SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta elektrotechniky a informatiky

Evidenčné číslo: FEI-12307-794

Štatistická analýza textov pre potreby kryptoanalýzy

Dizertačná práca

2016 **Mgr. Jozef Kollár**

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta elektrotechniky a informatiky

Evidenčné číslo: FEI-12307-794

Štatistická analýza textov pre potreby kryptoanalýzy

Dizertačná práca

Študijný program: Aplikovaná informatika
Študijný odbor: 9.2.9. aplikovaná informatika
Školiace pracovisko: Ústav informatiky a matematiky
Školiteľ: prof. RNDr. Otokar Grošek, PhD.

Bratislava 2016 **Mgr. Jozef Kollár**

Slovenská technická univerzita v Bratislave **Fakulta elektrotechniky a informatiky**
Ústav informatiky a matematiky

ZADANIE DIZERTAČNEJ PRÁCE

Autor práce:	Mgr. Jozef Kollár
Študijný program:	Aplikovaná informatika
Študijný odbor:	9.2.9. aplikovaná informatika
Evidenčné číslo:	FEI-12307-794
ID študenta:	794
Vedúci práce:	prof. RNDr. Otokar Grošek, PhD.
Miesto vypracovania:	Ústav informatiky a matematiky FEI STU

Názov práce: **Štatistická analýza textov pre potreby kryptoanalýzy**

Špecifikácia zadania: Kvantitatívna lingvistika využíva viaceré postupy, ktoré doteraz neboli uvažované pre potreby kryptoanalýzy klasických šifier. Cieľom práce je preskúmať možnosti ich využitia a porovnať dosiahnuté výsledky so známymi metódami. Získané výsledky aplikujte pri kryptoanalýze niektorej doteraz nerozlúštenej klasickej šifry.

Úlohy

1. Analyzujte textové štatistiky využiteľné v kryptoanalýze.
2. Pokúste sa o nájdenie lingvistických štatistík charakterizujúcich smer čítania.
3. Praktická ukážka kryptoanalýzy doteraz nerozlúštenej klasickej šifry.

Literatúra:

1. G. Wimmer, G. Altmann, L. Hřebíček, S. Ondrejovič, S. Wimmerová: Úvod do analýzy textov. Veda, 2003, Bratislava
2. L. Kubáček: Confidence Limits for Proportions of Linguistic Entities . Journal of Quantitative Linguistics, Vol.1, No. 1, 1994.
3. O. Grošek, P. Zajac: Automated Cryptanalysis. Encyclopedia of Artificial Intelligence, 2009
4. O. Grošek, P. Zajac: Automated Cryptanalysis of Clasical Ciphers . Encyclopedia of Artificial Intelligence, 2009

Dátum zadania: 06. 12. 2010

Dátum odovzdania: 31. 08. 2016

Mgr. Jozef Kollár
riešiteľ

prof. RNDr. Otokar Grošek, PhD.
vedúci pracoviska

prof. RNDr. Otokar Grošek, PhD.
garant študijného programu

Prehlasujem, že prácu som vypracoval samostatne, len s pomocou literatúry uvedenej v zozname literatúry a konzultácií so školiteľom.

V Bratislave, 30. 8. 2016 .....

Týmto by som sa chcel poďakovať Prof. Grošekovi za jeho odbornú pomoc, poskytnuté materiály, množstvo rád a nápadov, ako aj za jeho ochotu a trpezlivosť pri vedení práce.

Štatistická analýza textov pre potreby kryptoanalýzy

Abstrakt

Cieľom tejto práce je hľadať a skúmať nové metódy kvantitatívnej lingvistiky a ich aplikácie v kryptoanalýze. Samotná práca je rozdelená na tri kapitoly, ktoré zodpovedajú úloham zadania práce.

Prvá kapitola je, okrem krátkeho historického úvodu, venovaná štatistickým indexom a metódam využívaným v kryptoanalýze. Štatistické indexy uvádzané v tejto kapitole nie sú originálne, všetky už sú známe a v nejakej podobe publikované. Niektoré z uvedených výsledkov, ako sú napríklad tabuľka frekvencií znakov 1.1, odsek o hĺbke závislosti znakov (1.3.4) a odsek o priemernej dĺžke slov (1.3.5), sú však vlastné a boli robené na veľkých vzorkách textov. Jedine spôsob určovania miery závislosti dvoch znakov, v odseku o hĺbke závislosti znakov, je podľa našich vedomostí originálny.

V druhej kapitole sme hľadali metódu, pomocou ktorej by sa dal určiť smer čítania neznámeho textu. Pôvod tohto problému je výsostne praktický a vyplynul z reálnej potreby určiť smer čítania textu v Rohonczi kódexe. Rohonczi kódex je historický dokument v rozsahu približne 400 strán, ktorý obsahuje text, o ktorom sa nič nevie. Nepoznáme jazyk, nevieme či a akým spôsobom je text zašifrovaný a dokonca ani nevieme, či sa jedná o zmysluplný text, alebo len nejaký historický podvrh. Predpokladajme, že sa jedná o zmysluplný text. Ak by sa kryptológovia chceli pokúsiť o jeho lúštenie, potrebovali by predovšetkým vedieť, ktorým smerom je tento text písaný, aby mali správnu nádväznosť riadkov a mohli text ďalej analyzovať. Preto sme sa na vzorkách otvoreného textu v rôznych jazykoch pokúšali nájsť spôsob určenia smeru čítania textu bez toho, aby sa využívala štruktúra a vlastnosti jazyka.

Tretia kapitola je popisom úspešného útoku na sovietsku šifru VIC, ktorá je v literatúre označovaná za veľmi komplexnú a „neprelomiteľnú“ šifru ([22], str. 670–671). O šifre VIC existuje viacero publikácií, spomínaných v texte, avšak žiadna z nich sa nezaoberala jej analýzou. Rovnako sa žiadna nám známa publikácia nezaoberala útokom na šifru VIC. Nami popísaný útok je prakticky realizovateľný v reálnom čase.

On Statistical Language analysis and Cryptanalysis

Abstract

The aim of this thesis is investigation of new methods of quantitative linguistics and their application in cryptanalysis. The material is divided into three chapters.

The first chapter, asside from a short historical introduction, is dedicated to statistical indices and their applications in cryptanalysis. The indices introduced in this chapter are not new and have already been known and published in some form. Some of the results related to these indices, however, such as the letter frequency table 1.1, the section 1.3.4 about depth of letter dependency, and the section 1.3.5 on average word length, are original and were processed on big text samples. As regards methods, only the one of determining the measure of interdependence of two letters is original to the best of our knowledge.

In the second chapter we considered the problem of developing a method of determinig the text reading direction on an unknown text. The origin of this problem is purely practical and resulted from the need to determine the text reading direction in the Rohonczi codex. This codex is a historical manuscript consisting of about 400 pages containing a text on which nothing is known. We do not know the language of the text, we do not know whether and how the text is encrypted and we do not even know whether it is meaningful or just a hoax. Assuming meaningfulness, if cryptologists wanted to crack the cipher they would need to know in the first place the writing direction to get the correct line-flow to analyse the text. Therefore we tried, on open text samples in various languages, to find methods of determining the text direction without using the structure and properties of the language.

The third chapter describes a succesful attack against the soviet VIC cipher, which is mentioned as the most complex and "non-crackable" hand cipher in the literature ([22], pp. 670–671). There are several publications handling the VIC cipher, but none is dedicated to the cipher cryptanalysis and to the attack at the cipher. The attack at the VIC cipher described here is a real-time one.

Obsah

Niektoré základné definície a pojmy používané v práci

- **TSA**=„Telegraph Standard Alphabet“. Je to štandardná telegrafná abeceda. Táto abeceda má dve alternatívy. Jedna pozostáva len z 26 písmen anglickej abecedy `A...Z` a druhá, rozšírená, obsahuje aj medzeru a má 27 znakov. Z historicko-technických dôvodov sa aj iné ako anglické texty pred šifrovaním často zapisovali pomocou TSA abecedy.
- **Matematická štatistika**[1] je veda skúmajúca javy na rozsiahlom súbore dát a hľadajúca tie vlastnosti javov, ktoré sa prejavujú práve na rozsiahlych súboroch dát. Zjednodušene povedané, štatistika sa zaoberá vyhodnocovaním experimentálne získaných dát.
- **Náhodný výber** je n-tica $\mathbb{X} = (X_1, X_2, ..., X_n)$ nezávislých náhodných veličín s rovnakým rozdelením daným distribučnou funkciou $F_X(x)$ ([46], str. 156).
- **Výberová štatistika** je (vhodná) funkcia $T(\mathbb{X})$ náhodného výberu. Je to vlastne náhodná veličina ([46], str. 163). Výberovú štatistiku budeme v texte označovať aj ako **štatistický index**.

Poznámka: Ak v texte budeme písať o *náhodnej premennej*, tak sa vždy bude jednať o diskrétnu náhodnú premennú s konečným počtom nadobúdaných hodnôt.

[1] [46], str. 139, alebo [1], alebo prednášky Prof. Grošeka.

Príklady z kryptoanalýzy:

- Skúmame rôzne texty $t_1, ..., t_n$ nad TSA abecedou. Všetky tieto texty sú písané v tom istom jazyku. Relatívnu početnosť znaku `A` v texte t_i si označíme p_{Ai}. Výberovou štatistikou potom môže byť napríklad aritmetický priemer relatívnych početností znaku `A` v skúmaných textoch:

$$T_A = \frac{1}{n} \cdot \sum_{i=1}^{n} p_{Ai}$$

  Jej stredná hodnota je $E(T_A) = p_A$, kde p_A je pravdepodobnosť výskytu znaku `A` v príslušnom jazyku. V prípade T_A sa jedná o nevychýlený odhad tejto pravdepodobnosti.

- Skúmame vzorku prvých n znakov nejakého dlhého textu nad TSA abecedou. Nech náhodná premenná n_i vyjadruje početnosť znaku i v skúmanej vzorke. Samozrejme bude platiť $\sum_{i=1}^{26} n_i = n$. Potom výberovou štatistikou môže byť napríklad index koincidencie[2]:

$$IC = \sum_{i=1}^{26} \frac{n_i(n_i - 1)}{n(n-1)} .$$

  V prípade IC sa jedná o asymptoticky nevychýlený odhad pravdepodobnosti toho, že ak z veľkej vzorky textu náhodne vyberieme dva znaky, tak tieto budú rovnaké[3].

[2]Bude spomínaný podrobnejšie v ďalšom texte.
[3]Viď strana 33 a ďalej.

Kapitola 1

Úvod

Táto kapitola obsahuje stručné historické pozadie skúmanej problematiky a popis úlohy štatistickej analýzy textov a je zhrnutím už známych výsledkov a metód v kryptoanalýze. Vlastné výsledky budú uvedené až v ďalších dvoch kapitolách. V celej práci sa v použitých príkladoch obmedzíme na klasické, t.j. z prevažnej časti ručné, šifry. Pri moderných šifrách sú úlohy štatistickej analýzy textu rovnaké ako pri klasických šifrách, avšak metódy sú odlišné. Jednotlivými štatistickými indexami, ich vlastnosťami a ich využitím pri lúštení šifier sa budeme zaoberať v ďalších častiach.

1.1 Historické pozadie

Kryptografia je takmer rovnako stará ako písmo. Najstaršie dochované pamiatky dokladajúce utajovanie informácií majú zhruba 3500 rokov a pochádzajú z Mezopotámie. Dá sa predpokladať, že čoskoro potom ako začali ľudia informácie utajovať, našli sa aj takí, čo sa tieto tajomstvá snažili odhaliť. Tým vznikli dva odbory ľudskej činnosti, ktoré až o tisícročia neskôr dostali pomenovanie „kryptografia“, „kryptoanalýza“ a spoločne tvoria vedný odbor nazývaný „kryptológia“. Zo staroveku sa žiaľ nezachovali takmer žiadne informácie o spôsoboch lúštenia vtedajších šifier. Jedinou (nám známou) výnimkou je spartské Skythalé. S pravdepodobnosťou hraničiacou s istotou ale môžme predpokladať, že vtedajšie metódy lúštenia nemali nič spoločné s analýzou textu a štatistikou.

Prvé dochované informácie o lúštení šifier pochádzajú až z 8. storočia nášho letopočtu. Podľa informácií z Kahnovej knihy ([22], str. 97) napísal arabský autor Abū 'Abd al-Rahmām al-Khalil ibn Ahmad ibn Tammām al Farāhidi al-Zadi al Yahmadi, žijúci približne v rokoch 718/719 až 786/791, knihu „Kitāb al-mu'ammā“, čo v preklade znamená „Kniha tajných jazykov“.

V tejto knihe popísal spôsob akým lúštil šifry. Z nášho pohľadu je na jeho metóde zaujímavý fakt, že bola založená na relatívnej početnosti znakov abecedy. Trvalo to teda zhruba dve tisícročia, kým si ľudia všimli, že texty písané v bežných hovorových jazykoch nie sú úplne náhodné zhluky písmen, slov a viet, ale podliehajú zákonitostiam, ktoré sa dajú vyjadriť rečou matematiky. A nielen to, ale tieto zákonitosti sa dajú využiť aj pri lúštení šifier. Spočiatku bolo toto umenie rozvíjané v oblasti blízkeho východu a stredomoria obývaného arabmi. Do Európy dorazilo až o niekoľko storočí neskôr.

V podstate počas celého stredoveku a veľkej časti novoveku boli jediné štatistické vlastnosti textu, ktoré sa využívali pri lúštení šifier len relatívne početnosti znakov. Až koncom renesancie sa dokázateľne začali využívať aj relatívne početnosti bigramov, trigramov a prípadne slov a fráz. Vyššie n-gramy než sú trigramy sa pred érou počítačov príliš nevyužívali, kvôli vysokej náročnosti ich spracovávania. Iné štatistiky pri kryptoanalýze sa začali objavovať až koncom 19. storočia, počas 1. svetovej vojny a po nej. V čase 2. svetovej vojny zažívala kryptoanalýza klasických šifier, ale už aj prvých šifrovacích strojov svoj vrchol. S týmto súvisí aj rozvoj štatistickej analýzy textov, ktorá bola nevyhnutná najmä pri lúštení šifier generovaných šifrovacími strojmi. Okrem toho sa vedci aj v oblasti lingvistiky, čiže oblasti nesúvisiacej priamo so šiframi a kryptoanalýzou, začali v druhej polovici 20. storočia venovať matematickým modelom bežných jazykov, ich axiomatizácii a podrobnej analýze. S uvedenými skutočnosťami súvisí významný nárast záujmu o štatistickú analýzu textov, predovšetkým v posledných desaťročiach minulého storočia, najmä v súvislosti s pokusmi o automatizáciu prekladu či prepisovanie hovoreného slova na text.

1.2 Problém rozpoznávania jazyka

Ravi Ganesan, Alan T. Sherman: *Statistical Techniques for Language Recognition: An Introduction and Guide for Cryptanalysis*

Tento článok obsahuje veľmi zaujímavé informácie, pokiaľ ide o využitie štatistických metód v kryptoanalýze. Ďalším jeho pozitívom je množstvo odkazov na zdroje z oblasti štatistického spracovania textu.

V úvodnej časti článku autori spomínajú staršie výsledky, ktoré boli ich motiváciou. Jedná sa prevažne o výsledky Friedmana a Sinkova, nimi zavedené štatistiky a ich rôzne obmeny a to ako sa dajú využiť pri kryptoanalýze.

V ďalšej časti autori popisujú model jazyka[1], ktorý neskôr používajú pri

[1]Pod jazykom, prípadne prirodzeným jazykom, sa tu rozumie nielen bežný hovorový jazyk ako je slovenčina, angličtina, nemčina a pod. Jazykom rozumieme akýkoľvek reťazec

riešení formulovaných problémov. Jedná sa o Markovov model jazyka, ktorý je veľmi dobre implementovateľný a dostatočne dobre aproximuje reálny jazyk. V článku je pri riešení problémov využitý len Markovov model jazyka prvého rádu.

Následne autori formulujú štyri problémy a zaoberajú sa ich riešením. Tieto problémy sú:

1. Rozpoznanie známeho jazyka.
2. Odlíšenie známeho jazyka od rovnomerne rozdeleného šumu.
3. Odlíšenie neznámeho šumu 0-ho rádu od neznámeho jazyka 1-ho rádu.
4. Detekovanie nerovnomerne rozdeleného neznámeho jazyka.

Uvedené štyri problémy autori riešia pomocou štatistických metód testovaním hypotéz na Markovovských reťazcoch. Vždy sa testuje nulová hypotéza H_0 proti alternatívnej hypotéze H_1. Podľa podmienok úlohy sa určí kritická oblasť a podľa toho či testovacia štatistika padne mimo kritickej oblasti alebo do nej, sa prijme alebo zamietne hypotéza H_0. V prípade zamietnutia H_0 sa príjme hypotéza H_1. Ku každému z uvedených problémov je v článku uvedený aj ilustračný príklad aplikácie príslušného problému v kryptoanalýze. Tieto ilustračné príklady tu teraz tiež uvedieme pre lepšie objasnenie formulovaných problémov a ich aplikácie v kryptoanalýze. Vo všetkých problémoch sa testovaný text reprezentuje pomocou Markovovského reťazca prvého rádu a matica jeho prechodových pravdepodobností sa bude označovať P_M. Známy jazyk sa takisto bude reprezentovať Markovovským reťazcom prvého rádu a jeho matica prechodových pravdepodobností sa bude označovať P_B. Matica prechodových pravdepodobností rovnomerne rozdeleného náhodneho šumu sa bude označovať P_U (všetky jej prvky sú rovnaké). Príklady a podrobnejší popis jednotlivých problémov sú nasledovné:

1. **Nájdenie C programu medzi nezašifrovanými súbormi.**
   C program je jazyk s dobre definovanou štruktúrou. Tento autori modelujú pomocou Markovovského reťazca prvého rádu. Ostatné súbory sú neznáme, ale sú nezašifrované a predpokladá sa, že takisto majú nejakú štruktúru (TXT, PS, PDF,...). Nulová hypotéza bude $H_0 : P_M = P_B$ a alternatívna hypotéza bude $H_1 : P_M \neq P_B$.
2. **Hľadanie DES kľúča zašifrovaného súboru v známom jazyku.**
   V tomto prípade poznáme jazyk, v ktorom je písaný zašifrovaný text.

znakov s definovanou štruktúrou. Takže pod jazykom rozumieme aj programovacie jazyky, obrázky (napr. JPG súbory), textové súbory (napr. PDF, PS súbory) atď.

Ak súbor zašifrovaný pomocou DES dešifrujeme s použitím nesprávneho kľúča, tak výsledok sa bude podobať rovnomerne rozdelenému náhodnému šumu. Takže nulová hypotéza bude $H_0 : P_M = P_B$ a alternatívna hypotéza bude $H_1 : P_M = P_U$.

