III: The Arabic Linguistic Tradition (History of Linguistic Thought)*. Routledge, 1997. ISBN 9780415140621.

VULANOVIĆ, R. On Measuring Language Complexity as Relative to the Conveyed Linguistic Information. *SKY Journal of Linguistics*. 2007, 20. ISSN 1456-8438.

VYSKOČILOVÁ, K. *Tvorba specializovaného korpusu banátské češtiny a jazyková analýza vybraných jevů*. Diplomní práce, Univerzita Karlova, 2014.

WARD, W. A. The Semitic Root Hwy in Ugaritic and Derived Stems in Egyptian. *Journal of Near Eastern Studies*. 1969, 28, 4, s. 265–267. ISSN 00222968.

WIT, E.-J. C. – GILLETTE, M. What is Linguistic Redundancy? Technical report, The University of Chicago, 1999. Dostupné z: <http://www.math.rug.nl/~ernst/linguistics/redundancy3.pdf>.

WOLFRAM, S. *A New Kind of Science*. Wolfram Media Inc., 2002. ISBN 1-57955-008-8.

WORTHINGTON, M. Clause grouping in Neo-Assyrian. On the evidence of the direct speech marker mā. *Orientalia*. 2006, 75, 4, s. 334–358. ISSN 00305367.

WOUTERS, A. G. Design Explanation: determining the constraints on what can be alive. *Erkenntnis*. 2007, 67, 1, s. 65–80. ISSN 1572-8420. doi: 10.1007/s10670-007-9045-2.

WOUTERS, A. G. *Explanation Without A Cause*. Disertace, Utrecht University, 1999.

WYLLYS, R. E. Empirical and Theoretical Bases of Zipf's Law. *Library Trends*. 1981, 30, 1, s. 53–64. ISSN 1559-0682.

YNGVE, V. H. *From Grammar to Science: New Foundations for General Linguistics*. John Benjamins Publishing Company, 1996. ISBN 90-272-21618.

YNGVE, V. H. – WĄSIK, Z. (Ed.). *Hard-Science Linguistics (Open Linguistics)*. Continuum, 2004. ISBN 08-264-6114-X.

YNGVE, V. H. *Linguistics as a Science*. A Midland book; MB-402. Indiana University Press, 1986. ISBN 9780253334398.

ZEMÁNEK, P. *Vývoj arabštiny*. Univerzita Karlova, Filozofická fakulta, 2007. ISBN 978-80-7308-201-7.

ZEMÁNEK, P. – MILIČKA, J. Quotations, Relevance and Time Depth: Medieval Arabic Literature in Grids and Networks. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)*, s. 17–24. Association for Computational Linguistics, 2014. ISBN 978-1-937284-88-6.

ZENIL, H. (Ed.). *A Computable Universe: Understanding & Exploring*. World Scientific, 2012. ISBN 978-981-4447-78-2.

ZUSE, K. *A Computable Universe: Understanding & Exploring*, Calculating Space. World Scientific, 2012. ISBN 978-981-4447-78-2.

Přílohy

Příloha A

Odvození vzorců

A.1 Vzorec 5.1

Pravděpodobnost, že kanál přejde ze stavu B do stavu G již po prvním bitu, tedy že $n = 1$ je roven q. Tedy:

$$f(1) = q \tag{A.1}$$

Pravděpodobnost, že kanál přejde ze stavu B do stavu G po n-tém bitu, je vyjádřena pravděpodobností, že kanál do stavu G zatím nepřešel a zároveň q:

$$f(n) = (q \cap \ 1 - \sum_{i=1}^{n-1} f(i)) \tag{A.2}$$

Což můžeme převést na rekurentní vzorec:

$$f(n) = q(1 - \sum_{i=1}^{n-1} f(i)) \tag{A.3}$$

Kteryžto výraz zjednodušíme následujícím způsobem: abychom se zbavili sumy, nejprve provedeme umělou úpravu:

$$f(n) = q(1 - f(n-1) - \sum_{i=1}^{n-2} f(i)) \tag{A.4}$$

Výraz $\sum_{i=1}^{n-2} f(i)$ si můžeme ze vzorce A.13 vyjádřit následovně:

$$\sum_{i=1}^{n-2} f(i) = 1 - \frac{f(n-1)}{q} \tag{A.5}$$

A tento vzorec následně dosadit do rovnice A.14:

$$f(n) = q(1 - f(n-1) - 1 + \frac{f(n-1)}{q}) \tag{A.6}$$

A následně upravit:

$$f(n) = f(n-1)(1-q) \tag{A.7}$$

Tento vzorec lze převést na nerekurentní formu (odpovídající vzorci 5.1):