3. **Hľadanie kľúča Bealovej šifry na neznámom texte.**
   Ak text zašifrovaný Bealovou šifrou dešifrujeme s použitím nesprávneho kľúča, tak výsledok bude šum nultého rádu. V prípade dešifrovania správnym kľúčom bude mať výsledok vyšší rád, ale o zašifrovanom texte nemáme žiadne informácie, takže nepoznáme jeho štruktúru. Nulová hypotéza bude $H_0 : \text{rád}(M) = 0$ a alternatívna hypotéza bude $H_1 : \text{rád}(M) = 1$.

4. **Hľadanie REDOC II kľúča zašifrovaného neznámeho textu.**
   Ak text zašifrovaný REDOC II šifrou dešifrujeme s použitím nesprávneho kľúča, tak výsledok sa bude podobať rovnomerne rozdelenému náhodnému šumu. V prípade dešifrovania správnym kľúčom bude mať výsledok nejakú štruktúru, o ktorej však nemáme žiadne informácie. Nulová hypotéza bude $H_0 : P_M = P_U$ a alternatívna hypotéza bude $H_1 : P_M \neq P_U$.

Toto sú teda problémy riešené v článku. Ako vidno aplikovanie jednotlivých problémov v kryptoanalýze sa čiastočne prekrýva a naviac sa dajú formulovať ďalšie problémy relevantné pri kryptoanalýze.

Riešenia uvedených problémov v článku sa dajú rozdeliť na dve skupiny. Jednu skupinu tvorí 2. problém samostatne. Obe jeho hypotézy sú jednoduché a podľa Neyman-Pearsonovej lemy existuje asymptoticky najsilnejší test[2] riešiaci tento problém. Zvyšné tri problémy tvoria druhú skupinu. V každom z nich je aspoň jedna z hypotéz zložená a vo všeobecnosti neexistuje pre ne najsilnejší test. Tieto tri problémy sú riešené pomocou testov maximálnej vierohodnosti.

V záverečnej časti článku autori ešte stručne uvažujú nad variáciami riešených problémov. Ide najmä o to, ako aplikovať uvedené testy na krátke alebo zašumené texty. Tieto otázky v článku sú len postavené, ale nie riešené. Okrem toho je v záverečnej časti článku formulovaných aj niekoľko otvorených problémov. Jedná sa napríklad o použitie iných, prípadne rozšírených modelov jazyka pri riešení uvedených problémov, ďalej problém toho aký rád Markovovského reťazca je optimálny pri modelovaní toho, ktorého jazyka a podobne.

Na tento článok ešte nadväzuje článok „Heuristic Language Analysis: Techniques and Applications“ [33]. Autor tohto článku sa rozhodol napísať

[2] Aspoň tak silný ako akýkoľvek iný test.

program na automatickú kryptoanalýzu, využívajúci výsledky z [11] a prakticky overiť výsledky. Implementované boli len testy na prvé dva problémy, t.j. rozpoznanie známeho jazyka a odlíšenie známeho jazyka od rovnomerne rozdeleného náhodného šumu. Automatická kryptoanalýza bola testovaná na štyroch šifrách. Podľa záverov uvádzaných v [33] sú testy pre oba problémy funkčne približne zhodné.

1.3 Štatistické indexy využívané pri kryptoanalýze

1.3.1 Abeceda

Použitá abeceda, či už otvoreného alebo zašifrovaného textu, je jednou zo základných štatitstických vlastností textu. Jej využiteľnosť prí lúštení je veľmi obmedzená. Samostatne sa prakticky nepoužíva. Avšak v spojení s ďalšími informáciami získanými z postranných kanálov, prípadne v spojitosti s inými štatistikými indexami môže aj znalosť použitej abecedy prispieť k lúšteniu zašifrovanej správy.

Pri skúmaní abecedy daného jazyka nemáme veľa možností. Môžeme skúmať kvantitatívne vlastnosti, t.j. počet znakov abecedy daného jazyka, alebo kvalitatívne vlastnosti, t.j. aké znaky príslušná abeceda obsahuje. Pri skúmaní abecied rôznych jazykov veľmi rýchlo zistíme, že jednotlivé jazyky sa abecedami viac či menej líšia. Napríklad angličtina používa sadu znakov, ktorá sa označuje TSA. Klasická latinská abeceda z TSA neobsahuje znaky J, U a W. Nemecká abeceda má oproti TSA navyše znaky s „umlautom“ Ä, Ö a Ü a ostré „es“ ß. Holandská abeceda má oproti TSA naviac znak ÿ, ktorý vznikol spojením znakov *ij*. A takto by sme mohli pokračovať ďalej. Žiadne dva rôzne stredoeurópske jazyky nepoužívajú tú istú abecedu a to ani v prípade takých príbuzných jazykov akými sú slovenčina, čeština a slovinština, prípadne ruština, ukrajinčina a bulharština a pod. Preto by sa zjednodušene dalo povedať, že ak poznáme použitú abecedu, tak vieme o aký jazyk sa jedná. Nie je to však pravda. Problém spočíva v tom, že reálne texty takmer vždy obsahujú aj cudzie slová, prípadne číslovky a teda znaky, ktoré nepatria do abecedy jazyka textu. Napríklad slovenská abeceda ([35]) neobsahuje písmená W a X, ale tieto sa v cudzích slovách, menách a názvoch bežne vyskytujú aj v slovenských textoch: text, Xantipa, web, Wagner,...

V praxi väčšina ručných šifier využívala len TSA abecedu, prípadne TSA abecedu s pridaním zopár špeciálnych znakov. Jeden z dôvodov obmedzenia sa na TSA abecedu boli obmedzené technické prostriedky na prenos správ, akými boli telegraf, diaľnopis atď. Ďalší dôvod bol ten, aby sa na základe

použitej abecedy nedal jednoducho určiť použitý jazyk správy.

1.3.2 Frekvencie znakov

Ako bolo spomenuté v časti 1.1, v 8. storočí už bolo známe, že frekvencie znakov v každom jazyku nie sú náhodné, ale dodržiavajú isté zákonitosti. Toto bolo publikované v knihe „Kitāb al-mu'ammā“ a postupne sa rozšírilo z orientu aj do Európy. Na základe znalosti frekvencie znakov a jazyka otvoreného textu je možné lúštiť šifry akými sú jednoduchá zámena a jednoduché kódy. Tieto dva typy šifier boli do dôb renesancie v podstate jediné široko používané spôsoby šifrovania.

V tabuľke 1.1 na strane 17 sú pre ilustráciu uvedené frekvencie znakov angličtiny, nemčiny, slovenčiny a čestiny. Tieto frekvencie boli robené na textoch v príslušných jazykoch, ktoré všetky mali približne 5 miliónov znakov. Výsledky zodpovedajú tomu, čo sa bežne zvykne uvádzať v literatúre. Prekvapí snáď len pomerne nízka frekvencia písmena E v nemčine. V niektorých zdrojoch sa zvyknú udávať hodnoty až 18–20 %. Tento rozdiel môže byť čiastočne spôsobený tým, že v našej vzorke textu sme používali plnú nemeckú abecedu, obsahujúcu aj znaky Ä, Ö, Ü a ß. Bežne sa ale nemecké texty zvyknú transkribovať spôsobom Ä=AE, Ö=OE, Ü=UE a ß=SS, čo je gramaticky korektné. V takom prípade narastú frekvencie znakov A, E, O, U a S a na našej vzorke by to potom vyzeralo nasledovne:

znak	DE
A	4.801 %
E	15.710 %
O	2.104 %
U	3.563 %
S	6.226 %

Vidíme, že frekvencia znaku E narástla takmer na 16 % a ďalší rozdiel oproti iným udávaným hodnotám sa už dá vysvetliť len výberom a rozsahom vzorky textu.

1.3.3 Frekvencie n-gramov

Frekvencie n-gramov sa pri kryptoanalýze využívajú podobným spôsobom ako frekvencie znakov abecedy. Okrem toho sa dajú veľmi dobre využiť aj pri zisťovaní použitého jazyka.

znak	EN	DE	SK	CZ
a	6.554 %	4.414 %	8.154 %	6.312 %
á	0.000 %	0.000 %	1.483 %	1.718 %
ä	0.000 %	0.464 %	0.081 %	0.000 %
b	1.196 %	1.596 %	1.560 %	1.436 %
c	2.087 %	2.600 %	1.659 %	2.033 %
č	0.000 %	0.000 %	1.002 %	0.647 %
d	3.663 %	4.560 %	2.889 %	3.093 %
ď	0.000 %	0.000 %	0.188 %	0.043 %
e	10.485 %	14.534 %	6.770 %	6.997 %
é	0.000 %	0.000 %	0.458 %	0.729 %
ě	0.000 %	0.000 %	0.000 %	1.182 %
f	1.918 %	1.318 %	0.189 %	0.160 %
g	1.566 %	2.603 %	0.214 %	0.142 %
h	5.505 %	4.257 %	2.080 %	2.089 %
i	5.420 %	6.754 %	4.934 %	3.584 %
í	0.000 %	0.000 %	0.774 %	1.976 %
j	0.109 %	0.164 %	1.394 %	1.979 %
k	0.524 %	0.971 %	3.124 %	3.327 %
l	3.090 %	3.018 %	4.120 %	4.501 %
ľ	0.000 %	0.000 %	0.388 %	0.000 %
Ĺ	0.000 %	0.000 %	0.011 %	0.000 %
m	2.041 %	2.166 %	2.883 %	2.830 %
n	5.545 %	8.695 %	4.512 %	4.883 %
ň	0.000 %	0.000 %	0.154 %	0.037 %
o	6.106 %	1.892 %	7.708 %	6.475 %
ó	0.000 %	0.000 %	0.044 %	0.023 %
ô	0.000 %	0.000 %	0.135 %	0.000 %
ö	0.000 %	0.245 %	0.000 %	0.000 %
p	1.359 %	0.544 %	2.541 %	2.566 %
q	0.102 %	0.039 %	0.003 %	0.000 %
r	4.908 %	5.823 %	3.764 %	2.696 %
ŕ	0.000 %	0.000 %	0.018 %	0.000 %
ř	0.000 %	0.000 %	0.000 %	0.826 %
s	5.155 %	5.570 %	3.842 %	3.707 %
š	0.000 %	0.000 %	0.802 %	0.858 %
ß	0.000 %	0.378 %	0.000 %	0.000 %
t	7.409 %	4.994 %	3.658 %	4.323 %
ť	0.000 %	0.000 %	0.631 %	0.045 %
u	2.254 %	3.094 %	2.223 %	2.624 %
ú	0.000 %	0.000 %	0.601 %	0.069 %
ů	0.000 %	0.000 %	0.000 %	0.258 %
ü	0.000 %	0.527 %	0.000 %	0.000 %
v	0.816 %	0.694 %	3.403 %	3.427 %
w	1.869 %	1.310 %	0.010 %	0.019 %
x	0.132 %	0.633 %	0.020 %	0.014 %
y	1.403 %	0.127 %	1.234 %	1.701 %
ý	0.000 %	0.000 %	0.725 %	0.647 %
z	0.054 %	0.960 %	1.677 %	1.598 %
ž	0.000 %	0.000 %	0.829 %	1.084 %
␣	18.732 %	15.621 %	17.113 %	17.351 %

Tabuľka 1.1: Frekvencie znakov angličtiny, nemčiny, slovenčiny a češtiny

V kryptografii sa začali používať šifry využívajúce polygramové substitúcie[3] až počas renesancie. Takže až vtedy sa objavila potreba poznať frekvencie n-gramov pre $n > 1$ a vzhľadom na to, že frekvencie znakov už vtedajší lúštitelia bežne používali, nebol prechod k n-gramom ničím výnimočným. Malo to však jeden technický háčik, ktorý bol v tom čase len veľmi ťažko riešiteľný. Ak používame jazyk, ktorý má napríklad 26-znakovú abecedu, tak frekvenciu znakov nie je problém zistiť. Avšak bigramov bude v takomto jazyku 676 a vo všeobecnosti n-gramov bude 26^n. Ich zisťovanie je preto technicky veľmi náročné a potrebujeme naň podstatne väčšiu vzorku textu, než aká nám stačí na určenie frekvencie znakov.

Až do éry počítačov sa preto v praxi málokedy pracovalo s frekvenciami viac než trigramov. Ale už tieto veľmi účinným spôsobom zefektívňujú lúštenie jednoduchej substitúcie, bigramových šifier a aj umožňujú rozpoznávanie použitého jazyka.

1.3.4 Hĺbka závislosti znakov textu a jej význam pre zisťovanie entropie

Pod *hĺbkou závislosti dvoch znakov textu* sa rozumie maximálna hodnota $k \in \mathbb{N}$, pre ktorú ešte existuje korelácia medzi znakmi z_n a z_{n+k}, pričom index znaku tu označuje jeho pozíciu v texte. Túto veličinu skúmali najmä jazykovedci, ale má zmysel aj pre kryptológov.

Pre zmysluplné texty už dávno bolo ukázané, že jednotlivé znaky abecedy síce majú svoje relatívne početnosti výskytu, ale nie sú navzájom nezávislé. Ako príklad si môžeme uviesť angličtinu. V angličtine za písmenom Q vždy nasleduje písmeno U. To je príklad 100 % korelácie dvoch znakov idúcich bezprostredne za sebou. Ako iný príklad si uvedieme písmeno T. V angličtine za ním nasleduje písmeno H omnoho častejšie než akékoľvek iné písmeno abecedy. Takže to máme iný príklad korelácie dvoch po sebe idúcich znakov, pričom miera tejto korelácie je menšia než 100 %. Podobne sú navzájom korelované písmena aj v akomkoľvek inom zmysluplnom jazyku.

Uvedená korelácia dvoch znakov textu existuje aj pre znaky, ktoré sú navzájom vzdialené o viac než len jednu pozíciu v texte. Pre rôzne jazyky už bola empiricky zisťovaná hĺbka korelácie znakov, čiže hľadala sa maximálna hodnota k, pre ktorú ešte znaky z_n a z_{n+k} boli korelované.

V ďalšom texte si ukážeme ako sa dá z dostatočne veľkej vzorky textu zistiť miera korelácie dvoch znakov a pomocou toho určiť hĺbka závislosti znakov tohto textu.

[3] Väščinou sa jednalo len o bigramové substitúcie.

	A	B	C	.....	X	Y	Z
A	p_{AA}	p_{AB}	p_{AC}	.....	p_{AX}	p_{AY}	p_{AZ}
B	p_{BA}	p_{BB}	p_{BC}	.....	p_{BX}	p_{BY}	p_{BZ}
C	p_{CA}	p_{CB}	p_{CC}	.....	p_{CX}	p_{CY}	p_{CZ}
⋮				⋱			
⋮				⋱			
⋮				⋱			
X	p_{XA}	p_{XB}	p_{XC}	.....	p_{XX}	p_{XY}	p_{XZ}
Y	p_{YA}	p_{YB}	p_{YC}	.....	p_{YX}	p_{YY}	p_{YZ}
Z	p_{ZA}	p_{ZB}	p_{ZC}	.....	p_{ZX}	p_{ZY}	p_{ZZ}

Tabuľka 1.2: Tabuľka relatívnych početností dvojíc znakov.

Najskôr si ukážeme ako určíme mieru korelácie dvoch znakov vzdialených o k pozícií a to či pre danú hodnotu $k \in \mathbb{N}$ sú ešte tieto dva znaky textu korelované. V uvedenom popise budeme pre jednoduchosť používať len TSA abecedu, čiže v ďalej bude abeceda $\mathcal{A}$ =TSA. Postup by bol ale úplne rovnaký aj pre akúkoľvek inú abecedu.

Z danej vzorky textu s abecedou $\mathcal{A}$ zistíme relatívne početnosti p_X pre všetky $X \in \mathcal{A}$. Potom vyhľadáme postupne všetky dvojice znakov vzdialených o k pozícií a zostavíme si tabuľku ich relatívnych početností. Príkladom takejto tabuľky je tabuľka 1.2.

Pokiaľ potrebujeme len vedieť, či sú už dvojice znakov nezávislé, tak si musíme pre každú dvojicu znakov XY[4] zistiť jej podmienenú pravdepodobnosť[5]. Túto vypočítame nasledovne:

$$p(Y|X) = \frac{p_{XY}}{p_X}, \quad \text{kde } p_X \text{ je relatívna početnosť znaku } X \tag{1.1}$$

Keď už máme pre všetky dvojice XY vypočítané ich podmienené pravdepodobnosti a zapíšeme si ich do podoby tabuľky, tak sa stačí pozrieť na riadky tabuľky. Ak by dvojice znakov už neboli korelované, tak v každom riadku tabuľky by sme mali mať relatívne početnosti výskytu znakov zo zodpovedajúceho stĺpca.

[4]Máme na mysli dvojicu znakov, ktoré su v texte vzdialené o $k \in \mathbb{N}$ pozícií.

[5]Pre pravdepodobnosť náhodného javu Y za podmienky náhodného javu X v texte používame dve rôzne označenia. V časti 1.3.9, vzorec (1.7), používame označenie $p_X(Y)$, prevzaté zo Shannonových článkov. Tu používame označenie $p(Y|X)$, ktoré je bežnejšie.

My ale chceme nielen zistiť, či sú znaky vzdialené o k pozícií korelované, ale chceme aj vyjadriť mieru ich korelácie, aby sme videli ako sa táto mení so zmenou hodnoty k. Vzhľadom na to, že hodnoty náhodných premenných X a Y sú znaky abecedy, nemôžeme na vyjadrenie ich korelácie použiť Pearsonov alebo Spearmanov koeficient. Vhodnou mierou závislosti našich náhodných premenných X a Y je *koeficient kontingencie*. V našom prípade sa koeficient kontingencie vypočíta podľa vzorca:

$$\varphi(X,Y) = \sum_{\forall X \times Y \in \mathcal{A}^2} \frac{p_{XY}^2}{p_X . p_Y} - 1 \tag{1.2}$$

Podľa [25] (str. 75) platí veta:

Veta 1 *Nech X a Y sú ľubovoľné náhodné premenné, ktoré nadobúdajú len konečný počet hodnôt. Nech X nadobúda m rôznych hodnôt a Y nadobúda n rôznych hodnôt a nech $m \leq n$. Potom platí:*

1. $\varphi(X,Y) = 0$ *práve vtedy, keď sú náhodné premenné X a Y nezávislé*
2. $\varphi(X,Y) \leq n - 1$
3. $\varphi(X,Y) = n-1$ *práve vtedy, keď existuje funkcia f taká, že $Y = f(X)$.*

V našom prípade obe náhodné premenné nadobúdajú hodnoty z množiny $\mathcal{A}$, pre ktorú platí $|\mathcal{A}| = n$. Preto vetu 1 môžeme upraviť do podoby:

Dôsledok 1 *Nech X a Y sú ľubovoľné náhodné premenné, ktoré nadobúdajú konečný počet hodnôt z množiny $\mathcal{A}$, pre ktorú platí $|\mathcal{A}| = n$. Potom platí:*

1. $\varphi(X,Y) = 0$ *práve vtedy, keď sú náhodné premenné X a Y nezávislé*
2. $\varphi(X,Y) \leq n - 1$
3. $\varphi(X,Y) = n-1$ *práve vtedy, keď existuje funkcia f taká, že $Y = f(X)$.*

Ak naviac chceme vyjadriť mieru korelácie náhodných premenných X a Y na intervale $\langle 0, 1 \rangle$, tak koeficient kontingencie vydelíme koeficientom $|\mathcal{A}| - 1$: $\psi(X,Y) = \frac{\varphi(X,Y)}{|\mathcal{A}|-1}$. Potom platí:

Dôsledok 2 *Nech X a Y sú ľubovoľné náhodné premenné, ktoré nadobúdajú konečný počet hodnôt z množiny $\mathcal{A}$, pre ktorú platí $|\mathcal{A}| = n$. Potom platí:*

1. $\psi(X, Y) = 0$ *práve vtedy, keď sú náhodné premenné X a Y nezávislé*
2. $\psi(X, Y) \leq 1$
3. $\psi(X, Y) = 1$ *práve vtedy, keď existuje funkcia f taká, že $Y = f(X)$.*

Týmto sme si ukázali ako vypočítame mieru korelácie dvoch znakov vzdialených o $k \in \mathbb{N}$ pozícii v texte a teraz si ukážeme experimentálne získané výsledky tejto korelácie pre rôzne jazyky.

Testy, aby ich výpovedná hodnota bola čo najväčšia, budeme robiť na rovnakom texte v deviatich rôznych jazykoch. Jedná sa o text románu *Dvadsaťtisíc míľ pod morom* od Julesa Verna. Jazyky, v ktorých máme tento text k dispozícii, sú:

- Angličtina (EN)
- Holandčina (NL)
- Polština (PL)
- Čeština (CZ)
- Maďarčina (HU)
- Španielčina (ES)
- Francúzština (FR)
- Nemčina (DE)
- Taliančina (IT)

Všetky uvedené jazykové verzie boli prevedené do TSA abecedy vynechaním diakritiky a interpunkcie. Len v nemeckých textoch, pretože tie to umožňujú, bola spravená transkribcia obvyklým spôsobom, t.j. Ä=AE, Ö=OE, Ü=UE a ß=SS. S výnimkou taliančiny sú dĺžky textov vo všetkých jazykoch približne rovnaké a pohybujú sa medzi približne 660 kB pri českej verzii po približne 824 kB pri francúzskej verzii. Len talianska verzia mala približne 388 kB, čím bola výrazne kratšia. Zrejme sa jednalo o nejakú skrátenú verziu románu.

Závislosť znakov bola zisťovaná pomocou skriptu napísaného v Pythone, spôsobom popísaným v predošlých odstavcoch. Algoritmus je triviálny a časová náročnosť spracovania súborov uvedenej veľkosti bola zanedbateľná. Koeficient kontingencie sme zisťovali dvoma spôsobmi. Najskôr pre 27 znakovú TSA abecedu, čiže 26 písmen a medzera a potom pre 26 znakovú TSA abecedu obsahujúcu len písmená. Výsledky oboch verzií boli takmer rovnaké a odchýľky boli minimálne. Maximálna odchýlka bola pri závislosti dvoch znakov idúcich bezprostredne za sebou a táto odchýlka bola u všetkých textov približne 1 %[6]. Pre znaky vzdialené o $k \geq 2$ pozícii potom táto odchýlka

[6]Pokiaľ koeficient kontingencie vyjadrujeme v percentách od 0 % po 100 %.

veľmi rýchlo klesala s rastúcim k. Takže tu uvedieme výsledky len pre verziu pracujúcu s 27 znakovou TSA abecedou, obsahujúcou aj medzeru, čo je v podstate text dobre čitateľný aj človekom.

Jazyk / Vzdialenosť	CZ	DE	EN	ES	FR	HU	IT	NL	PL
	Hodnoty koeficientu kontingencie uvedené v %								
$k = 1$	4.666	6.297	4.573	4.714	4.904	4.323	5.196	6.715	5.511
$k = 2$	1.208	2.104	1.485	1.776	1.813	1.660	3.002	2.483	1.413
$k = 3$	0.426	0.773	0.576	0.588	0.673	0.835	1.449	0.897	0.452
$k = 4$	0.243	0.366	0.239	0.271	0.379	0.247	0.371	0.432	0.161
$k = 5$	0.127	0.167	0.120	0.165	0.258	0.146	1.879	0.194	0.105
$k = 6$	0.069	0.107	0.080	0.089	0.091	0.066	0.105	0.118	0.061
$k = 7$	0.044	0.072	0.046	0.056	0.065	0.046	0.074	0.067	0.038
$k = 8$	0.045	0.039	0.031	0.044	0.036	0.022	0.046	0.043	0.027
$k = 9$	0.024	0.032	0.021	0.025	0.027	0.021	0.037	0.025	0.023
$k = 10$	0.018	0.024	0.019	0.019	0.023	0.012	0.026	0.018	0.021
$k = 11$	0.018	0.021	0.013	0.014	0.017	0.010	0.038	0.020	0.011
$k = 12$	0.011	0.014	0.007	0.011	0.013	0.007	0.019	0.011	0.019
$k = 13$	0.007	0.008	0.008	0.008	0.009	0.006	0.018	0.018	0.008
$k = 14$	0.005	0.007	0.005	0.006	0.007	0.005	0.012	0.008	0.005
$k = 15$	0.006	0.004	0.005	0.006	0.006	0.005	0.011	0.011	0.006
$k = 16$	0.008	0.004	0.005	0.004	0.006	0.006	0.008	0.018	0.004
$k = 17$	0.005	0.004	0.004	0.004	0.005	0.005	0.007	0.008	0.004
$k = 18$	0.005	0.004	0.004	0.003	0.004	0.003	0.007	0.012	0.005
$k = 19$	0.005	0.003	0.004	0.005	0.004	0.004	0.006	0.004	0.004
$k = 20$	0.004	0.003	0.003	0.003	0.003	0.003	0.019	0.004	0.004

Tabuľka 1.3: Koeficient kontingencie pre znaky vzdialené o k pozícií v texte, zisťovaný na románe **Jules Verne:** *20 000 míľ pod morom*

Vo všetkých uvedených jazykoch sme nechali koeficient kontingencie počítať pre znaky vzdialené $k = 1$ až $k = 100$ pozícii v texte. Bežne sa v jazykovednej literatúre uvádza závislosť znakov v maximálnej vzdialenosti 20 až 50 pozícií v texte, pričom údaje kolíšu podľa zdroja a metodika zisťovania týchto hodnôt v prevažnej väčšine prípadov nie je uvedená. Takže vzdialenosť 100 sa javí ako dostatočne vysoká hranica.

My sme si za hranicu, kedy už znaky nebudeme považovať za závislé, stanovili hodnotu koeficientu kontingencie 0.005 %. Takže $k \in \mathbb{N}$, pre ktoré hodnota koeficientu kontingencie prvý raz klesla pod túto hranicu, považujeme za vzdialenosť, pri ktorej už dva znaky textu nie sú závislé.

Iný spôsob ako stanoviť túto hranicu by bol sledovať diferencie medzi koeficientami kontingencie pre po sebe idúce hodnoty k a keď tieto diferencie klesnú pod nejakú stanovenú hranicu, určiť príslušné $k \in \mathbb{N}$ za hĺbku závis-

losti znakov. Hodnoty koeficientu kontingencie totiž, ako vidno z tabuľky 1.3, neklesajú monotónne, ale mierne kolíšu.

Pri prvom uvedenom spôsobe stanovenia hranice závislosti, ktorý sme použili my, vidno z tabuľky 1.3, že táto hranica sa pohybuje medzi $k = 15$ až $k = 20$. Výnimkou je taliansky text, ktorý nami stanovenú hranicu nedosiahol, aj keď tendencia koeficientu kontingencie bola celkovo klesajúca. Pri talianskom texte tento koeficient pri rastúcom k kolísal podstatne viac než pri ktoromkoľvek inom jazyku. Väčšinou sa pohyboval v rozsahu $\langle 0.006 - 0.007 \rangle$, ale každých 5–10 krokov pre hodnotu k vyskočil o rád vyššie. V nasledovnej tabuľke je uvedené kolísanie hodnôt koeficientu kontingencie pre $k = 21$ až $k = 100$, ako aj jeho dominantná hodnota:

Jazyk	Rozptyl	Dominantná hodnota
CZ	$\langle 0.003, 0.007 \rangle$	0.004 %
DE	$\langle 0.003, 0.004 \rangle$	0.003 %
EN	$\langle 0.002, 0.003 \rangle$	0.003 %
ES	$\langle 0.002, 0.004 \rangle$	0.003 %
FR	$\langle 0.002, 0.004 \rangle$	0.003 %
HU	$\langle 0.003, 0.005 \rangle$	0.003 %
IT	$\langle 0.005, 0.067 \rangle$	0.006 %
NL	$\langle 0.003, 0.004 \rangle$	0.003 %
PL	$\langle 0.002, 0.006 \rangle$	0.003 %

Pod dominantnou hodnotou tu máme na mysli hodnotu, ktorú nadobudol koeficient kontingencie pre väčšinu hodnôt k. Pri niektorých jazykoch ako boli napríklad francúzština, nemčina, angličtina to bolo takmer pre všetky hodnoty k s niekoľkými málo výnimkami. Len pre taliančinu bolo kolísanie podstatne väčšie, aj keď aj tam pre viac než polovicu hodnôt k nadobudol koeficient kontingencie dominantnú hodnotu a celkovo len v 23 prípadoch vyskočil o rád vyššie, pričom väčšinou to bolo na hodnotu 0.010 a len v 3 prípadoch to boli hodnoty 0.039, 0.053 a 0.067. Pre zaujímavosť sme spravili aj výpočet koeficientu kontingencie pre taliančinu pre hodnoty $1 \leq k \leq 1000$. Ani pre hodnoty $k > 100$ neklesla veľkosť koeficientu kontingencie pod hranicu 0.005 %. Stále mala prevažne hodnotu 0.006 %, akurát odchýlky od tejto hodnoty boli s rastúcim k stále výnimočnejšie.

V ôsmych z deviatich jazykoch, ktoré sme skúmali, je hranica závislosti dvoch znakov v texte, vzdialenosť 15 až 20 pozícii. Za touto hranicou sa už hodnota koeficientu kontingencie prakticky nemení. V podstate aj pre taliančinu si môžme ako hranicu závislosti zobrať $k = 20$ a nespravíme príliš veľkú chybu, vzhľadom na to, že pre $k \geq 21$ má v talianskom texte koeficient kontingencie väčšinou hodnotu 0.006.

To, že hodnota koeficientu kontingencie sa takmer nemení, znamená, že v tabuľke relatívnych početností dvojíc znakov so vzdialenosťou k (tabuľka 1.2, strana 19) budú vo všetkých riadkoch približne rovnaké hodnoty. Z toho potom vyplýva, že aj pri výpočte podmienenej entropie nemusíme brať n-gramy väčšie než je hranica závislosti dvoch znakov v texte, pretože ak je hĺbka závislosti v nejakom jazyku napríklad 17, tak podmienené entropie 17-gramu a vyšších n-gramov sa nebudú prakticky líšiť, čo vyplýva zo spôsobu, akým sa počíta podmienená entropia n-gramu.

Na záver sme ešte skúsili vypočítať hĺbku závislosti pre rôzne jazyky na vzorkách textov, ktoré mali dĺžku približne 5 miliónov znakov. Výsledky koeficientu kontingencie boli pre hodnoty $k \in \{1, 2, 3\}$ takmer rovnaké ako v prípade románu *20 000 míľ pod morom*. Pre vyššie hodnoty k koeficient kontingencie klesal rýchlejšie, dosahoval pri niektorých jazykoch až výrazne nižších hodnôt a hĺbka závislosti nám pri našom spôsobe jej určovania vyšla menšia. Jej hodnoty uvádzame v tabuľke 1.4. Vzorku textu pre španielčinu sme nemali, avšak mali sme naviac vzorky pre latinčinu a slovenčinu.

Jazyk	Hĺbka závislosti	Koeficient kontingencie [%]
CZ	10	0.003
DE	12	0.004
EN	11	0.004
FR	13	0.003
HU	9	0.004
IT	12	0.004
LA	18	0.004
NL	11	0.003
PL	12	0.003
SK	11	0.003

Tabuľka 1.4: Hĺbka závislosti na vzorkách zhruba 5 miliónov znakov

Ako už bolo spomenuté skôr, hĺbka závislosti znakov vyšla menšia než pri jej zisťovaní na texte románu *20 000 míľ pod morom*. Dalo by sa to vysvetliť tým, že vzorky textov boli niekoľkonásobne[7] väčšie a nesúrodejšie. Pozostávali z viacerých samostatných textov prozaického charakteru. V románe *20 000 míľ pod morom* sa nachádza mnoho dlhých čísloviek vypisovaných slovom, veľa pomerne dlhých geografických názvov a pod. V ňom sú priemerné dĺžky slov väčšie než je priemer v príslušnom jazyku a z toho zrejme vyplýva aj väčšia hĺbka závislosti znakov. Na vzorkách veľkých približne 5 miliónov

[7] V priemere 6–7 násobne.

znakov sa hĺbka závislosti pohybovala od 9 po 13. Jedinou zvláštnou výnimkou bol latinský text, pri ktorom hĺbka závislosti vyskočila až na 18. Ale tento latinský text mal aj priemernú dĺžku slov výrazne väčšiu, než bola priemerná dĺžka slov vo všetkých ostatných jazykoch. Priemerným dĺžkam slov sa budeme stručne venovať v nasledujúcej časti 1.3.5.

1.3.5 Priemerná dĺžka slov v danom jazyku

Priemernou dĺžkou slov sa jazykovedci zaoberajú pri zisťovaní štatistických vlastností jazyka a aj pri lúštení klasických šifier má znalosť priemernej dĺžky slov svoje využitie. My teraz určíme priemerné dĺžky slov na vzorkách textov, na ktorých sme zisťovali hĺbku závislosti znakov. Potom sa pozrieme, či sa medzi týmito dvoma veličinami vyskytuje nejaká evidentná závislosť.

Dĺžky slov v románe **Jules Verne:** *20 000 míľ pod morom*				
Jazyk	Počet slov	Max. dĺžka	Priem. dĺžka	Najdlhšie slovo
CZ	112 570	25	5.08692	čtyřiadvaceticentimetrové
DE	115 422	29	5.45025	vierundzwanzigpfuenderkanonen
EN	142 639	18	4.74215	unintentionally
ES	139 915	19	4.79686	extraordinariamente
FR	152 886	18	4.61121	chronométriquement
HU	113 884	23	5.84950	kormánynyilatkozatokkal
IT	63 289	39	5.25265	ventiquattromilaquattrocentoquarantotto
NL	131 081	25	4.98261	verzekeringmaatschappijen
PL	117 786	23	5.54507	najniebezpieczniejszych

Tabuľka 1.5: Priemerné dĺžky slov v rôznych jazykoch #1

V tabuľke 1.5 máme priemerné dĺžky slov v deviatich rôznych jazykoch, zisťované na texte románu *20 000 míľ pod morom*[8]. V tabuľke 1.6 na strane 26 máme priemerné dĺžky slov v desiatich rôznych jazykoch, zisťované na vzorkách textov veľkosti približne 5 miliónov znakov. V oboch tabuľkách máme aj niektoré ďalšie údaje ako sú: počet slov vo vzorke textu, dĺžka najdlhšieho slova a najdlhšie nájdené slovo.

Ako je vidno z uvedených tabuliek, priemerné dĺžky slov v skoro všetkých skúmaných jazykoch sa pohybujú medzi 5–6 znakmi. Rozdiely sú až na 1. a 2. desatinnom mieste. Pri vzorkách textov obsahujúcich približne 5 miliónov znakov, priemerné dĺžky slov mierne poklesli. Tento fakt, ako sme už spomenuli na konci predošlej časti, je pravdepodobne spôsobený tým, že v romáme *20 000 míľ pod morom* je veľa dlhých čísloviek vypisovaných slovom a veľa

[8]Priemerná dĺžka slov španielčiny zodpovedá hodnote, ktorá sa uvádza v článku [19].

Dĺžky slov v rôznych jazykoch na vzorkách približne 5 milíonov znakov				
Jazyk	Počet slov	Max. dĺžka	Priem. dĺžka	Najdlhšie slovo
CZ	827 517	24	4.76327	krestanskodemokratickeho
DE	849 138	29	5.42575	allerweltbedienungscandidaten
EN	1 072 286	18	4.33856	characteristically
FR	1 037 635	19	4.37497	revolutionnairement
HU	702 033	35	5.45014	kilenczezerkilenczszazkilenczvenket
IT	901 940	32	4.76653	tredicimilaseicentocinquantanove
LA	917 660	21	5.68835	israelitisimpendebant
NL	937 859	27	4.75772	grootwaardigheidsbekleeders
PL	1 113 871	23	3.31882	kilkakrocstotysiecznego
SK	813 919	27	4.84367	tisicdevatstopatdesiatdevat

Tabuľka 1.6: Priemerné dĺžky slov v rôznych jazykoch #2

pomerne dlhých geografických názvov, čo zvýšilo priemerné dĺžky slov. Aj najdlhšie slová v maďarskom a talianskom texte, s dĺžkou cez 30 znakov, v tabuľke 1.6, sú číslovky vypisované slovom.

Výnimočná bola zmena v priemernej dĺžke slov pre vzorky poľských textov. Táto z hodnoty 5.54507 pre román *20 000 míľ pod morom*, spadla až na hodnotu 3.31882 pre vzorku textu veľkosti 5 miliónov znakov. Tým sa stala poľština jazykom s najkratšou priemernou dĺžkou slova. Pritom pri bežnom pohľade na použitú vzorku poľského textu, obsahoval tento aj dlhé slová a slová obvyklej dĺžky. V poľskom texte sa ale nachádza nezvyklo veľa jedno-, dvoj- a trojpísmenkových spojok, predložiek a slov. Ako príklad môžu slúžiť: a, i, o, w, z, co, do, ma, na, no, od, te, to,ze, ..., pričom dvojpísmenkových slov sa v texte nachádza výrazne viac než tu uvádzame a trojpísmenkových slov sú len na prvých 2 stranách textu desiatky. Pritom vzorka poľských textov je zostavená zo štrnástich bežných textov prozaického charakteru. Z týchto textov je 13 pôvodných poľských textov[9] a 1 text je preklad anglického textu do poľštiny[10].