$$\boxed{f(n) = q(1-q)^{n-1}} \tag{A.8}$$

Kterýžto vzorec můžeme dokázat matematickou indukcí, neboť:

$$f(1) = q(1-q)^{1-1} = q \tag{A.9}$$

A zároveň (dosadíme do rekurentního vzorce A.17):

$$q(1-q)^{n-1} = q(1-q)^{n-1-1}(1-q) \tag{A.10}$$

$$q(1-q)^{n-1} = q(1-q)^{n-1} \tag{A.11}$$

Q. E. D.

A.2 Vzorec 5.6

Stejně jako v příloze A.1, pravděpodobnost, že kanál přejde ze stavu B do stavu G po n-tém bitu, je vyjádřena pravděpodobností, že kanál do stavu G zatím nepřešel a zároveň q, ovšem v tomto případě tato pravděpodobnost není konstanta, ale funkce, proto ji značíme $q(n)$:

$$f(n) = (q(n) \cap 1 - \sum_{i=1}^{n-1} f(i)) \tag{A.12}$$

Což můžeme převést na rekurentní vzorec:

$$f(n) = q(n)(1 - \sum_{i=1}^{n-1} f(i)) \tag{A.13}$$

Kterýžto výraz zjednodušíme následujícím způsobem: abychom se zbavili sumy, nejprve provedeme umělou úpravu:

$$f(n) = q(n)(1 - f(n-1) - \sum_{i=1}^{n-2} f(i)) \tag{A.14}$$

Výraz $\sum_{i=1}^{n-2} f(i)$ si můžeme ze vzorce A.13 vyjádřit následovně:

$$\sum_{i=1}^{n-2} f(i) = 1 - \frac{f(n-1)}{q(n-1)} \tag{A.15}$$

A tento vzorec následně dosadit do rovnice A.14:

$$f(n) = q(n)(1 - f(n-1) - 1 + \frac{f(n-1)}{q(n-1)}) \tag{A.16}$$

A následně upravit:

$$f(n) = f(n-1)\ q(n)\ \frac{1 - q(n-1)}{q(n-1)} \tag{A.17}$$

Z tohoto vzorce vyjdeme. Nyní pomocí ekvivalentních úprav vyjádříme $q(n)$:

$$q(n) = \frac{f(n)q(n-1)}{f(n-1)(1 - q(n-1))} \tag{A.18}$$

Za funkci $f(n)$ dosadíme normální rozdělení, čímž získáme rovnici:

$$q(n) = \frac{\frac{e^{-\frac{(n-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}}{\frac{e^{-\frac{(n-1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}} \frac{q(n-1)}{1 - q(n-1)} \tag{A.19}$$

Tu zjednodušíme[1] na:

$$q(n) = \frac{q(n-1)\, e^{\frac{2\mu-2n+1}{2\sigma^2}}}{1-q(n-1)} \tag{A.20}$$

Aby rozdělení délek dávek bylo rovno normálnímu rozdělení, je tedy nutno, aby parametr q_i byl určován následující rekurentní funkcí:

$$q_1 = f_1 = \frac{e^{-\frac{(1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{A.21}$$

$$q_i = \frac{q_{i-1}\, e^{\frac{\mu-i+\frac{1}{2}}{\sigma^2}}}{1-q_{i-1}} \tag{A.22}$$

Tuto funkci můžeme využít při generování šumu s normální distribucí dávek.

[1] $$q(n)\frac{1-q(n-1)}{q(n-1)} = \frac{e^{-\frac{(n-\mu)^2}{2\sigma^2}}}{e^{-\frac{(n-1-\mu)^2}{2\sigma^2}}} = e^{\frac{(n-1-\mu)^2}{2\sigma^2}-\frac{(n-\mu)^2}{2\sigma^2}} = e^{\frac{\mu^2-2\mu n+2\mu+n^2-2n+1-\mu^2+2\mu n-n^2}{2\sigma^2}}$$

Příloha B

Rozbor hraničních hlásek

B.1 Arabština

Zmínka o tom, že v arabštině existuje třída hlásek, které jsou kořenové, tedy nemohou se vyskytovat v afixech (kterým se říká *az-zawāʾid*, viz poznámku 3 na straně 108), si zasluhuje krátké pozastavení — ostatně zasluhovala by si i rozsáhlejší studii, na kterou zde ovšem není prostor.

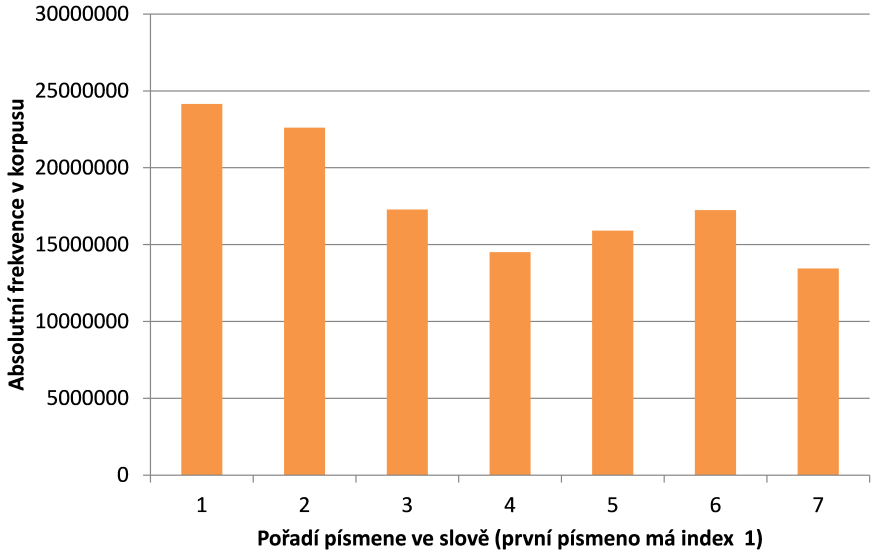
Otázka, kterou zde budeme řešit, zní: mají tyto hlásky skutečně rozdílnou distribuci na hranicích slova a uprostřed slova? Bohužel nemáme k dispozici plně vokalizované texty,[1] takže zjištěná měření se budou týkat grafické podoby textu, přičemž se budeme soustředit pouze na souhlásky. I když v arabštině zápis textu s fonologickou realizací značně souvisí, raději budeme výsledky interpretovat spíše jako pilotní studii.

Graf na obrázku B.1 ukazuje, že písmena z třídy *az-zawāʾid* se sice koncentrují na okraji slova, nicméně ona koncentrace není nijak přesvědčivá — tedy že se velmi často vyskytují i v kořeni slova nebo jako infixy.

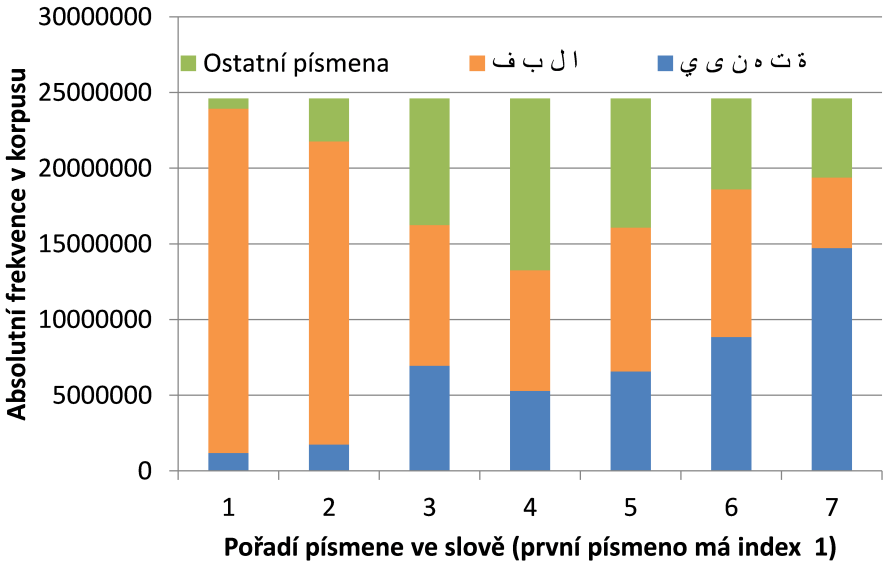
Přesvědčivějších výsledků dosáhneme, když tuto třídu rozdělíme na poloviny: na písmena, která se koncentrují na začátku slova, a na ta, která mají tendenci být na jeho konci. Dále vytřídíme písmena, která se nekoncentrují ani na jednom konci slov. Na začátku slova se koncentrují *ʾalif, bāʾ, lām, wāw* a *fāʾ*, zatímco na konci slova najdeme *tāʾ marbūṭu, tāʾ, nūn, hāʾ, yāʾ* a *ʾalif maksūru*.[2] Zvláštní postavení má *mīm*, které v kratších slovech obsazuje zejména třetí pozici (jakožto předpona *mu-* a *ma-*) a pozici na konci slova jako součást přípojných zájmen, zatímco v delších se distribuuje spíše jako kořenová hláska.

[1] Měření jsem provedl na korpsu CLAUDia, bližší popis tohoto korpusu najdete v (Zemánek – Milička, 2014).