Pozrime si teraz poradia testovaných jazykov podľa ich priemernej dĺžky slov. Tieto sú uvedené v tabuľke 1.7 na strane 27. Na oboch testovaných vzorkách sú tieto poradia takmer identické. Prehodili sa len poradia angličtiny a francúzštiny, pričom rozdiely v ich priemernej dĺžke slova sú minimálne a výrazne sa zmenilo poradie poľštiny, ktorá z takmer najväčšej priemernej dĺžky slov prešla na najkratšiu priemernú dĺžku slov.

Čo je pre nás však podstatné je fakt, že vo výsledkoch pre priemernú dĺžku slov a hĺbkou závislosti znakov v danom jazyku sa nejaví žiadna evidentná

[9]Napríklad: **Henryk Sienkiewicz:** *Pan Wołodyjowski*, *Quo vadis*, *Ogniem i mieczem.*
[10]**Arthur Conan Doyle:** *Tajemnica Baskerville – dziwne przygody Sherlocka Holmes*

Zoradené od najdlhšej po najkratšiu priemernú dĺžku slova										
Poradie	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Román **Jules Verne:** *20 000 míľ pod morom*										
Jazyk	HU	PL	DE	IT	CZ	NL	ES	EN	FR	–
Vzorka približne 5 miliónov znakov										
Jazyk	LA	HU	DE	SK	IT	CZ	NL	FR	EN	PL

Tabuľka 1.7: Poradia jazykov podľa priemernej dĺžky slov

a priama korelácia. Ak si zoberieme 3 jazyky s najväčšou alebo 3 jazyky s najmenšou priemernou dĺžkou slova, tak zistíme, že v hĺbke závislosti sa tieto vyskytujú rovnako medzi jazykmi s najväčšou ako aj medzi jazykmi s najmenšou hĺbkou závislosti znakov.

1.3.6 Od kappa indexu po index koincidencie

V tejto časti si popíšeme štatistické indexy, ktoré sa začali používať v súvislosti s lúštením polyalfabetických šifier. Z hľadiska kryptoanalýzy je najdôležitejší spomedzi týchto indexov, index koincidencie označovaný niekedy aj ako IC alebo I.C.

Index koincidencie je vo svojej podstate veľmi jednoduchý koeficient. Je to vlastne pravdepodobnosť toho, že ak si v skúmanom texte náhodne vyberieme dva znaky, tak tieto budú rovnaké. Keď si počet znakov skúmaného textu označíme n a symbolom n_i budeme označovať početnosť znaku i v skúmanom texte, tak potom IC bude:

$$IC = \sum_i \frac{n_i(n_i - 1)}{n(n-1)} .$$

Tento jednoducho vyzerajúci index je, podobne ako frekvencie znakov, pre každý jazyk charakteristický. Pri kryptoanalýze sa využíva hlavne:

1. na zisťovanie mono- alebo polyalfabecity šifry,
2. na určenie odhadu dĺžky kľúča pri šifrách s periodickým heslom,
3. a na hrubý odhad jazyka otvoreného textu.

Okrem toho sa využíva aj na rad ďalších testov súvisiacich s kryptoanalýzou. Uvedené indexy sú zavedené a dobre popísané v knihách:

1. **Friedman W. F.:** *Military Cryptanalysis I., II., III., IV.* ([5]–[8])

   Táto kniha je zaujímavá hneď z niekoľkých dôvodov. Verzia, ktorú máme k dispozícii [5]–[8], pochádza zo stránky NSA, bola vydaná v roku 1938 a sú to štyri Friedmanove príručky vojenskej kryptoanalýzy. Podľa informácií z iných zdrojov sa ale dá predpokladať, že aspoň časť týchto štyroch príručiek bola napísaná ešte v 20-tych, prípadne začiatkom 30-tych rokov minulého storočia. Informácia z Wikipédie, že tieto príručky boli napísané v rokoch 1956 až 1959 a publikované v rokoch 1957 až 1977 je preto buď chybná, alebo sa týka nejakej inej, prepracovanej verzie uvedených príručiek. Každopádne pre verejnosť boli Friedmanove príručky odtajnené až v 80-tych rokoch minulého storočia a boli opäť vydané v roku 1984.

   Uvedené štyri knihy boli napísané ako učebnice kryptoanalýzy pre vojenských kryptológov. Vzhľadom na dobu, v ktorej boli napísané, sa tieto učebnice zaoberajú klasickými šiframi a zhŕňajú všetky dovtedy známe poznatky o ich lúštení. Sú tam teda preberané frekvencie znakov v rôznych jazykoch, frekvencie bigramov a vo všeobecnosti n-gramov, ďalších štatistických vlastností jazyka a aj to, ako sa využívajú pri lúštení. V III. časti sú preberané aj rôzne štatistické testy, napr. test monoalfabecity šifry, test koincidencie a pod. Čo nás teraz ale bude zaujímať, je index koincidencie (IC). O IC sa môžme dočítať prakticky v každej kryptologickej učebnici a väčšinou sa tam stretávame s formuláciami typu „používanie IC zaviedol Friedman“, prípadne „IC objavil údajne Friedman“, avšak nikde som sa doteraz nestretol s citovaním zdroja, kde bol pôvodne IC definovaný a zavedený. Niekedy sa ako zdroj citujú uvedené Friedmanove príručky alebo jeho ďalšia kniha [9], ktorá má IC priamo v názve. Avšak ani v jednom z týchto zdrojov nie je IC definovaný a používaný v dnešnej podobe. Dnes používaná podoba IC sa vyvinula z κ, χ a Φ testov. Všetky tieto testy sú zavedené a popísané vo Friedmanových knihách. Napr. v II. časti Military Cryptanalysis ([6], str. 113–114) je popisovaná koincidencia znakov textu, čo je vlastne κ index. Tento κ index sa dá jednoducho definovať ako odhad pravdepodobnosti toho, že dva rovnako dlhé texty v tom istom jazyku budú mať na rovnakých pozíciach rovnaké znaky. V III. diely Friedmanovej príručky ([7], od str. 58) sa potom popisujú spomínané testy a ich využitie v kryptoanalýze. Na strane 94 v [7] je uvedený test monoalfabecity, tzv. Φ test, ktorý vlastne už využíva dnešnú podobu IC, akurát to tam nie je formálne definované ako samostatný index. Namiesto toho

je tam definovaný Φ test a stredná hodnota pre Φ v tvare:

$$f_A(f_A - 1) + f_B(f_B - 1) + ... + f_Z(f_Z - 1) = \kappa_p.N.(N - 1)$$

pričom f_i označuje počet výskytov znaku i, N je celkový počet znakov a κ_p je κ index skúmaného jazyka. Ľavú stranu uvedenej rovnice Friedman označuje $E(\Phi)$, čo je stredná hodnota indexu Φ. IC, tak ako ho definujeme a používame dnes, je v uvedenej rovnici vlastne κ_p. To síce nezodpovedá Friedmanovmu chápaniu κ indexu, ale poukazuje to na fakt, že všetky uvedené indexy navzájom úzko súvisia a κ index je aproximácia IC (samozrejme to platí aj naopak). Spomínanými testami sa podrobnejšie budeme zaoberať až v časti týkajúcej sa Bauerovej knihy, pretože tam sú tieto indexy a testy veľmi dobre popísané a je tam aj vysvetlený spôsob a dôvody toho ako sa od nich dospelo ku dnešnej podobe IC.

2. **William F. Friedman:** *The Index of Coincidence and its Applications in Cryptanalysis* ([9])

3. **Bauer, F. L.:** *Entzifferte Geheimnisse* ([2])

   Táto kniha patrí medzi najlepšie príručky kryptológie a existuje okrem nemeckého originálu aj v anglickom preklade. Autor je nielen matematik, ale aj bývalý profesionálny kryptológ a má v danej oblasti bohaté skúsenosti, čo je na kvalite textu poznať. Kniha je rozdelená na dve základné časti. Prvá časť sa venuje kryptografii a druhá časť sa venuje kryptoanalýze. Nás bude zaujímať druhá časť knihy, pretože štatistické metódy kryptoanalýzy tvoria jej podstatnú časť. Autor tu zhŕňa a podrobne popisuje všetky doteraz známe poznatky, počnúc anatómiou jazyka, cez rôzne indexy a testy až po ich praktické využitie pri lúštení šifier. Podrobne sú tu popísané aj všetky indexy a testy spomínané v časti o Friedmanových knihách *Military Cryptanalysis* a je aj vysvetlené ako a prečo sa od Friedmanových indexov κ, χ, Φ prešlo ku dnešnému IC, čo tu v stručnosti popíšeme.

   Pri uvádzaní výsledkov z Bauerovej knihy budeme pomocou T, resp. T' označovať skúmané texty (postupnosť znakov). Písmenom M budeme označovať dĺžku skúmaných textov. Pokiaľ nebude uvedené inak, tak predpokladáme, že tieto texty sú nad rovnakou abecedou, ktorej počet znakov budeme označovať písmenom N.

   Autor v kapitole „Kappa und Chi“ začína κ indexom. Ako už bolo spomenuté, tento index sa dá jednoducho popísať ako pravdepodobnosť

toho, že dva texty rovnakej dĺžky a nad rovnakou abecedou, budú mať na rovnakých miestach rovnaké znaky.

Ak si texty označíme $T = (t_1, t_2, ..., t_M)$ a $T' = (t'_1, t'_2, ..., t'_M)$, pričom $M > 1$, tak potom κ index[11] týchto dvoch textov sa formálne definuje ako:

$$\kappa(T, T') = \sum_{i=1}^{M} \frac{\delta(t_i, t'_i)}{M},$$

kde

$$\delta(x, y) = \begin{cases} 1, & \text{ak } x = y; \\ 0, & \text{inak.} \end{cases}$$

Už z definície je zrejmé, že $\kappa(T, T') \leq 1$ a $\kappa(T, T') = 1$ práve vtedy keď $T = T'$. Friedman si pravdepodobne ako prvý všimol, že tento index je závislý od jazyka a je pre každý jazyk charakteristický. Napr. pre angličtinu sa udávajú hodnoty 6.5 %–6.9 % a pre nemčinu sa udávajú hodnoty 7.5 %–8.3 % ([2], str. 325). κ index je pritom invariantný voči transpozíciam, jednoduchým substitúciam a aj niektorým typom polyalfabetických substitúcii. Dá sa preto využiť pri ich lúštení. Ak predpokladáme, že skúmané texty boli generované nezávislými stochastickými zdrojmi Q a Q' a poznáme pravdepodobnosti výskytu jednotlivých znakov (p_i bude označovať pravdepodobnosť výskytu znaku i v príslušnom zdroji), tak vieme vyjadriť strednú hodnotu κ indexu ([2], str. 325):

$$E\left(\kappa(T, T')\right)_{QQ'} = \sum_i p_i . p'_i$$

V prípade, že oba skúmané texty pochádzajú z rovnakého zdroja Q platí:

$$E\left(\kappa(T, T')\right)_{Q} = \sum_i p_i^2$$

a platí:

$$\frac{1}{N} \leq E\left(\kappa(T, T')\right)_{Q} \leq 1$$

pričom N je počet znakov použitej abecedy a ľavá rovnosť sa nadobúda len pri rovnomernom rozdelení znakov (t.j. $p_i = \frac{1}{N}$ pre $\forall i$). Ak

[11] κ index pravdepodobne ako prvý zaviedol Friedman vo svojej knihe „Index of Coincidence“ [9]. Friedman κ index označoval aj ako „index koincidencie“, resp. I.C., čo je trocha mätúce, pretože sa nejedná o dnešné chápanie IC. Presný rok vydania Friedmanovej knihy je nejasný. V elektronickej verzii nie je uvedený a v Bauerovej knihe je v nemeckom originále uvedený rok 1922 a v anglickom preklade je na tom istom mieste uvedený rok 1925. V iných zdrojoch sa dokonca uvádza rok 1920.

poznáme očakávané (stredné) hodnoty κ indexu pre rôzne jazyky, tak vieme jednoducho realizovať κ test, pomocou ktorého vieme spraviť hrubý odhad jazyka a to aj pri textoch zašifrovaných niektorými typmi šifier, vzhľadom na invariantnosť κ indexu voči nim.

V ďalšej kapitole autor pokračuje zavedením χ indexu. Podobne ako pri predošlom κ indexe vychádzame z dvoch textov $T = (t_1, t_2, ..., t_M)$ a $T' = (t'_1, t'_2, ..., t'_M)$ rovnakej dĺžky. Označme si pomocou m_i náhodnú premennú vyjadrujúcu počet výskytov znaku i v texte T a analogicky m'_i v texte T'. Potom samozrejme platí:

$$\sum_i m_i = \sum_i m'_i = M$$

kde M je dĺžka skúmaných textov. χ index sa následne definuje ako:

$$\chi(T, T') = \sum_i \frac{m_i.m'_i}{M^2}$$

Ak sa na uvedenú definíciu pozrieme, tak ihneď vidno, že sa vlastne jedná o odhad strednej hodnoty κ indexu, vzhľadom na to, že $E\left(\frac{m_i}{M}\right) = p_i$. Rovnako ako κ index, má aj χ index vlastnosť invariantnosti voči niektorým typom šifier (napr. transpozícia, jednoduchá substitúcia,...). Okrem toho platí $\chi(T, T') \leq 1$, pričom $\chi(T, T') = 1$ jedine vtedy, keď dĺžka textov je 1. Keď sú znaky textu rovnomerne rozložené, čiže ak $m_i = \frac{M}{N}$ (M je dĺžka textu a N je počet znakov použitej abecedy), tak $\chi(T, T') = \frac{1}{N}$. Dôležitý prípad, ktorý je ďalším krokom ku dnešnému chápaniu IC, je hodnota χ indexu pre $T = T'$:

$$\psi(T) = \chi(T, T) = \sum_i \frac{m_i^2}{M^2} \tag{1.3}$$

Podobne ako pre strednú hodnotu κ indexu, aj pre ψ index platia nerovnosti:

$$\frac{1}{N} \leq \psi(T) \leq 1$$

pričom rovnosť na ľavej strane sa nadobúda len pre rovnomerne rozdelené znaky textu ($m_i = \frac{M}{N}$ pre $\forall i$) a rovnosť na pravej strane nastáva len v prípade, že text má len 1 znak. ψ index sa dá interpretovať ako pravdepodobnosť toho, že ak z textu dva razy po sebe náhodne vyberieme jeden znak (s návratom), tak v oboch prípadoch vyberieme rovnaký znak. Táto veličina je opäť pre každý jazyk charakteristická. Pokiaľ sú skúmané texty generované nezávislými stochastickými zdrojmi Q a Q'

a poznáme pravdepodobnosti výskytu jednotlivých znakov (p_i, p'_i), tak vieme vypočítať strednú hodnotu χ indexu ([2], str. 328):

$$E\left(\chi(T,T')\right)_{QQ'} = \sum_i p_i.p'_i$$

Ako vidno, stredné hodnoty κ a χ indexov sú rovnaké. Pokiaľ sú oba texty generované rovnakým stochastickým zdrojom $(Q = Q')$, tak platí:

$$E\left(\kappa(T,T')\right)_Q = E\left(\chi(T,T')\right)_Q = E\left(\chi(T,T)\right)_Q = E\left(\psi(T)\right)_Q = \sum_i p_i^2$$

V kapitole „Das Kappa-Chi-Theorem“ ([2], str. 329) sa autor zaoberá skúmaním uvedených indexov na textoch T a $T^{(r)}$, kde $T^{(r)}$ označuje text T cyklicky zrotovaný o r miest doprava. Cieľom tejto kapitoly je dôkaz vety uvádzajúcej rovnosť[12]:

$$\frac{1}{M}\sum_{r=0}^{M-1}\kappa(T^{(r)},T') = \chi(T,T')$$

Uvedená rovnosť vyjadruje vzťah medzi indexami κ a χ. Bezprostredným dôsledkom uvedenej rovnosti je:

$$\frac{1}{M}\sum_{r=0}^{M-1}\kappa(T^{(r)},T) = \psi(T) \tag{1.4}$$

Posledná rovnosť (1.4) dáva do priameho vzťahu indexy κ a ψ.

Napokon v kapitole „Das Kappa-Phi-Theorem“ ([2], str. 330) autor zavádza Φ index a zdôvodňuje jeho zavedenie. Pritom Φ index presne zodpovedá dnešnému chápaniu IC, t.j. indexu koincidencie. Postup a dôvody jeho zavedenia sú nasledovné. V prípade, že skúmané texty sú identické $(T = T')$, tak platí:

$$\kappa(T^{(0)},T) = \kappa(T,T) = 1 \tag{1.5}$$

Naproti tomu je pre $r \neq 0$ hodnota $\kappa(T^{(r)},T)$ väčšinou podstatne nižšia. Preto pri počítaní priemernej hodnoty predstavuje prípad $r = 0$ extrém a je vhodné ho vynechať a počítať len zvyšných $M-1$ prípadov:

$$\frac{1}{M-1}\sum_{r=1}^{M-1}\kappa(T^{(r)},T)$$

[12]Táto veta vyjadruje len vzťah medzi κ a χ indexom na skúmaných textoch $T^{(r)}, T'$. Preto nie je podstatné akými zdrojmi sú tieto texty generované.

Odtiaľ potom nasledovnou úpravou dostaneme:

$$\frac{1}{M-1}\cdot\sum_{r=1}^{M-1}\kappa(T^{(r)},T)=\frac{1}{M-1}\cdot\left(\sum_{r=0}^{M-1}\kappa(T^{(r)},T)-1\right)=...$$

Posledná úprava vyplýva z použitia rovnosti (1.5). Ďalej pomocou dôsledku (1.4) „Kappa-Chi vety“, ktorý vyjadruje vzťah medzi $\kappa(T^{(r)},T)$ a $\psi(T)$ dostaneme:

$$...=\frac{1}{M-1}\cdot(M.\psi(T)-1)=\frac{1}{M-1}\cdot\left(\sum_{i=1}^{N}\left(\frac{m_i^2}{M}\right)-1\right)=...$$

Posledná rovnosť vyplýva z definície $\psi(T)$ v (1.3). Ďalej dostávame:

$$...=\frac{1}{M-1}\cdot\frac{1}{M}\cdot\left(\sum_{i=1}^{N}\left(m_i^2\right)-M\right)=\frac{1}{M-1}\cdot\frac{1}{M}\cdot\left(\sum_{i=1}^{N}\left(m_i^2-m_i\right)\right)=$$

Posledná úprava využíva fakt, že $\sum_{i=1}^{N}m_i=M$. Napokon dostaneme:

$$...=\frac{1}{M-1}\cdot\frac{1}{M}\cdot\left(\sum_{i=1}^{N}m_i.(m_i-1)\right)$$

Preto si definujeme nový index (náhodnú premennú) ako:

$$\Phi(T)=\sum_{i=1}^{N}\frac{m_i.(m_i-1)}{M.(M-1)}$$

pre ktorý platí:

$$\Phi(T)=\frac{1}{M-1}\cdot\sum_{r=1}^{M-1}\kappa(T^{(r)},T)$$

Používanie indexu $\Phi(T)$ má oproti $\psi(T)$ tú výhodu, že prípady $m_i=0$ a $m_i=1$ nám k počítanej sume nič nepridávajú. Inými slovami, málo početné znaky nám neovplyvnia výslednú hodnotu. Preto je Φ index vhodnejší pre kratšie texty.

Ako už bolo spomenuté skôr, Φ index je to isté ako dnešné chápanie indexu koincidencie IC a líši sa to od Friedmanovho „indexu koincidencie“, ktorý bol v našom (aj Friedmanovom) označení κ index. Do kryptoanalýzy tento index síce zaviedol Friedman, ale bez príslušného pomenovania. Dnešné pomenovanie IC dal Φ indexu zrejme až

Kullback ([2], str. 330) a on aj presadil jeho širšie použitie pre kryptoanalýzu. Všetky uvedené indexy (κ, χ, ψ, Φ) sú ale príbuzné a ich hodnoty sú pre zodpovedajúce si texty podobné. Ide však o to, na čo ktorý z nich je vhodný a za akých podmienok. Kullback presadil používanie Φ testu popri χ teste na základe stochastických dôvodov. Ukázal, že pre potreby odlíšenia jazyka dáva Φ test lepšie výsledky, t.j. výraznejšie odlišuje jazyky. Vzťahy medzi Φ a ψ indexom, ako aj výpočet strednej hodnoty pre Φ index sú uvedené v [2], str. 331:

$$\psi(T) = \frac{M-1}{M} \cdot \Phi(T) + \frac{1}{M}$$

$$\psi(T) - \Phi(T) = \frac{1-\Phi(T)}{M} = \frac{1-\psi(T)}{M-1}$$

Z toho potom vyplýva, že $\Phi(T) \leq \psi(T)$. Okrem toho má Φ index rovnaké vlastnosti invariantnosti voči šifrám ako ψ index. Ak je skúmaný text dĺžky M generovaný stochastickým zdrojom Q a pravdepodobnosti výskytu jednotlivých znakov sú p_i, tak stredná hodnota Φ indexu je:

$$E\left(\Phi(T)_Q^{(M)}\right) = \frac{M}{M-1} \cdot \sum_{i=1}^{N} p_i \cdot \left(p_i - \frac{1}{M}\right)$$

Na rozdiel od predchádzajúcich indexov je stredná hodnota $\Phi(T)$ závislá od dĺžky textu. Ďalej platí:

$$E\left(\Phi(T)_Q^{(M)}\right) \geq \begin{cases} \frac{1}{N} \cdot \frac{M-N}{M-1}\,, & \text{ak} \quad M > N\,; \\ 0\,, & \text{ak} \quad M \leq N\,. \end{cases}$$

Ak dĺžku textu zväčšujeme do nekonečna, tak platí:

$$E\left(\Phi(T)_Q^{(\infty)}\right) = \sum_{i=1}^{N} p_i^2 = E(\psi(T))$$

V kapitole „Symmetrische Funktionen der Zeichenhäufigkeiten“ ([2], str. 331) sa Bauer zaoberá špeciálnou triedou symetrických funkcií[13] závislých od frekvencie znakov abecedy v skúmanom texte. Takéto symetrické funkcie sú totiž invariantné voči niektorým už uvádzaným typom šifier (jednoduché substitúcie, transpozície,...). Najjednoduchšou takou funkciou je $\sum_{i=1}^{N} m_i^2$, kde m_i je počet výskytov znaku i. Bauer

[13] Symetrické funkcie sú funkcie invariantné voči permutácii svojich argumentov.

potom uvádza a čiastočne aj skúma špeciálnu triedu symetrických funkcií, ktorá je zovšeobecnením už uvedených indexov[14]:

$$\psi_a(T) = \begin{cases} \left(\sum_{i=1}^{N} \left(\frac{m_i}{M}\right)^a\right)^{\frac{1}{a-1}} & \text{pre } \ 1 < a < \infty \\ e^{\sum_{i=1}^{N} \frac{m_i}{M} \cdot \ln\left(\frac{m_i}{M}\right)} & \text{pre } \ a = 1 \\ \max_{i=1}^{N} \left(\frac{m_i}{M}\right) & \text{pre } \ a = \infty \end{cases}$$

Ak predpokladáme platnosť rovnosti: $\sum_{i=1}^{N} \frac{m_i}{M} = 1$, čo je v prípade kryptoanalýzy logický predpoklad, tak $\psi_2(T) = \psi(T)$. Naviac pre $1 \leq a \leq \infty$ platí: $\psi_a(T) = \frac{1}{N}$ práve vtedy, keď sú všetky m_i rovnaké, t.j. všetky znaky abecedy sa v skúmanom texte vyskytujú rovnako často.

Funkcia $-\mathrm{ld}\,\psi_a(T)$ sa nazýva Renyiho a-entropia textu T:

$$-\mathrm{ld}\,\psi_a(T) = \begin{cases} \frac{1}{1-a} \cdot \mathrm{ld}\,\left(\sum_{i=1}^{N} \left(\frac{m_i}{M}\right)^a\right) & \text{pre } \ 1 < a < \infty \\ -\sum_{i=1}^{N} \frac{m_i}{M} \cdot \mathrm{ld}\,\left(\frac{m_i}{M}\right) & \text{pre } \ a = 1 \\ -\max_{i=1}^{N} \mathrm{ld}\,\left(\frac{m_i}{M}\right) & \text{pre } \ a = \infty \end{cases}$$

Renyiho 1-entropia $-\mathrm{ld}\,\psi_1(T)$ je to isté ako Shannonova entropia a Renyiho 2-entropia $-\mathrm{ld}\,\psi_2(T)$ by sa podľa [2], str. 332 mala nazývať Kullbackova entropia. Všetky tieto funkcie tiež majú využitie pri kryptoanalýze.

V nasledujúcej kapitole nazvanej „Periodenanalyse“ ([2], str. 333) sa autor zaoberá využitím uvedených indexov pri určovaní dĺžky hesla pri šifrách s periodickým heslom. Tieto šifry boli veľmi dlho považované za bezpečné. Až v roku 1863 publikoval pruský dôstojník F. W. Kasiski článok popisujúci metódu ich lúštenia[15]. Do prvej tretiny 20. storočia to bola jediná známa metóda na lúštenie šifier s periodickým heslom. V porovnaní s neskoršími metódami bola ale menej efektívna. V dvadsiatych rokoch minulého storočia zaviedol Friedman používanie κ indexu na určovanie periódy šifier s periodickým heslom a v tridsiatych

[14]Táto problematika súvisí aj s tzv. f-divergenciami. To sú funkcie vyjadrujúce podobnosť, resp. nepodobnosť pravdepodobnostných modelov. f-divergenciám je venovaná kapitola „Divergencia pravdepodobnostných modelov“ v knihe [43] (str. 75–91).

[15]Ako sa neskôr ukázalo, pravdepodobne už 10 rokov pred Kasiskim podobnú metódu používal aj Babbage. Babbage ale svoje výsledky nepublikoval a o tom, že Kasiského metódu poznal, existujú len indície.

rokoch Kullback túto metódu vylepšil zavedením Φ indexu. Obe spomenuté metódy lúštenie šifier s periodickým heslom nielen zefektívnili, ale umožnili ho aj realizovať strojovo. Dodnes je index koincidencie, čiže Φ index, najúčinnejší nástroj na lúštenie šifier s periodickým heslom.

Bauerova kniha pokrýva veľmi široké spektrum pokiaľ ide o kryptoanalýzu a štatistické vlastnosti jazyka a obsahuje mnoho, z nášho pohľadu zaujímavých informácií. Tu uvedené indexy sú len jedným príkladom za všetky.

4. **Kullback, S.:** *Statistical Methods In Cryptanalysis* ([32])

5. **Konheim, A.:** *Cryptography: A Primer* ([29])

Táto kniha je štandardná učebnica kryptológie. Jej prednosťou pre nás je to, že sa zaoberá aj niektorými štatistickými vlastnosťami textu a štatistickými metódami využiteľnými pri kryptoanalýze. Naviac autor problematiku detailne vysvetľuje a ilustruje mnohými úlohami.

Ako príklad použitia štatistických metód v kryptoanalýze tu z uvedenej knihy spomenieme časť o lúštení polyalfabetických šifier ([29], str. 143–189), ktorá sa zaoberá rovnakými problémami ako sme uvádzali aj pri Friedmanových knihách a Bauerovej knihe.

V kapitole „The Phi Test“ autor popisuje Φ test podobne ako je to spravené aj v Bauerovej knihe. Za zmienku stojí to, že Φ index definuje nasledovným spôsobom:

$$\Phi(X) = \sum_{0 \le t < m} N_t\,(X)(N_t(X) - 1)$$

kde $X = (x_0, x_1, ..., x_{n-1})$ je skúmaný text pozostávajúci z n znakov a $N_t(X)$ označuje počet výskytov znaku $t \in \mathbb{Z}_m$ v texte X. Ako z uvedeného vidno, tento Φ index sa odlišuje od Φ indexu zavedeného v Bauerovej knihe, t.j. od indexu koincidencie. Rozdiel je ale len v multiplikatívnej konštante $\frac{1}{n(n-1)}$. Preto sú tieto dva indexy principiálne zhodné a možno s nimi narábať a používať ich rovnakým spôsobom.

V nasledujúcich dvoch kapitolách autor ukazuje použitie Φ testu na monoalfabetické a polyalfabetické substitúcie a podrobne odvádza strednú hodnotu a varianciu Φ indexu. Jedná sa vlastne o test toho, či je použitá šifra monoalfabetická alebo polyalfabetická, čo tento test pomerne spoľahlivo dokáže odhaliť.

V kapitole „Using Φ to Estimate the Period r“ ([29], str. 154) je na príkladoch ukázané ako sa pomocou Φ indexu určuje perióda, pri šifrách

s periodickým heslom (Vigenère, Beaufort,...). Určenie periódy je s použitím indexu koincidencie (Φ indexu) pomerne jednoduchá záležitosť. Podstatné je, aby pomer dĺžky skúmaného textu a dĺžky periódy bol dostatočne veľký.

V ďalších dvoch kapitolách autor zavádza κ index rovnakým spôsobom ako je to spravené aj v Bauerovej knihe a Friedmanových knihách. Potom ukazuje ako sa aj pomocou κ indexu určuje perióda šifry s periodickým heslom. Vzhľadom na úzku súvislosť κ a Φ indexov je táto možnosť zrejmá. Avšak v hraničných situáciach Φ test dáva lepšie výsledky.

V kapitole „Key Expansion“ sa autor zamýšľa nad tým, že ak sa pri šifrách s periodickým heslom pomer dĺžky textu a dĺžky kľúča blíži k 1, tak bezpečnosť šifry sa blíži k bezpečnosti OTP[16]. Tým sa kryptoanalýza stáva veľmi obtiažna. Preto je pri šifrovaní žiaduce predĺžiť kľúč pomocou rôznych expanzných algoritmov. Jedna z možností expanzie kľúča je prešifrovanie OT pomocou šifry s periodickým heslom viacnásobne a s kľúčmi rôznych (vhodných) dĺžok. V nasledujúcich štyroch kapitolách potom autor predvádza kryptoanalýzu dvojnásobnej Vigenèrovej šifry. Samozrejme takáto kryptoanalýza by sa dala spraviť aj pri viacnásobnej Vigenèrovej šifre, avšak jej zložitosť veľmi rýchlo narastá.

Z Konheimovej knihy ešte spomenieme kapitolu „The Variance of Φ“ uvedenú v dodatkoch ([29], str. 385). V nej je podrobne odvodený výpočet variancie Φ indexu, používanej v texte knihy.

1.3.7 Modelovanie jazyka a Shannonova entropia

Claude E. Shannon: *A Mathematical Theory of Communication*

Tento Shannonov článok [38] bol na svoju dobu revolučným dielom. Autor sa v ňom nezaoberá kryptoanalýzou, ale ako zamestnanec Bellových laboratórií skúma z matematického pohľadu prenos informácií. Článok sa zaoberá najmä prenosovými kanálmi, ich prenosovou kapacitou, kódovaním informácie a podobnými záležitosťami, ktoré z hľadiska kryptoanalýzy nie sú príliš zaujímavé. Popri tom sú ale v tomto článku rozoberané rôzne modely jazyka a definované pojmy entropia a redundancia, ktoré majú svoje využitie aj v kryptoanalýze.

V prvej časti článku nazvanej „Discrete Noiseless Systems“ sa autor zaoberá aj reprezentáciou diskrétneho zdroja správ ([38], str. 385). Pod diskrét-

[16] OTP=„One-Time-Pad“ je šifra založená na jednorazovej tabuľke, čiže istý variant Vernamovej šifry.

nym zdrojom rozumieme taký zdroj, ktorý produkuje správu po jednotlivých symboloch (bity, ASCII znaky, slová...). Toto je zaujímavé aj z hľadiska kryptoanalýzy, pretože je to ekvivalentný problém s vytváraním modelov jazyka. Shannon modeluje diskrétny zdroj správ pomocou stochastických procesov. Skúma aj modely založené na Markovovských reťazcoch a automatoch, ktoré graficky znázorňuje formou orientovaných grafov. Vzhľadom na dobu vzniku článku sú použité Markovovské reťazce obmedzené maximálne na rád 2. Na strane 388 je uvedená veľmi pekná séria šiestich príkladov správ generovaných rôznymi modelmi.

1. Model s nezávislými a rovnomerne rozloženými znakmi.
2. Model s nezávislými znakmi, ale ich frekvencia zodpovedá jazyku (angličtine).
3. Markovovský model 1. rádu, t.j. každý znak je závislý len od bezprostredne predchádzajúceho znaku a frekvencie bigramov zodpovedajú jazyku.
4. Markovovský model 2. rádu, t.j. frekvencie trigramov zodpovedajú jazyku.
5. Model založený na generovaní celých slov, pričom slová sú nezávislé a ich frekvencia zodpovedá jazyku.
6. Markovovský model 1. rádu založený na slovách, t.j. generujú sa celé slová, pričom každé slovo závisí len na bezprostredne predchádzajúcom a prechodové pravdepodobnosti zodpovedajú jazyku.

Na uvedených príkladoch je pekne vidno aproximáciu jazyka príslušným modelom. V poslednom modeli dávajú už celé časti viet zmysel aj keď text ako celok zmysel nedáva. Shannon uvádza, že dostatočne komplexným stochactickým zdrojom sa dá ľubovoľne presne modelovať diskrétny zdroj. To je jedna z oblastí, ktorá je veľmi zaujímavá z hľadiska kryptoanalýzy. Na konci tejto kapitoly Shannon píše, že skúmanie lepších aproximácií jazyka by bolo veľmi zaujímavé, ale je to príliš pracné – vzhľadom na vtedy neexistujúce dostupné a dostatočne výkonné počítače.

V kapitole „Graphical Representation of a Markoff Process“ ([38], str. 389) autor formálne definuje Markovovský proces a zavádza jeho reprezentáciu orientovanými grafmi. Táto reprezentácia má tú výhodu, že sa v nej jednoducho overuje ergodickosť procesu. Dve podmienky postačujúce na to, aby zodpovedajúci proces bol ergodický sú uvedené na strane 391. Tieto podmienky sú nasledovné:

1. Graf musí byť súvislý, t.j. nesmú existovať dve oddelené časti A a B také, že z vrcholov časti A sa nevieme dostať do vrcholov časti B.

2. Dĺžky všetkých orientovaných cyklov v grafe musia byť nesúdeliteľné.

Pokiaľ graf reprezentujúci Markovovský proces spĺňa uvedené vlastnosti, tak zodpovedajúci proces je ergodický. Na konci tejto kapitoly je ešte definovaný stacionárny Markovovský proces. Význam týchto pojmov je dnes mierne upravený, ale podstata zostáva nezmenená.

Napokon v kapitole „Choice, Uncertainty and Entropy“ ([38], str. 392) sa dostávame k tomu najpodstatnejšiemu – zavedeniu entropie. Motivácia zavedenia entropie bola nasledovná. Diskrétny zdroj správ modelujeme pomocou Markovovského procesu. Vieme nejakým spôsobom „zmerať“ koľko informácie sme vygenerovali? Predpokladajme, že máme množinu n javov, ktoré môžu nastať s pravdepodobnosťami $p_1, p_2, ..., p_n$. Tieto pravdepodobnosti sú jediná informácia, ktorú máme k dispozícii. Vieme nájsť mieru neurčitosti výsledku? Ak takáto miera existuje (označíme ju $H(p_1, p_2, ..., p_n)$), tak je rozumné od nej požadovať nasledovné vlastnosti:

1. Miera $H(p_1, p_2, ..., p_n)$ by mala byť spojitá v premennej p_i, pre $\forall i$.

2. Ak sú všetky p_i rovnaké, čiže ak $p_i = \frac{1}{n}$ pre $\forall i \in \{1, 2, ..., n\}$, tak $H(p_1, p_2, ..., p_n)$ by mala byť rastúca spolu s n.

3. Ak sa voľba výsledku dá rozložiť na po sebe nasledujúce voľby, tak pôvodná $H(p_1, p_2, ..., p_n)$ by mala byť váženým súčtom čiastkových H' jednotlivých volieb.

V Appendixe II ([38], str. 419) Shannon dokázal nasledovnú vetu:

Veta 2 *Jediná funkcia $H(p_1, p_2, ..., p_n)$ spĺňajúca tri podmienky uvedené vyššie je funkcia:*

$$H(p_1, p_2, ..., p_n) = -K. \sum_{i=1}^{n} p_i . \log_a p_i$$

kde $K > 0$ je konštanta a základ logaritmu je $a > 1$.

Funkcia $H(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i . \mathrm{ld}\, p_i$, kde $\mathrm{ld}\, x = \log_2 x$, sa nazýva *entropia* (alebo *Shannonova entropia*) a hrá dôležitú úlohu v teórii informácií, kde vyjadruje mieru neurčitosti. Ako sa neskôr ukázalo, táto veličina je dôležitá aj pri štatistickej analýze textov a je pre jednotlivé jazyky charakteristická podobne ako index koincidencie.
Pre entropiu, ktorú v ďalšom texte budeme pre stručnosť označovať len symbolom H, platí:

1. $H = 0$ práve vtedy, ak niektoré $p_j = 1$ a pre $\forall i \neq j$ platí $p_i = 0$. Čiže ak výsledok pokusu (v našom prípade generovaná správa) vieme s istotou predpovedať, entropia (neurčitosť) je nulová.