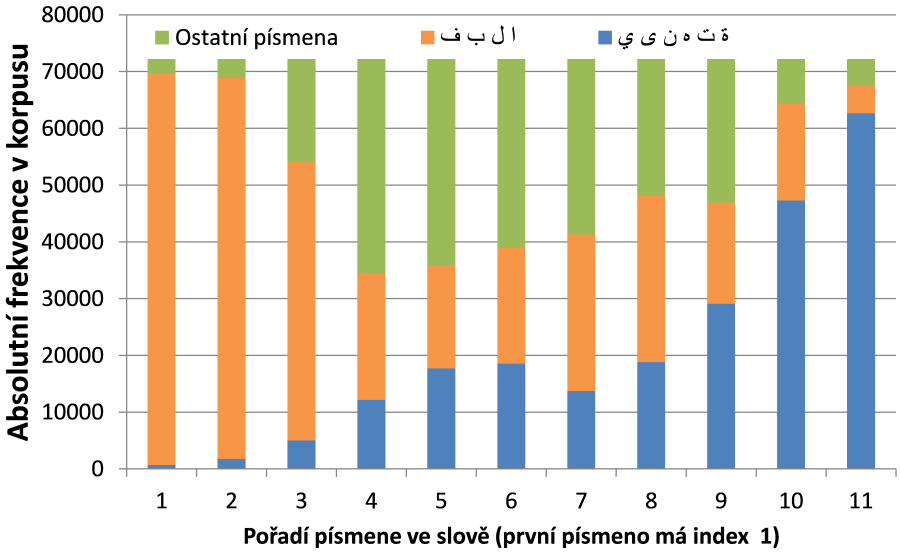
[2] Pro čtenáře obeznámeného s arabským imperfektem může být překvapivé, že *tāʾ* a *yāʾ* se na začátku slova vyskytuje málokdy.



Obrázek B.1: Frekvence písmen reprezentující *az-zawāʾid* (*yāʾ, tāʾ, sīn, mīm, nūn, wāʾ, ʾalif, bāʾ, fāʾ, lām, kāf*) na jednotlivých pozicích ve všech sedmipísmenných slovech.



Obrázek B.2: Frekvence písmen na jednotlivých pozicích ve všech sedmipísmenných slovech. Oranžová: písmena, která se s větší pravděpodobností vyskytují na začátku slova (*ʾalif, bāʾ, lām, wāw, mīm* a *fāʾ*); modrá: písmena, která se s větší pravděpodobností vyskytují na konci slova (*tāʾ marbūṭa, tāʾ, nūn, hāʾ, yāʾ* a *ʾalif maksūra*); zelená: ostatní písmena.



Obrázek B.3: Frekvence písmen na jednotlivých pozicích ve všech jedenáctipísmenných slovech. Oranžová: písmena, která se s větší pravděpodobností vyskytují na začátku slova (*ʾalif, bāʾ, lām, wāw* a *fāʾ*); modrá: písmena, která se s větší pravděpodobností vyskytují na konci slova (*tāʾ marbūṭa, tāʾ, nūn, hāʾ, yāʾ* a *ʾalif maksūra*); zelená: ostatní písmena.

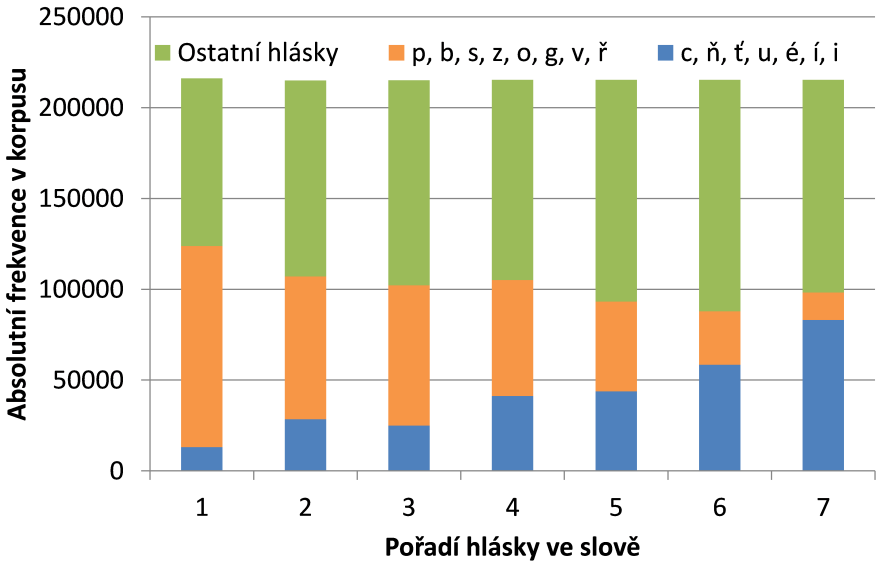
Grafy B.2 a B.3 jsou příkladem toho, jak se tyto skupiny písmen skutečně distribuují, a ukazují, že toto rozdělení má smysl. Tato předběžná hypotéza však potřebuje další testování.