2. Entropia je maximálna ak $p_i = \frac{1}{n}$ pre $\forall i \in \{1, 2, ..., n\}$.

3. Majme dve náhodné premenné x a y[17]. Prvá nech nadobúda m možných hodnôt i a druhá nech nadobúda n možných hodnôt j. Nech $p(i, j)$ je pravdepodobnosť, že prvá premenná má hodnotu i a súčasne druhá má hodnotu j. Potom entropia združenej náhodnej premennej bude:

$$H(x, y) = -\sum_{i,j} p(i, j).\log p(i, j)\,,$$

ale entropie jednotlivých premenných x a y samostatne budú:

$$H(x) = -\sum_{i,j} p(i, j).\log \sum_{j} p(i, j)$$

a

$$H(y) = -\sum_{i,j} p(i, j).\log \sum_{i} p(i, j)\,.$$

Z toho sa potom dá ukázať platnosť nerovnosti:

$$H(x, y) \leq H(x) + H(y)$$

pričom rovnosť platí len vtedy, keď sú náhodné premenné x a y nezávislé, t.j. $p(i, j) = p(i).p(j)$.

4. Akékoľvek vyrovnávanie hodnôt pravdepodobností $p_1, p_2, ..., p_n$ zväčšuje výslednú entropiu. Pod vyrovnávaním hodnôt tu rozumieme „spriemerovanie“ hodnôt p_i pomocou váhových indexov a_{ij}:

$$p'_i = \sum_{j} a_{ij} p_j$$

pričom pre váhové indexy platí $\sum_i a_{ij} = \sum_j a_{ij} = 1$ a $a_{ij} \geq 0$. Pri takomto vyrovnávaní hodnôt p_i entropia rastie vždy s výnimkou prípadu, keď sa jedná len o permutáciu hodnôt p_i ([38], str 395).

[17] V tejto časti budeme náhodné premenné označovať malými písmenami $x, y, ...$ v súlade so Shannonovým článkom a nie veľkými písmenami $X, Y, ...$ ako je to obvyklé.

5. Predpokladajme, že máme dve náhodné premenné x a y. Potom vieme definovať podmienenú pravdepodobnosť $p_i(j) = \frac{p(i,j)}{\sum_j p(i,j)}$ toho, že y nadobúda hodnotu j za podmienky, že x nadobúda hodnotu i pre konkrétne i, j. Následne definujeme podmienenú entropiu y ako vážený priemer entropií y pre každú hodnotu x, v závislosti od pravdepodobnosti nastatia x ([38], str. 395):

   $$H_x(y) = -\sum_{i,j} p(i,j).\log p_i(j)$$

   Táto veličina vyjadruje priemernú neurčitosť náhodnej premennej y ak poznáme x. V článku autor ukazuje, že platí rovnosť:

   $$H(x,y) = H(x) + H_x(y)$$

6. Z predchádzajúcich bodov dostávame:

   $$H(x) + H(y) \geq H(x,y) = H(x) + H_x(y) \quad \Longrightarrow \quad H(y) \geq H_x(y)$$

V článku ([38], str. 398) Shannon ešte definuje *relatívnu entropiu* ako pomer entropie a maximálnej hodnoty, ktorú môže entropia nadobudnúť. Na základe toho potom zavádza pojem *redundancie*, čo je 1 mínus relatívna entropia. Tieto veličiny sú pre nás takisto zaujímavé. Zvyšok článku sa zaoberá prenosom informácie, kompresiou, prenosovými kanálmi ďalšími záležitosťami, ktoré z hľadiska štatistickej analýzy textu nie sú podstatné.

Ako už bolo spomenuté skôr, entropia je pri analýze textu užitočná veličina. Pre jazyky je charakteristická, takže sa pomocou nej dá spraviť odhad jazyka textu (aj keď len veľmi hrubý – podstatne nepresnejší než napr. podľa IC). Pri analyzovaní neznámeho textu sa entropia skôr používa na testy toho, či sa jedná o nejaký zašifrovaný zmysluplný text, alebo len o náhodný šum. Za príklady môžu slúžiť Voynichov manuskript alebo Rohonczi kódex. Okrem toho sa pomocou entropie dá definovať pseudometrika:

$$\varrho(x,y) = H_x(y) + H_y(x)\,.$$

Podobným spôsobom sa dá pomocou entropie a podmienenej entropie definovať aj vzdialenosť rozdelenia pravdepodobností skúmaného textu od rovnomerného rozdelenia. Tieto problémy skúmal už Kullback a sú rozoberané aj v [16] ako aj v prednáškach „Teoretické základy informatiky“ (Grošek, Mikuš).

1.3.8 Určovanie entropie jazyka

Claude E. Shannon: *Prediction and Entropy of Printed English*

Po zavedení Shannonovej entropie v článku *A Mathematical Theory of Communication* [38], ktorý bol spomenutý v predošlej časti, nastal problém ako určiť entropiu jazyka. Problémom určovania entropie jazyka sa zaoberá mnoho článkov a publikácií a podrobnejšie to spomenieme aj my v druhej kapitole tejto práce. Tažkosť pri určovaní entropie jazyka spočíva v tom, že pri jej výpočte nie je možné mechanicky použiť Shannonov vzorec na výpočet entropie a podmienenej entropie textu.

Shannon sám sa týmto problémom zaoberal v článku [39] a český preklad jeho článku je uverejnený v zborníku [45] na stranách 75–88. Z hľadiska entropie textu sa tu v porovnaní s [38] nenachádza nič nové. Zaujímavý je ale experimentálny spôsob určovania entropie jazyka. Entropia sa tu určuje na základe hádania textu a počtu pokusov a omylov potrebných na rekonštrukciu textu. V článku je tento spôsob podrobne popísaný a sú v ňom uvedené príklady aj experimentálne získané výsledky. Začína sa hádaním znakov, a potom sa preberá aj hádanie bigramov, trigramov, celých slov a porovnávajú sa získané výsledky. Takto sa dajú stanoviť spodná aj horná hranica a na ich základe potom pomerne presne určiť entropia jazyka. Pri analyzovaní zašifrovaného textu sa v článku popisovaná metóda dá použiť na určenie počtu tzv. *nepravých kľúčov* ([18], str. 62). Naviac je to zaujímavý spôsob analýzy otvoreného textu.

1.3.9 Entropia a podmienená entropia

V časti 1.3.7 na strane 37 sme si už spomenuli články, v ktorých bol pojem entropie prvý raz spomenutý v súvislosti s kódovaním textu. Teraz si stručne ukážeme ako sa teoreticky počíta entropia textu.

Entropia, nazývaná aj *miera neurčitosti* alebo *miera/stupeň chaosu*, je veličina vyjadrujúca množstvo informácie, ktorú sme schopní získať z nejakého zdroja. Jej zavedenie spadá do 19. storočia a súvisí s termodynamikou. Do informatiky bola entropia zavedená až v polovici 20. storočia.

Existuje viacero typov entropie. My sa tu budeme zaoberať výlučne entropiou definovanou Shannonom v jeho známom článku [38] z roku 1948. Túto entropiu budeme nazývať *Shannonova entropia* alebo *entropia textu*. Zavedenie Shannonovej entropie súvisí s teóriou kódovania a veľkosťou kódu potrebného na prenos informácie.

Pre ilustráciu si predstavme, že máme nejaký zdroj, ktorý nám generuje znaky TSA abecedy, pričom znak a_i má pravdepodobnosť výskytu p_i. Potom

Shannonovu entropiu takéhoto zdroja vypočítame podľa vzorca:

$$H_1 = -\sum_{\forall i} p_i . \log_2 p_i \tag{1.6}$$

Uvedený vzorec pre entropiu platí v prípade, že po prvé znaky abecedy kódujeme binárne a po druhé generovanie jednotlivých znakov zdrojom je navzájom nezávislé. V takom prípade entropia H_1 vyjadruje priemerné množstvo informácie na jeden znak v bitoch. Vo vzorci (1.6) sa entropia obvykle označuje len ako H. V tomto prípade sme sa odchýlili od obvyklého označenia, aby sme boli kompatibilní s označením, ktoré budeme používať v ďalšej časti.

Moderné počítače takmer výlučne používajú binárne kódovanie[18], a preto sa dvojkový logaritmus pri výpočte entropie považuje takmer za samozrejmosť. Ak by sme ale používali ternárny kód, tak by sme entropiu museli rátať pomocou trojkových logarimov, v prípade oktálneho kódu pomocou osmičkových logaritmov atď.

Ku kódovaniu znakov sme sa už vyjadrili. Ako to je ale so vzájomnou nezávislosťou generovaných znakov. V prípade znakov zmysluplného textu, v akomkoľvek jazyku, tieto znaky nie sú navzájom nezávislé, ale veľmi silno závisia na predchádzajúcich znakoch. Existujú dokonca kombinácie znakov, ktoré nemôžu v danom jazyku nasledovať bezprostredne za sebou, alebo zriedkavejšie existujú kombinácie znakov, ktoré v danom jazyku musia nasledovať po sebe[19]. Preto sa zmysluplný text nedá dobre modelovať pomocou generátora, ktorý jednotlivé znaky generuje navzájom nezávislé. Zmysluplnému textu podstatne lepšie zodpovedá markovovský generátor.

Pri počítaní entropie zmysluplného textu nestačí preto uvažovať jednotlivé znaky ako je to vo vzorci (1.6), ale je nutné zohľadniť aj ich závislosť na predchádzajúcich znakoch. Tu vstupuje do hry *podmienená entropia*:

$$H_n = -\sum_{\forall i,j} p(b_i, j) . \log_2 p_{b_i}(j) \tag{1.7}$$

Podmienenú entropiu n-gramov tiež zaviedol Shannon. Uvedený vzorec je prevzatý z článku [39] a význam označení je nasledovný:

- b_i je $(n-1)$-gram znakov skúmaného textu
- j je znak skúmaného textu

[18] Výnikou je napr. ruský počítač *Setun* z roku 1958, ktorý používal trojkové kódovanie, tzv. ternárny počítač. Ternárne kódovanie bolo v mnohých smeroch podstatne výhodnejšie, avšak hlavne z technických dôvodov ho vytlačilo binárne kódovanie.

[19] Napríklad v angličtine `q+u`.

- (b_i, j) je n-gram znakov skúmaného textu, ktorý dostaneme zlepením $(n-1)$-gramu b_i a znaku j
- $p(b_i, j)$ je pravdepodobnosť výskytu n-gramu (b_i, j) v skúmanom texte
- $p_{b_i}(j)$ je podmienená pravdepodobnosť výskytu znaku j po $(n-1)$-grame b_i v skúmanom texte

Podmienená entropia H_n zo vzorca (1.7) potom vyjadruje priemerné množstvo informácie na jeden znak textu pri znalosti predchádzajúcich $(n-1)$ znakov textu. Rovnako ako pri H_1 aj H_n sa vyjadruje v bitoch.

Podmienená entropia n-gramov definovaná vzorcom (1.7) sa niekedy rozpisuje nasledovným spôsobom:

$$H_n = -\sum_{\forall i,j} p(b_i, j).\log_2 p_{b_i}(j) = -\sum_{\forall i,j} p(b_i, j).\log_2 \frac{p(b_i, j)}{p(b_i)} =$$

$$= -\sum_{\forall i,j} p(b_i, j).(\log_2 p(b_i, j) - \log_2 p(b_i)) =$$

$$= -\sum_{\forall i,j} p(b_i, j).\log_2 p(b_i, j) + \sum_{\forall i}\sum_{\forall j} p(b_i, j).\log_2 p(b_i) =$$

$$= -\sum_{\forall i,j} p(b_i, j).\log_2 p(b_i, j) + \sum_{\forall i} \log_2 p(b_i).\sum_{\forall j} p(b_i, j) =$$

$$= -\sum_{\forall i,j} p(b_i, j).\log_2 p(b_i, j) + \sum_{\forall i} p(b_i).\log_2 p(b_i)$$

V uvedených úpravách sme použili vzťah pre výpočet podmienenej pravdepodobnosti:

$$p_{b_i}(j) = \frac{p(b_i, j)}{p(b_i)}\,,$$

kde $p(b_i)$ označuje pravdepodobnosť výskytu $(n-1)$-gramu b_i. Okrem toho sme využili ešte aj fakt, že:

$$\sum_{\forall j} p(b_i, j) = p(b_i)\,,$$

čiže súčet pravdepodobností výskytu n-gramov (b_i, j) cez všetky možné znaky j sa rovná pravdedobnosti výskytu $(n-1)$-gramu b_i.
Ak si zavedieme pojem *entropie n-gramu*:

$$E_n = -\sum_{\forall i,j} p(b_i, j).\log_2 p(b_i, j)\,,$$

tak podmienenú entropiu n-gramu vypočítame zo vzorca:

$$H_n = E_n - E_{n-1} \tag{1.8}$$

Musíme si pritom uvedomiť dôležitý rozdiel medzi entropiou n-gramu E_n a podmienenou entropiou n-gramu H_n. Entropia n-gramu E_n je len pomocný pojem a pri skúmaní textov nemá zmysel sa ňou zaoberať. Je to veličina vyjadrujúca priemerné množstvo informácie na jeden n-gram v bitoch. Oproti tomu podmienená entropia n-gramu H_n, ktorú sme už spomínali skôr, je veľmi dôležitý pojem pri skúmaní textov, ktorý úzko súvisí s entropiou jazyka.

Zavedením pojmu entropie n-gramu možeme vypočítať podmienenú entropiu entropiu n-gramu podľa vzťahu (1.8). Ten ďalej môžeme prepísať do podoby:

$$\begin{aligned}
H_1 &= E_1 \\
H_2 &= E_2 - E_1 = E_2 - H_1 \\
H_3 &= E_3 - E_2 = E_3 - H_2 - H_1 \\
H_4 &= E_4 - E_3 = E_4 - H_3 - H_2 - H_1 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
H_n &= E_n - E_{n-1} = E_n - H_{n-1} - H_{n-2} - \ldots - H_1
\end{aligned} \tag{1.9}$$

Vzorec (1.9) sa v niektorých prácach uvázda ako pomôcka v súvislosti so zisťovaním podmienenej entropie n-gramu, avšak z hľadiska výpočtu to nemá žiaden podstatný prínos v porovnaní s priamym použitím vzorca (1.7).

Za predpokladu, že jazyk môžme považovať za Markovovský zdroj r-tého rádu, bude existovať limita H_n. Táto limita potom bude *entropia jazyka* a platí:

$$H = \lim_{n\to\infty} H_n = H_r \tag{1.10}$$

Čiže entropia jazyka je limita podmienených entropií n-gramov daného jazyka, pre n idúce do nekonečna. Dobrou správou je fakt, že uvedená limita veľmi rýchlo konverguje. Je to spôsobené tým, že hĺbka závislosti znakov v zmysluplnom texte, ktorú sme spomenuli v časti 1.3.4 na strane 18, nie je príliš veľká. To znamená, že od istej, nie moc veľkej, hodnoty n sa už hodnota H_n nebude príliš meniť. Zlou správou je, že na zistenie entropie danej vzorky textu je to prakticky nepoužiteľné, čo vysvetlíme v ďalšej časti.

1.3.10 Prečo sa Shannonov vzorec nedá priamo použiť na danú vzorku textu

To ako sa teoreticky dá určiť entropia jazyka sme si ukázali v predošlej časti. Entropia jazyka sa dá vypočítať podľa vzorca (1.10) ako limita podmienených

entropií n-gramov daného jazyka. Čo to však skrýva za pojmom „*entropia jazyka*“? Entropia sa zvykne nazývať aj miera neurčitosti a vyjadruje napríklad množstvo informácie na jeden znak. Ako už ukázali viacerí autori, medzi nimi sám Shannon v článkoch venovaných entropii (napr. [38], [40]), prípadne Jaglom v [20], alebo aj Kontoyiannis v [30], veľkosť entropie silne závisí od typu textu, jeho literárneho štýlu a aj od použitého jazyka, resp. veľkosti abecedy použitého jazyka. Preto hovoriť všeobecne o entropii jazyka je zavádzajúce. Entropia jazyka by sa dala najlepšie chápať tak, že si ako vzorku zoberieme úplne všetky texty napísané v danom jazyku, čo je prakticky nemožné, a na nej určíme veľkosť entropie.

Majme teda danú vzorku textu nejakého jazyka, pričom v tomto prípade pod vzorkou textu máme na mysli akúkoľvek veľkú vzorku textu, ktorú sme schopní získať a pracovať s ňou. Na takejto vzorke môžme zisťovať len entropiu a podmienenú entropiu znakov, bigramov a prípadne n-gramov pre menšie hodnoty $n \in \mathbb{N}$. O entropii, tak ako ju definuje vzorec (1.10), nemá zmysel hovoriť, pretože tú na žiadnej reálne dostupnej vzorke nie sme schopní určiť. Problém je v spôsobe akým sa entropia definuje a počíta. Vzorec (1.10) definuje entropiu jazyka, v tomto prípade vzorky textu, ako limitu podmienených entropií n-gramov[20] príslušného jazyka, opäť v tomto prípade vzorky textu. Háčik sa skrýva práve v počítaní podmienených entropií. Na ich výpočet potrebujeme poznať podmienené relatívne početnosti n-gramov príslušnej vzorky textu. Ak by sme si, pre jednoduchosť, zobrali len 26 znakovú TSA abecedu, tak:

1. Pre $n = 1$ máme 26 znakov a na zistenie ich relatívnych početností nám pre väčšinu jazykov podľa [31] postačí 300–500 znakov.

2. Pre $n = 2$ máme 676 bigramov a opäť podľa [31] nám na určenie ich relatívnych početností bude stačiť vzorka 7000–13000 znakov, čo nie je problém získať.

3. Podobne ako v predošlých bodoch, môžme pokračovať ďalej aj pre $n = 3$ až $n = 7$. Pri $n = 7$ máme vyše 8.10^9 7-gramov, čo je už dosť veľká hodnota. Na to aby sme pokryli všetky tieto 7-gramy a dostali dostatočne presné hodnoty ich relatívnych početností, potrebujeme vzorku textu s aspoň 80.10^9 až 10^{11} znakmi v príslušnom jazyku. Spracovať takúto vzorku textu je možné, ale získať takto veľkú vzorku zmysluplného textu v akomkoľvek jazyku je prakticky nemožné[21].

[20] Za predpokladu, že táto limita existuje.