B.2 Čeština

Narozdíl od arabštiny češtinu můžeme poměrně jednoduše automaticky převést na fonologický přepis (nebo tento přepis alespoň aproximovat), budeme tedy nadále mluvit o českých hláskách (včetně samohlásek). Měření jsem provedl na korpusu děl Karla Čapka.

Po změření četnosti hlásek na různých pozicích se jeví jako perspektivní množina hlásek *p, b, s, z, o, g, v* a *ř*, jejíž prvky se objevují zejména na začátcích slov, a množina *c, ň, ť, u, é, í* a *i*, jejíž prvky se objevují zejména na koncích slov. Jak vidíme na grafu B.4, výsledek není tak přesvědčivý jako v arabštině, což ale může být dáno nikoli rysy jazyka, ale tím, že jde o měření svým způsobem jiného jevu (písmena bez krátkých samohlásek versus kompletní hláskový inventář). Navíc též délka slov je těžko porovnatelná (i v arabštině byl výsledek méně přesvědčivý pro kratší slova a pro delší česká slova je použitý korpus příliš krátký).

Zajímavé byly výsledky pro hlásky *e* a *f*, které obsazují zejména pozice na začátku a na konci slova (*f* jakožto spodobené *v*), podobně zajímavě se chovají hlásky *č*, *n* a *ú*.



Obrázek B.4: Frekvence písmen na jednotlivých pozicích ve všech sedmihláskových slovech. Oranžová: hlásky, které se s větší pravděpodobností vyskytují na začátku slova; modrá: hlásky, které se s větší pravděpodobností vyskytují na konci slova; zelená: ostatní hlásky.

* * *

V každém případě můžeme uzavřít tvrzením, že distribuce hlásek rozhodně není ve slově rovnoměrná.

Příloha C

Podrobnosti anotace a ukázka anotovaného textu

C.1 Český korpus

Tagování českého textu probíhalo následujícím způsobem: nejprve jsem z textu extrahoval všechna různá slova (tedy seznam slovních typů). Tento seznam (každé slovo na novém řádku) následně Zuzana Komrsková morfematicky analyzovala, přičemž dodržovala formalismus, kdy se před každý morf předřáží značka, která udává jeho kvalitu.

Předpona	žádná značka
Kořen	§
Přípona	-
Koncovka	=

Tabulka C.1: Seznam značek na označování morfů.

Navíc byla před každé slovo předražena značka roztřiďující je na autosémantika, synsémantika a vlastní jména.

Autosémantické slovo	žádná značka
Synsémantické slovo	;
Vlastní jméno	€

Tabulka C.2: Seznam značek na označování slov.

Díky poměrně nízké ambiguitě češtiny bylo možné následně podle morfematicky segmentovaného slovníku automaticky segmentovat přímo text. Následně jsem ručně

označil konce vět a souvětí, přičemž jsem bral v potaz vnořené vedlejší věty, například „Muž, který takto mluvil, stál se vzpřímenou hlavou“ chápeme jako dvě věty, z nichž jedna má 5 slov a druhá 3.

Program Menzerath.exe (též je součástí balíčku, který společně se segmentovaným slovníkem i textem naleznete na adrese `http://milicka.cz/disertace.zip`) s tím počítá a dokáže měřit počty segmentů rekurzivně. V tabulce C.3 najdete značky, které ukončují segmenty na jednotlivých úrovních.

Úroveň hlásek	žádná značka
Úroveň morfů	-
Úroveň slov	mezera
Úroveň vět	+
Úroveň souvětí	.
Úroveň odstavců	§
Značky pro vloženou větu	{ }

Tabulka C.3: Seznam značek oddělujících segmenty.