[21] Najväčšie textové korpusy spomínané v článkoch o entropii, napr. v [30], majú veľkosť 500 miliónov slov. Jednalo sa o anglické korpusy. Ak si zoberieme priemernú dĺžku slova

4. Pre $n \geq 8$ je aj počet n-gramov aj veľkosť vzorky, ktorá by bola potrebná na spoľahlivé určenie ich relatívnych početností, mimo rozsah možného. Spracovanie vzorky by pre menšie hodnoty $n = 8, 9, \ldots$ ešte bolo možné, aj keď veľmi náročné. Ale získanie dostatočne veľkej vzorky zmysluplného textu je už pri týchto hodnotách nemožné.

Vezmime si ako príklad hodnotu $n = 7$ a 7-gramy. Predpokladajme, že máme vzorku textu, obsahujúcu 10 miliónov znakov. To približne zodpovedá dvadsiatim 300-500 stranovým románom. Na TSA abecede existuje 8 031 810 176 teoreticky možných 7-gramov. Pritom na vzorke 10 miliónov znakov máme teoreticky možných 9 999 994 7-gramov v prípade, že berieme do úvahy prekrývajúce sa 7-gramy. Pomer týchto dvoch hodnôt je zhruba 0.00124505. To znamená, že vzorkou 10 miliónov znakov sme schopní teoreticky pokryť len približne 0.12 % teoreticky možných 7-gramov. Aj ked uvážime, že zďaleka nie všetky teoreticky možné 7-gramy môžu v danom jazyku existovať, stále je táto hodnota príliš malá. To znamená, že v tabuľke relatívnych početností 7-gramov budú takmer všetky položky nulové. Pri väčších hodnotách n bude percento pokrytia teoreticky možných n-gramov na vzorke textu veľmi rýchlo konvergovať k nule pri akejkoľvek reálnej vzorke zmysluplného textu. Ak teraz uvážime ako sa počítajú podmienené relatívne početnosti a podmienená entropia tak je zrejmé, že hodnota podmienenej entropie na reálne dostupných vzorkách textu, musí pri aplikácii Shannonovho vzorca takisto veľmi rýchlo konvergovať k nule.

To by znamenalo nulovú entropiu jazyka a keďže entropia je definovaná aj ako *miera neurčitosti*, znamenalo by to, že v jazyku s nulovou entropiou vieme so 100 % istotou predpovedať nasledujúci znak textu pri znalosti predošlých n znakov textu. Také čosi samozrejme nie je pre žiadny hovorový jazyk možné, takže entropia jazyka pre žiaden hovorový jazyk nie je a nemôže byť nulová. Napriek tomu sa nájde dosť vedeckých článkov, kde sa entropia jazyka počíta presne týmto spôsobom a kde sa dokazuje, že je nulová. Ako príklad si môžme uviesť článok [36], str. 43. Tam autorka zisťuje a porovnáva pomocou Shannonovho vzorca entropiu ruštiny a kazachštiny. Sama v texte uvádza, že skúma 500 znakové vzorky textov vedeckého, obchodného, žurnalistického atď. charakteru. Na týchto 500 znakových vzorkách potom počíta H_1 až H_6, pričom z textu nie je úplne jasné, či sa jedná o podmienenú entropiu n-gramov, alebo len entropiu n-gramov. Pravdepodobne ide o druhý uvedený prípad. V každom prípade ale 500 znaková vzorka nie je postačujúca už ani pre hodnotu $n = 2$ bez ohľadu na to, ktorú z týchto entropií autorka zisťuje.

5 znakov, čo je viac než angličtina má, tak je to 2 500 miliónov znakov. Najväčšie korpusy teda majú okolo 2.5 miliardy znakov a my by sme pre $n = 7$ potrebovali vzorky 80 až 100 miliárd znakov.

Na strane 43 sú dve tabuľky pre ruštinu a kazachštinu, z ktorých vidno, že s rastúcim n entropia prudko klesá k nule. Z toho potom autorka vyvodzuje ďalšie závery a porovnáva rôzne štýly textu v uvedených dvoch jazykoch. Len na okraj k tomuto článku ešte treba podotknúť, že autorka počítala entropiu pre ruštinu s 31 znakovou ruskou abecedou a pre kazachštinu so 43 znakovou kazachskou abecedou.

Iným príkladom nekorektného počítania entropie je článok [19]. V tomto článku sa nejedná o problém príliš krátkej vzorky textu, pretože autor zvolil iný prístup a entropiu zisťuje na modely jazyka. Hneď v úvode ale má chybu v definícii entropie jazyka a spôsobe jej počítania. Pritom sa odvoláva na Shannonov článok [38] a vzorec, avšak nesprávne ho prevzal a použil. V závere článku je potom tabuľka (Table X) entropií pre H_1 až H_{18} a tieto rovnako ako v [36] konvergujú k nule.

1.3.11 Ako sa dá určiť entropia vzorky textu

Za predpokladu, že máme nejakú väčšiu vzorku textu, môžme odhadnúť jej entropiu[22] jedným z nasledujúcich spôsobov:

1. Psychologicky, pomocou experimentu s reálnymi ľuďmi.
2. Na základe modelu jazyka zostaveného podľa danej vzorky.
3. Pomocou odhadov založených na kompresných algoritmoch.

Všetkými uvedenými postupmi sa už zaoberali viacerí autori. Psychologický pokus určovania entropie s reálnymi ľuďmi realizoval a popísal napr. Shannon v článku [39]. Jeho postup upravili a výsledky vylepšili Cover a King v článku [3]. Určovaním entropie na základe modelov jazyka sa opäť zaoberal už Shannon v [38] a aj viacerí ďalší autori. A rovnako aj určovaním entropie na základe odhadov postavených na kompresných algoritmoch sa zaoberali viacerí autori, napríklad Kontoyiannis v [30].

[22] Máme na mysli entropiu v zmysle vzorca (1.10).

Kapitola 2

Zisťovanie smeru čítania textu

Pri práci so zašifrovanými historickými dokumentami sa často stretávame s problémom určenia správneho smeru čítania textu. Príkladmi takéhoto typu problému sú Voynichov manuskript (Obr. 2.1) a Rohonczi kódex (Obr. 2.2).

Obr. 2.1: Ukážka strany Voynichovho manuskriptu.

V oboch uvedených prípadov nevieme:

- či sa jedná o nejaký zmysluplný zašifrovaný text alebo o hoax,
- aký typ šifry bol použitý,

Obr. 2.2: Ukážka strany Rohonczi kódexu.

- akým jazykom je písaný otvorený text (pokiaľ ide o zmysluplný text),
- ktorým smerom sa text číta.

Našou snahou teraz nebude riešiť všetky uvedené problémy, ale budeme sa venovať len poslednému, t.j. pokúsime sa určiť správny smer čítania textu, čo je základom k ďalším pokusom určovania zmysluplnosti a lúštenia textu. Budeme teda predpokladať, že máme k dispozícii nejaký zašifrovaný zmysluplný text a nepoznáme jazyk otvoreného textu. Pritom smer čítania textu nemusí priamo závisieť na použitom jazyku, pretože môže byť zmenený aj použitou šifrou. Čo sa však v oboch uvedených príkladoch predpokladať dá je to, že je zachovaná štruktúra dokumentu, čiže identickým miestam otvoreného textu zodpovedajú identické miesta v zašifrovanom texte. To znamená, že je použitý nejaký typ jednoduchej substitúcie.

2.1 Súčasný stav výskumu

Pokiaľ je nám známe, tak o tomto probléme doposiaľ nebolo nič publikované. Dajú sa nájsť články a publikácie, ktoré sa zaoberajú určovaním smeru čítania textu, ale všetky tieto veci sa týkajú spracovania elektronických textov, prevažne HTML stránok. V takom prípade sa nejedná o určovanie smeru čítania neznámeho textu, ale o zjednodušenú verziu problému určenia použitého jazyka otvoreného (elektronického) textu. Toto sa deje na základe použitých fontov, kódovania, testovania štruktúr z preddefinovanej sady jazykov

atď. Jedná sa teda o strojové rozpoznávanie smeru čítania textu, prípadne použitého jazyka, takže ide o úlohu, ktorá je pre človeka triviálna. Takýto problém je pomerne uspokojivo vyriešený. Je to však iný problém, než akým sa zaoberáme.

My máme text v neznámom jazyku, pričom sa môže jednať aj o zašifrovaný text v známom jazyku. O tomto texte a jazyku nemáme žiadne informácie, takže štruktúru jazyka, kódovanie a pod. využívať nemôžme. Jedinou informáciou, ktorú máme k dispozícii je samotný daný text a iba jeho štruktúra. Pekným príkladom takéhoto textu je už spomenutý Rohonczi kódex.

2.2 Riešenie problému

Pri všetkych známych jazykoch je vertikálny smer čítania textu zhora nadol. Horizontálny smer sa ale pre jednotlivé jazyky líši. Európske jazyky sa zapisujú zľava doprava. Semitské a aj niektoré ázijské jazyky sa zapisujú zprava doľava[1]. Budeme sa zaoberať otázkou, či existuje nejaký index otvoreného textu, pomocou ktorého sa dá určiť správny smer čítania textu. Inými slovami, či existuje index, ktorého hodnoty sa líšia v závislosti od smeru čítania. Ako príklad si na začiatok vezmime nasledovných pár riadkov textu[2]:

> navelkychvisacichhodinachvjedalniubaranaodbilajednapopolno
> ciplnymcistymmolovymzvukomzadunaluderadlhohucalvtichejaskoropra
> zdnejmiestnostivsetcivaznejsihostiauzdavnopoodchadzaliibasklepnik
> skuceravouhlavouakociganlenivosamotalmedzistolikmituatamzavad
> ilbokomostolickualebostolikmrastiltvarbrncalsipodfuzakoust

Ukážka textu je pre jednoduchosť vyčistená od medzier a diakritiky. Okrem toho sme takto získaný reťazec znakov náhodne rozdelili na riadky dĺžok 55 až 65 znakov[3].

Zamyslime sa teraz nad tým, na základe čoho môžeme zostrojiť index, ktorý by nám umožnil zistiť správny smer čítania textu. Aj keď experimenty budeme robiť na otvorenom texte, musíme vychádzať z predpokladu, že jazyk nepoznáme a nemôžeme preto využívať jeho štruktúru, ako je napríklad znalosť frekvencií n-gramov. Jediné, čoho sa v texte môžeme zachytiť sú zlomy riadkov. Pozrime si opäť našu stručnú ukážku textu:

> navelkychvisacichhodinachvjedalniubaranaodbilajednapopolno
> ciplnymcistymmolovymzvukomzadunaluderadlhohucalvtichejaskoropra

[1] Okrem toho niektoré ázijské jazyky, prípadne ich variácie, zapisujú text po stĺpcoch zhora nadol. Takýmito jazykmi sa ale nebudeme zaoberať.

[2] Jedná sa o úryvok z *Demokratov* od Janka Jesenského.

[3] Delenie dĺžok riadkov spôsobom 60 ± 5 nie je v tomto prípade nijako podstatné.

zdnejmiestnostivsetcivaznejsihostiauzdavnopoodchadzaliibasklepnik
skuceravouhlavouakociganlenivosamotalmedzistolikmituatamzavad
ilbokomostolickualebostolikmrastiltvarbrncalsipodfuzakoust

Pokiaľ text čítame v správnom smere (v uvedenom príklade zľava doprava), tak na zlomoch riadkov dostávame reťazce znakov, ktoré v danom jazyku dávajú zmysel. V uvedenej ukážke je takto zelenou farbou vyznačený zlom medzi 1. a 2. riadkom, na ktorom sa nachádza reťazec *„noci“*. Pokiaľ by bol skúmaný text dostatočne dlhý, je vysoký predpoklad, že tieto legitímne reťazce znakov sa budú opakovane nachádzať aj na riadkoch textu, mimo zlomov.

Keď ale text čítame v nesprávnom smere (v uvedenej ukážke zprava doľava), tak na zlomoch riadkov dostávame zväčša náhodné reťazce znakov, ktoré v danom jazyku nedávajú zmysel. Riadky textu v uvedenej ukážke majú rôznu dĺžku, ktorá bola volená náhodne. Výber reťazcov textu na zlomoch riadkov pri čítaní v nesprávnom smere zprava doľava je potom ekvivalentný s náhodnym výberom reťazcov znakov v texte a ich spájaní. V uvedenej ukážke je červenou farbou vyznačený zlom medzi predposledným a posledným riadkom, na ktorom sa nachádza reťazec *„ksts“* (čítaný v smere zprava doľava). Tento reťazec nie je v použitom jazyku legitímny, takže jeho výskyt na riadkoch textu mimo zlomov je málo pravdepodobný.

Pri hľadaní indexu umožňujúceho určenie správneho smeru čítania textu budeme vychádzať z uvedených faktov a začneme skúmaním reťazcov znakov na zlomoch riadkov v oboch smeroch čítania. Pri zisťovaní relatívnych početností n-gramov v skúmanej vzorke textu, nemôžeme tento text brať ako celok. Je to spôsobené neznalosťou správneho nadväzovania riadkov. Pri zisťovaní relatívnych početností n-gramov (čo je možné len v prípade rozsiahlejšej vzorky), resp. legitímnych[4] n-gramov v skúmanom texte, sa preto musíme obmedziť len na jednotlivé riadky bez prechodov medzi nimi. Skúmaný text prečítame teda oboma možnými smermi a v oboch prípadoch budeme zisťovať, či sa reťazce znakov na zlomoch riadkov, samozrejme vždy v príslušnom smere čítania, vyskytujú aj na riadkoch textu a či sú to teda legitímne reťazce v danom texte. V príklade na predošlej strane sme ukázali, že reťazce znakov na zlomoch riadkov, pri čítaní v nesprávnom smere, sú vlastne ekvivalentné náhodne vybratým a spojeným reťazcom znakov textu. Na základe tohoto faktu si postavíme hypotézu:

Hypotéza 1 *Pri čítaní textu v nesprávnom smere, bude na zlomoch riadkov „menej“[5] reťazcov, ktoré sa vyskytujú aj na riadkoch textu a sú teda pre príslušný text legitímne, než pri čítaní textu v správnom smere.*

[4]Pojem legitímny n-gram v skúmanom texte je zavedený v definícii 2 na strane 53.

[5]Spôsob akým to budeme merať bude upresnený až v časti 2.3.

Uvedenú hypotézu sa pokúsime potvrdiť alebo vyvrátiť skúmaním otvorených textov v rôznych jazykoch. Problémom bude nájsť spôsob ako vyhodnotiť počet výskytov opakujúcich sa legitímnych reťazcov znakov na zlomoch riadkov. Budeme hľadať vhodný index, ktorý dokáže rozlíšiť správny a nesprávny smer čítania textu. V nasledujúcich podkapitolách popíšeme nami skúmané indexy a výsledky dosiahnuté pomocou nich.

Predtým ako sa dostaneme k popisu indexov, si uvedieme niekoľko definícií.

Definícia 1 *Pod* ***dĺžkou textu****, označovanou d, budeme rozumieť celkový počet znakov skúmaného textu. Pokiaľ abeceda skúmaného textu obsahuje aj medzeru, tak aj táto sa ráta ako znak textu a medzery medzi slovami sa započítavajú do dĺžky textu.*

Definícia 2 *Pod* ***legitímnym n-gramom v skúmanom texte*** *budeme rozumieť taký n-gram, ktorý sa v skúmanom texte na riadkoch (mimo zlomov) vyskytuje aspoň raz.*

Definícia 3 *Pod* ***partíciami n-gramu na zlome riadkov*** *budeme rozumieť časti tohto n-gramu, ktoré ležia na jednom a druhom riadku jeho výskytu. Partície n-gramu na zlome riadkov budú vždy dve a súčet ich dĺžok bude n.*

2.3 Výber indexov

Aké indexy má zmysel používať na overenie postavenej hypotézy? Počítame len nájdené legitímne reťazce na zlomoch riadkov a pri ich počítaní môžeme zohľadňovať ich:

- počet,
- dĺžku,
- relatívnu frekvenciu v texte,
- spôsob rozdelenia na zlome,
- počet podreťazcov,
- prípadne ďalšie vlastnosti.

Ad hoc spôsobom sme si zvolili niekoľko indexov zohľadňujúcich tieto parametre a na otvorených textoch sme vyskúšali ich vhodnosť na určovanie smeru čítania textu.

Ako už bolo spomenuté skôr, budú sa kontrolovať reťazce na zlomoch riadkov, bude sa zisťovať, či sú tieto reťazce legitímne a následne sa bude určovať hodnota príslušného indexu.

Ihneď po začatí testovania sa veľmi rýchlo ukázalo, že má zmysel si definovať *najdlhší reťazec na zlome riadkov* nasledovným spôsobom:

Definícia 4 *Za **najdlhší reťazec na zlome riadkov** budeme považovať taký reťazec, ktorého kratšia partícia má najväčšiu dĺžku.*

Čiže za najdlhší reťazec na zlome riadkov sa nebude nutne považovať reťazec s najväčšou celkovou dĺžkou, ale reťazec, ktorého kratšia partícia má najväčšiu dĺžku. Tento spôsob definovania najdlhšieho reťazca na zlome riadkov bol potrebný preto, lebo v oboch smeroch čítania textu sa často na zlomoch vyskytujú reťazce typu[6] $1 + q$ alebo $p + q$, kde p je omnoho menšie než q. Uvedenou definíciou a vhodným zvolením indexu potom v praxi eliminujeme takmer všetky takéto nežiadúce prípady.

Ako sa počas ďalšieho testovania ukázalo, pri reťazcoch na zlomoch riadkov má zmysel zohľadnovať:

1. **Dĺžku reťazca**
   Dĺžku reťazca budeme v testovaných indexoch zohľadnovať tak, že reťazec dĺžky n budeme brať s multiplikatívnym keoficientom $n - 1$. Je to preto, lebo reťazec dĺžky n vieme na zlom riadkov umiestniť práve $n - 1$ spôsobmi.

2. **Pomer rozdelenia reťazca na zlome**
   Pomer rozdelenia reťazca na zlome riadkov budeme zohľadnovať tak, že ak je reťazec dĺžky n na zlome riadkov rozdelný na partície $n = p + q$, kde $p \leq q$, tak tento reťazec budeme brať s multiplikatívnym koeficientom $p - 1$.

3. **Relatívnu početnosť nájdeného reťazca na riadkoch**
   Relatívnu početnosť nájdených legitímnych reťazcov na zlomoch budeme zohľadňovať tak, že pre každý takýto reťazec zistíme na riadkoch textu relatívnu početnosť jeho výskytu. Táto relatívna početnosť bude potom váhou príslušného reťazca. To znamená, že reťazce s vysokou relatívnou početnosťou výskytu na riadkoch budú mať na zlomoch vyššiu váhu, než reťazce s nízkou relatívnou početnosťou výskytu na riadkoch.