a	;§a	se	;§se	na	;§na
krysař	§krys-ař	v	;§v	ale	;§ale
to	;§t=o	je	;§je	jeho	;§j=eho
že	;§že	hammeln	€§hammeln	agnes	€§agnes
by	;§by	jako	;§jako	bylo	;§by-l=o
co	;§co	z	;§z	o	;§o
s	;§s	tak	;§tak	byl	;§by-l
do	;§do	k	;§k	ho	;§ho
za	;§za	jsem	;§js=em	jak	;§jak
však	;§však	li	;§li	ještě	;§ještě
nebylo	;ne§by-l=o	mu	;§mu	její	;§j-ej=í
který	;§kter=ý	jörgen	€§jörgen	si	;§si
krysaře	§krys-ař=e	není	;§nen-í	po	;§po
krysaři	§krys-ař=i	také	;§tak-é	u	;§u
vše	§vš=e	ani	;§ani	byla	;§by-l=a
ji	;§j=i	města	§měst=a	oči	§oč=i
mnoho	§mnoh-o	ve	;§ve	pouze	;§pouze
srdce	§srdc=e	své	;§sv=é	kristián	€§kristián
něco	§ně§co	píšťala	§píšťal=a	smích	§smích
jste	;§js=te	nad	;§nad	sepp	€§sepp
slova	§slov=a	svou	;§sv=ou	aby	;§a§by
už	;§už	krysy	§krys=y	mi	;§mi
nic	§n=ic	pak	§pak	tu	§t=u
zde	§zde	před	;§před	pro	;§pro
kde	;§kde	než	;§než	nyní	§nyní
příliš	§příliš	vám	;§v=ám	které	;§kter=é
mohl	§moh-l	nikdy	ni§kdy	zdálo	§zdá-l=o
den	§den	i	;§i	kdo	;§kdo
píšťaly	§píšťal=y	přece	;§přece	byly	;§by-l=y
daleko	§dalek-o	snad	;§snad	strumm	€§strumm
tiše	§tiš-e	třeba	;§třeba	člověka	§člověk=a
když	;§když	která	;§kter=á	měl	§mě-l
náhle	§náhl-e	ní	;§ní	až	;§až
koppel	€§koppel	magister	§magister	ne	§ne
nikdo	;ni§kdo	šel	§še-l	ticho	§tich=o
dítě	§dít=ě	krysařova	§krys-ař-ov=a	možno	§možn-o
muž	§muž	opět	§opět	ruce	§ruc=e

a- vaš-e- jmén-o- +.§ne-jmen-uj-i- se- +js-em- ni-kdo- +.js-em- hůř- než- ni-kdo- +js-em- krys-ař- +.§muž- {kter-ý- tak-to- mluv-i-l- +}stá-l- se- vz-přím-en-ou- hlav-ou- před- vrat-y- dom-u- +v- n-iX-ž- se- do- sou-mrak-u- bjel-a-l-a- žen-sk-á- postav-a- +.hled-ě-l- na- ni- sv-ýma- temn-ýma- pátr-a-v-ýma- oč-ima- +.by-l- vysok-ý- a- štíhl-ý- štíhl-ejš-í- ještě- ve- sv-ém- při-léh-a-v-ém- samet-ov-ém- kabát-c-i- a- v- úzk-ýX-noh-avic-íX- +.j-eho- ruc-e- by-l-y- drob-n-é- a- jemn-é- jako- ruc-e- pan-í- +.ne-mňe-l- při- s-obje- ani- zbran-ě- ani- hol-e- +a-č- se- zdá-l-o- že- při-Xáz-í- z-dal-ek-a- po-cest-áX- +j-e-ž- vždy- jist-y- ne-by-l-y- +.s-vír-a-l- za-t-o- ně-co- dlouh-ého- +co- bud-i-l-o- z-vjed-av-ost- žen-y- +s- kter-ou- hovoř-i-l- +.by-l-a- t-o- píšťal-a- ciz-o-kraj-né- prác-e- +jak-é- dosud- ne-vid-ě-l-a- +.§krys-ař- +za-smá-l-a- se- žen-a- ve- dveř-íX- +.při-Xáz-í-te- včas- do- hameln- +.nen-í- t-u- krys-ař-e- +za-t-o- krys- je- t-u- mnoh-o- +.vy-svjetl-e-te- mi- krys-ař-i- +od-kud- se- ber-ou- krys-y- +.ne-bý-va-l-o-j-iX- dřív-e- +jak- n-ám- řík-aj-í- +.je- prav-d-a- +konč-i-l-a- s- ú-smně-v-em- +stař-í-lid-é- mín-í- +že- svjet- je- stále- horš-í- +.§krys-ař- po-krč-i-l- ramen-y- +.§od-kud- js-ou- +ne-ví-m- +.js-ou- však- v- každ-ém- z- vaš-iX- dom-ů- +.hlod-aj-í- bez- u-stá-n-í- +hlod-aj-í- dol-e- ve- sklep-íX- +hlod-aj-í- tam- +kde- j-iX- ne-vid-í-me- +.sta-n-ou- se- však- do-těr-n-ými- +a- stoup-aj-í- pak- výš-e- +.připrav-uj-e-te- host-in-u- svatb-u- kŕt-i-n-y- +co-ž- já- ví-m- +.před-stav-te- si- +že- se- náhl-e- při- host-in-ě- z-jeví- krys-y- s- dlouh-ými- boltc-i- a- dlouh-ými- knír-y- +.t-en-t-o- z-jev- kaz-í- Xuť- +Xáp-e-te- přece- +.§ano- +smá-l-a- se- žen-a- ve- dveř-íX- +.při- svatb-je- kateřin-in-ě- ob-jev-i-l-a- se- po-řád-n-á- krys-a- +.ženiX- by-l- bled-ý- jako- stěn-a- +a- kateřin-a- pad-l-a- do- mdlob- +.lid-é- ne-s-nes-ou- n-ic- tak- mál-o- jako- t-o- +co- j-im- kaz-í- Xuť- +od-hodl-a-jí- se- pak- za-vol-a-t- krys-ař-e- +.§připrav-uj-e-te- svatb-u- nebo-kŕt-i-n-y- +.o-táz-a-l- se- krys-ař- náhl-e- bez- pře-Xod-u- +.§žen-a- ve- dveř-íX- se- za-smá-l-a- hlas-it-ě- +.§js-te- ciz-in-ec- +je- zřejm-o- +že- js-te- ciz-in-ec- +.ne-js-em- v-dá-n-a- krys-ař-i- +.krys-ař- se- u-klon-i-l- +.§t-o- ne-vad-í- +ni-jak- t-o- ne-vad-í- +.nuže- po-vol-a-jí- krys-ař-e- +.krys-ař- písk-á- +a- písk-á- +až- vy-ved-e- vš-eX-n-u- havjeť- z- j-ej-íX- s-krýš-í- +.jd-e- za- n-ím- jako- o-mám-en-á- +.za-ved-e- j-i- do-řek-y- do- rýn-a- dunaj-e- havol-y- veser-y- +.a- dům- je- prost-ý- krys- +.§krys-ař- opjet- se- u-klon-i-l- +a- j-eho- hlas- za-Xvje-l- se- ně-jak-ou- elegi-í- +.žen-a- mlč-e-l-a- po-hráv-aj-íc- si- vjetv-ičk-ou- jasmín-u- +.ale- je- li- t-omu- tak- +ne-po-mysl-í- ni-kdo- na- krys-ař-e- +.krys-ař- ciz-in-k-o- je- muž- +kter-ý- ne-zůst-á-v-á- +ale- jd-e- +.lid-é- vid-í- ho- rád-i- při-Xáz-e-t- +.ale- od-Xáz-e-ti- ho- vid-í- mnoh-em- rad-ěji- +.§o-pravd-u- +řek-l-a- pouze- +.zně-l-o- t-o- jako- po-vz-buz-en-í- +a- snad- t-o- ani- po-vz-buz-en-í- ne-by-l-o- +.ale- krys-ař- Xáp-a-l- t-o- tak- +.bled-é- j-eho- tvář-e- se- z-barv-i-l-y- +by-l-a- by- t-o- moh-l-a- po-zor-ova-t- +ne-bý-ti- temn-a- +.cít-ím- t-o- ciz-in-k-o- +.lid-é- ne-mil-uj-í- krys-ař-e- +boj-í- se- ho- pouze- +.děvč-e- za-smá-l-o- se- opjet- +.a- č-ím- t-o- +že- jd-ou- krys-y- tak- slep-je- za- v-ámi- krys-ař-i- +. o-živ-l-a- t-ím-to- po-hyb-em- +. z-vjed-av-je- na- mluv-č-ího- i- na- píšťal-u- +.

C.2 Arabský korpus

Vzhledem k vysoké ambiguitě arabských slov a k tomu, že pro krátký text není velký rozdíl mezi počtem typů a tokenů, rozhodl jsem se, že morfematickému rozkladu podrobím přímo text a nikoli seznam slov jako v případě češtiny. Také jsem upustil od přímého označování hranic morfů, neboť arabská morfologie má nonkonkatenativní charakter (opět odkazuji na poznámku 8 na straně 36, kde se s ní můžete seznámit blíže) a morfémy jsou do sebe přímo vnořené (zejména kořen a to, co nazýváme *šablonou* (k arabské morfologii viz poznámku 8 na straně 36), tedy v semitistickém žargonu *kmen*), což je pro ruční tagování velmi pracné a nepřehledné. Místo toho třídím morfy do typických skupin a udávám pro každé jednotlivé slovo jejich počet (označení skupin naleznete v tabulce C.4).

Konektivní partikule *wa-* a *fa-*	W
Se jmény spojované předložky *bi-* a *li-*	B
Určitý člen (*al-*)	L
Kořen	KO
Kmen („šablona“)	KM
Slovesné afixy	VA
Morfém označující ženský rod u jmen	F
Jmenné koncovky	NA
(Přípojná) zájmena	Z
Ostatní (předpona *ma-* apod.)	O

Tabulka C.4: Vysvětlení zkratek v popiscích sloupců.

Třídění do skupin je více méně arbitrární a jeho hlavním úkolem je práci systematizovat, zpřehlednit a usnadnit její kontrolu, pro naši potřebu je důležitý konečný počet morfů, proto nezáleží na tom, který morf skončil v jaké přihrádce. Nicméně ono rozdělení je, myslím, dostatečně precizní na to, aby pomohlo například určit, které slovo je sloveso (to, které obsahuje morf *slovesné afixy*) a co je jméno určené členem nebo přípojným zájmenem.

Za slova obsahující kořen a kmen považuji taková, u nichž můžeme kořen použít k derivaci jiných slov za použití jiných kmenů a kmen můžeme použít k derivaci jiných slov za použití jiných kořenů. Takže například předložka *fi* je chápána jako jednomorfová (a morf je zařazen do kategorie *Ostatní*), zatímco předložka *bayna* je chápána jako výsledek interakce kořene *byn* (který se vyskytuje například ve slovese *tabāyana*) a šablony $X_1 \alpha X_2 X_3$ (podle které je utvořeno například slovo *ṭayr*). Ovšem přestože má tato předložka formu jména, neuznáváme jí právo na jmennou koncovku, třebaže její koncové *-a* bylo nejspíš původně jmennou akuzativní koncovkou. Samozřejmě si

dokáži představit spoustu jiných přístupů, které by segmentaci provedly úplně jinak, a bylo by zajímavé sledovat, jak by si s tímto úkolem poradili rodilí mluvčí.

Do značné míry arbitrární je i zvolená vokalizace: přepisuji ji přesně podle v textu zmíněné redakční úpravy, ovšem ani tak není úkol snadný, například slova *law* a *intahaynā* se ve vzájemném spojení čtou dohromady, přičemž se vynechává počáteční hamza u druhého zmíněného slova (takže *fa-law intahaynā* se spisovně čte *fa-lawintahaynā*), což se tradičně zapisuje jako *fa-lawi ʾntahaynā*, ovšem těžko můžeme jednoznačně určit, ke kterému z těchto dvou slov ona konektivní samohláska skutečně patří. Proto by bylo vhodné vyzkoušet, jestli bychom dosáhli stejných nebo obdobných výsledků, kdybychom místo recitátorské vokalizace *fa-lawi ʾntahaynā* používali *fa-law intahaynā*. Nicméně po krátké úvaze jsem se rozhodl tímto práci nezatěžovat, neboť různých variant segmentace a vokalizace je v podstatě neomezené množství, takže je to spíše námět na samostatnou studii.

Segmentaci na úrovni vět jsem se snažil provést tak, aby byla pokud možno srovnatelná se segmentací českého textu. Podobně jako v češtině, i v arabském textu jsem dával pozor na vnořené vedlejší věty, takže například úsek textu *fa-ḥaddiṯ-nī, ʾin raʾayta, ʿan iẖwāni ʾs-safāʾi* je chápán jako dvě věty, z nichž jedna obsahuje čtyři slova a druhá má slova dvě.

	W	B	L	KO	KM	VA	F	NA	Z	O	# hlásek	# morfů
fa-ḥaddiṯ-nī	1			1	1	1			1		10	5
,												
ʾin										1	3	1
raʾayta				1	1	1					7	3
,												
ʿan										1	3	1
iẖwāni				1	1			1			6	3
ʾs-safāʾi			1	1	1			1			7	4
kayfa				1	1						5	2
yubtadaʾa				1	1	1					9	3
tawāṣulu-hum				1	1			1	1		11	4
wa-yastamtiʿu	1			1	1	1					12	4
baʿḍu-hum				1	1			1	1		8	4
bi-baʿḍin		1		1	1			1			8	4
.												
qāla				1	1	1					4	3
ʾl-faylasūfu			1	1				1			10	3
:												
ʾinna										1	5	1
ʿāqila			1	1	1			1			6	4
lā										1	2	1
yaʿdilu				1	1	1					7	3
bi-ʾl-ʾiẖwāni		1	1	1	1			1			10	5
šayʾan				1	1			1			6	3
.												
fa-ʾl-ʾiẖwānu	1		1	1	1			1			10	5
humu									1		4	1
l-ʾaʿwānu			1	1	1			1			8	4
ʿala										1	4	1
l-ẖayri			1	1	1			1			6	4
kulli-hi									1	1	7	2
,												
wa-ʾl-muʾāsifūna	1		1	1	1			1			13	5
ʿinda										1	5	1
māḍā										1	4	1
yanūbu				1	1	1					6	3
mina										1	4	1
ʾl-makrūhi			1	1	1					1	8	4
.												