4. **Podreťazce**
   Jedna alternatíva počítania reťazcov na zlomoch riadkov je brať do úvahy len najdlhší reťazec na zlome. Druhá alternatíva je brať do úvahy aj všetky jeho podreťazce, t.j. na zlome riadku budeme brať do úvahy všetky legitímne reťazce od dĺžky 2 až po najdlhší reťazec na zlome. Toto je vlastne spôsob bonifikácie dlhších reťazcov – čím bude reťazec dlhší, tým bude jeho hodnota vyššia.

[6] Nasledujúci zápis znamená, že reťazec na zlome má partície dĺžok 1 a q, resp. p a q.

Uvedené spôsoby ohodnocovania reťazcov na zlomoch riadkov si názorne ilustrujeme na krátkej vzorke textu z románu *Demokrati*:

Príklad 1:

> kedsomsemprisielbolsompozaludkovomkataremalsomvelmicitlivyzalu
> dokdoktormizakazaltazkejedlafazuluhrachsosovicukapustuudeninya
> jasomneobycajneoblubovalakuratsosovicusudenymmasomcorobitnielen
> zaludokaleicharaktersommalslabynemoholsomsazdrzatnapukalsomsa

Na uvedenom príklade krátkeho textu teraz podrobne popíšeme indexy, ktoré budeme neskôr testovať:

Index I_1
Počet opakovaní legitímnych reťazcov na zlomoch.

Toto je najjednoduchší z testovaných indexov a vyjadruje počet zlomov riadkov skúmaného textu, na ktorých sa pri zvolenom smere čítania vyskytujú legitímne n-gramy dĺžky aspoň 3.
V príklade 1 sa, v správnom smere čítania, na zlome 1. a 2. riadku nachádza legitímny reťazec „žalúdok“, ktorý je v texte príkladu vyznačený červenou farbou. Okrem toho sa na zlome 2. a 3. riadku nachádza legitímny bigram „aj“. Tento však nebudeme uvažovať, pretože jeho dĺžka je menšia než 3 a legitímne bigramy sa v oboch smeroch čítania textu vyskytujú na zlomoch riadkov aj náhodne v pomerne veľkom počte. Takže pre smer čítania zľava-doprava máme $I_1 = 1$.

Index I_2
Počet opakovaní legitímnych reťazcov na zlomoch so zohľadnením ich dĺžky.

Tento index vyjadruje počet zlomov riadkov, na ktorých sa nachádzajú legitímne n-gramy dĺžky aspoň 3 a zároveň zohladňuje dĺžku týchto n-gramov.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec. Tento reťazec, slovo „žalúdok“, má dĺžku 7 znakov, takže sa bude počítať s váhou 6. Preto pre smer čítania zľava-doprava máme $I_2 = 6$.

Index I_3
Počet opakovaní legitímnych reťazcov na zlomoch so zohľadnením ich dĺžky a pomeru rozdelenia.

Index I_3 vyjadruje počet zlomov riadkov, na ktorých sa nachádzajú legitímne n-gramy dĺžky aspoň 3, zároveň zohladňuje ich dĺžku a pomer rozdelenia. Na každom zlome sa berie do úvahy len najdlhší reťazec.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec. Tento reťazec, slovo „žalúdok“, má dĺžku 7 znakov, takže jeho dĺžka sa bude brať s váhou 6. Okrem toho jeho kratšia partícia na zlome, reťazec „dok“, má dĺžku 3, a preto sa jeho pomer rozdelenia bude bonifikovať váhou 2. Takže pre smer čítania zľava-doprava máme $I_3 = 12$.

Index I_{3V}
Počet opakovaní legitímnych reťazcov na zlomoch so zohľadnením ich dĺžky, pomeru rozdelenia a podreťazcov.

Tento index vyjadruje počet zlomov riadkov, na ktorých sa nachádzajú legitímne n-gramy dĺžky aspoň 3 a zároveň zohladňuje ich dĺžku a pomer rozdelenia. Na zlomoch riadkov sa berú do úvahy všetky legitímne reťazce, t.j. najdlhší nájdený reťazec a všetky jeho podreťazce.
Hodnota tohto indexu sa počíta rovnako ako hodnota indexu I_3. Rozdiel je len v tom, že neberieme do úvahy len najdlhší legitímny n-gram na zlome riadkov, ktorým je v našom prípade slovo „žalúdok“, ale aj všetky jeho podreťazce:

- ud (1,0)
- udo (2,0)
- udok (3,0)
- lud (2,0)
- ludo (3,1)
- ludok (4,1)
- alud (3,0)
- aludo (4,1)
- aludok (5,2)
- zalud (4,0)
- zaludo (5,1)
- zaludok (6,2)

V zátvorkách vedľa podreťazcov sú uvedené váhy zodpovedajúce ich dĺžke a pomeru rozdelenia. Ako vidno, už zo spôsobu ohodnotenia dĺžky n-gramu vyplýva, že bigramy sa eliminujú váhou 0 a netreba ich explicitne vynechávať. Takže pre smer čítania zľava-doprava máme $I_{3V} = 38$.

Index I_4
Počet kratších partícií legitímnych reťazcov na zlomoch riadkov.

Pri indexe I_4 na každom zlome riadku nájdeme najdlhší legitímny n-gram. Tento n-gram bude rozdelený na partície dĺžky $p + q = n$, kde $p \leq q$. Pre každú hodnotu $p = 1, 2, \ldots$ zistíme počet kratších partícií príslušnej dĺžky na zlomoch riadkov daného textu. Index I_4 potom vypočítame podľa vzťahu:

$$I_4 = \sum_{\forall p} p^2.k_p$$

kde k_p je počet kratších partícii dĺžky p na zlomoch riadkov daného textu.
V príklade 1 sa, v správnom smere čítania, legitímne n-gramy nachádzajú len na zlome 1. a 2. riadku (slovo „žalúdok“) a na zlome 2. a 3. riadku (reťazec „aj“). Takže pre dĺžku kratšej partície $p_1 = 1$ máme $k_1 = 1$ legitímny bigram a pre dĺžku kratšej partície $p_3 = 3$ máme tiež len jeden legitímny 7-gram a $k_3 = 1$. Preto pre smer čítania zľava-doprava máme $I_4 = 10$.

Index I_5
Počet legitímnych reťazcov na zlomoch riadkov so zohľadnením ich relatívnych početností na riadkoch.

Pri indexe I_5 nájdeme legitímne n-gramy, dĺžky aspoň 3, na zlomoch riadkov a spravíme súčet relatívnych početností výskytu týchto n-gramov na riadkoch. Na každom zlome sa berie do úvahy len najdlhší reťazec.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec. Tento reťazec, slovo „žalúdok“, má dĺžku 7 znakov. Na riadkoch textu sa tento reťazec vyskytuje len raz, na začiatku 4. riadku, a celý text má 248 znakov. Takže pre smer čítania zľava-doprava máme $I_5 = 0.004132$.

Index I_{5V}
Počet legitímnych reťazcov na zlomoch riadkov so zohľadnením ich relatívnych početností na riadkoch a podreťazcov.

Tento index sa počíta rovnakým spôsobom ako index I_5, ale okrem najdlhšieho reťazca na zlome riadkov berieme do úvahy aj všetky jeho podreťazce. Týmto spôsobom sa vlastne bonifikuje dĺžka nájdeného reťazca.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec. Tento reťazec, slovo „žalúdok“, má dĺžku 7 znakov. Všetky jeho podreťazce sú uvedené v popise indexu I_{3V}. Pre smer čítania zľava-doprava máme potom hodnotu $I_{5V} = 0.057278$.

Index I_6
Počet legitímnych reťazcov na zlomoch riadkov so zohľadnením ich relatívnych početností na riadkoch a dĺžky.

Pri indexe I_6 nájdeme legitímne n-gramy, dĺžky aspoň 3, na zlomoch riadkov a spravíme súčet relatívnych početností výskytu týchto n-gramov na riadkoch s váhou $(n-1)$, kde n je dĺžka nájdeného reťazca na zlome. Na každom zlome sa berie do úvahy len najdlhší reťazec.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec. Tento reťazec, slovo „žalúdok“, má dĺžku 7 znakov, takže jeho dĺžka sa bude bonifikovať hodnotou 6. Na riadkoch textu sa tento reťazec vyskytuje len raz a to na začiatku 4. riadku a celý text má 248 znakov. Takže pre smer čítania zľava-doprava máme $I_6 = 0.024793$.

Index I_{6V}
Počet legitímnych reťazcov na zlomoch riadkov so zohľadnením ich relatívnych početností na riadkoch, dĺžky a podreťazcov.

Tento index sa počíta rovnakým spôsobom ako index I_6, ale okrem najdlhšieho reťazca na zlome riadkov berieme do úvahy aj všetky jeho podreťazce.

V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec, slovo „žalúdok“. Všetky jeho podreťazce sú uvedené v popise indexu I_{3V}. Pre smer čítania zľava-doprava máme potom hodnotu $I_{6V} = 0.204889$.

Index I_7
Počet legitímnych reťazcov na zlomoch riadkov so zohľadnením ich relatívnych početností na riadkoch, dĺžky a pomeru rozdelenia.

Pri indexe I_7 nájdeme legitímne n-gramy, dĺžky aspoň 3, na zlomoch riadkov, pričom budeme zohľadňovať ich dĺžku, pomer rozdelenie na zlome aj ich relatívnu početnosť na riadkoch. Na každom zlome sa berie do úvahy len najdlhší reťazec.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec. Tento reťazec, slovo „žalúdok“, má dĺžku 7 znakov, takže jeho dĺžka sa bude bonifikovať hodnotou 6. Jeho kratšia partícia má dĺžku 3, takže bonifikácia bude koeficientom 2. Na riadkoch textu sa tento reťazec vyskytuje len raz a to na začiatku 4. riadku a celý text má 248 znakov. Takže pre smer čítania zľava-doprava máme $I_7 = 0.049587$.

Index I_{7V}
Počet legitímnych reťazcov na zlomoch riadkov so zohľadnením ich relatívnych početností na riadkoch, dĺžky, pomeru rozdelenia a podreťazcov.

Tento index sa počíta rovnakým spôsobom ako index I_7, ale okrem najdlhšieho reťazca na zlome riadkov berieme do úvahy aj všetky jeho podreťazce.
V príklade 1 sa, v správnom smere čítania, len na zlome 1. a 2. riadku nachádza vyhovujúci legitímny reťazec, slovo „žalúdok“. Všetky jeho podreťazce, aj s bonifikáciami dĺžky a pomeru rozdelenia, sú uvedené v popise indexu I_{3V}. Pre smer čítania zľava-doprava máme potom hodnotu $I_{7V} = 0.156347$.

Všetky uvedené indexy sú postavené tak, že pri rastúcej dĺžke a rastúcom počte riadkov skúmaného textu môžu neobmedzene rásť. Spočiatku sa zdalo, že by bolo vhodné nejakým spôsobom normalizovať tieto indexy vzhľadom na počet riadkov, alebo dĺžku textu, alebo oboje. Avšak pri testovaní sa rýchlo ukázalo, že hodnoty žiadneho z uvedených indexov nie sú invariantné pre daný jazyk, danú dĺžku textu a/alebo počet riadkov, ako je tomu napr. pri relatívnej početnosti znakov daného jazyka. Takže normalizácia testovaných indexov by nedávala žiaden zmysel.

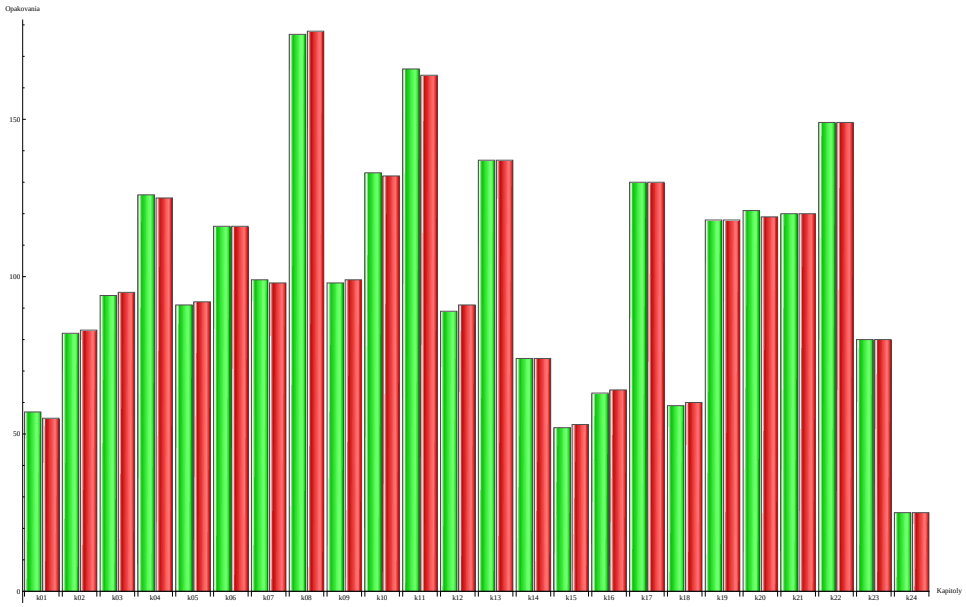
To že testované indexy s dĺžkou textu a počtom riadkov rastú v našom prípade nie je na závadu. My potrebujeme len rozlíšiť správny a nesprávny smer čítania skúmaného textu. Samotná hodnota indexu pre nás vôbec nie je podstatná. Budeme porovnávať hodnoty testovaných indexov v oboch smeroch čítania a budeme určovať, v ktorom smere čítania je táto hodnota väčšia. Takže samotná veľkosť v jednom z možných smerov čítania je irelevantná.

2.4 Ako sme testovali

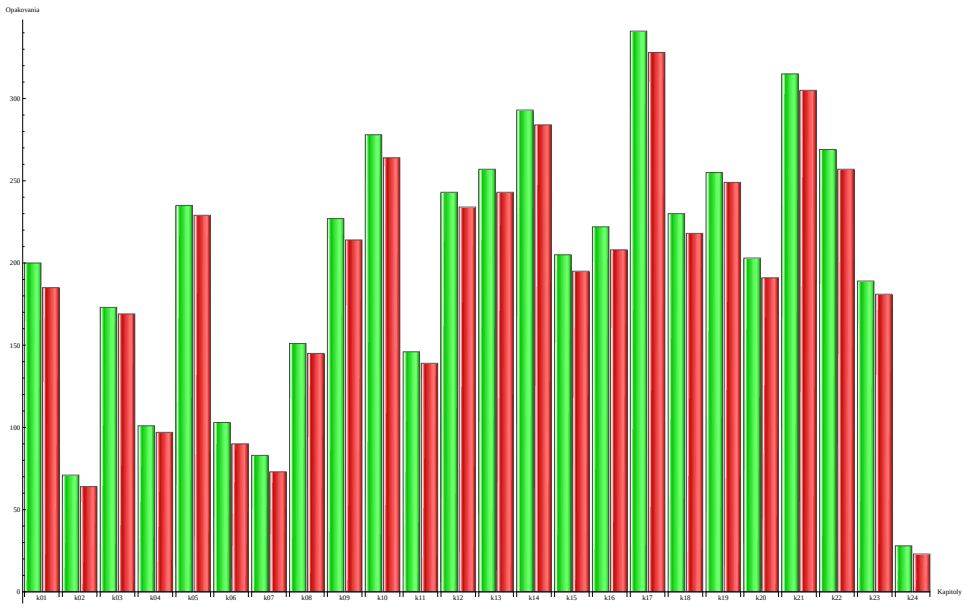
Testovanie sa začalo na slovenských textoch a až následne sa prešlo na testovanie ďalších jazykov. Najskôr sa robili testy jednorazovo, na jednotlivých kapitolách. Pritom riadkovanie textov bolo zachované. To znamenalo, že jednotlivé riadky textu končili vždy celým slovom. Na niektorých textoch boli výsledky testov, z nášho pohľadu, optimistické, na iných zasa zlé a pri niektorých textoch až katastrofálne. Pomer dobrých a zlých výsledkov testov bol zhruba vyrovnaný.

2.4.1 Nedelenie vs. delenie slov na konci riadkov

Pri podrobnejšom skúmaní príčin takýchto výsledkov sa ukázalo, že dôvodom nesprávnych výsledkov testu je práve nedelenie resp. delenie slov na koncoch riadkov. Na obrázkoch 2.3 a 2.4 (str. 61) sú graficky znázornené výsledky testu I_1 na všetkých 24 kapitolách románu *Demokrati*. Z obrázku 2.3 vidno, že v 7 prípadoch test správne určil smer čítania textu, 9 prípadov je vyrovnaných a v 8 prípadoch test nesprávne určil smer čítania textu. Je to spôsobené tým, že pri nesprávnom smere čítania, reťazec na zlome riadku je „zlepenec“ náhodne vybraných reťazcov z textu. Ak si pre jednoduchšie objasnenie zoberieme len bigram na zlome dvoch riadkov, tak tento dostaneme pri nesprávnom smere čítania zlepením dvoch náhodne vybraných znakov textu. Je zrejmé, že relatívna početnosť bigramu na zlome bude závisieť hlavne od relatívnych početností znakov v texte. Pokiaľ ale riadky vždy začínajú a končia kompletným slovom, tak bigram na zlome bude závisieť nie od re-



Obr. 2.3: Počet opakovaní legitímnych reťazcoch na zlomoch riadkov v románe *Demokrati*, pričom slová sa medzi riadkami nedelia.



Obr. 2.4: Počet opakovaní legitímnych reťazcoch na zlomoch riadkov v románe *Demokrati*, pričom text je rozdelený na riadky náhodne.

latívnych početností znakov v danom jazyku, ale od relatívnych početností znakov na začiatku a konci slov v danom jazyku. Tieto tri relatívne početnosti sa podstatne líšia[7]. Z tohoto faktu potom vyplývajú aj diametrálne rozdiely pri výsledkoch indexu I_1 medzi textami, kde riadky končia celým slovom a textami, kde sú slová na koncoch riadkov náhodne rozdelené.

2.4.2 Náhodné delenie textu na riadky

Na základe uvedených faktov, sme pozmenili spôsob testovania. V skúmaných textoch sme pred testovaním zrušili pôvodné riadkovanie a rozdelili sme ich na riadky dĺžky $60 \pm \delta$, kde δ bola náhodná celočíselná hodnota z intervalu $\langle -5, 5 \rangle$. To znamená, že dĺžky riadkov sa pohybovali medzi 55 až 65 znakmi. Pritom dochádza k deleniu slov medzi riadkami a to často aj gramaticky nesprávnym spôsobom.

Výsledky testu I_1 na kapitolách románu *Demokrati* sú zobrazené na obrázku 2.4. Z uvedeného obrázku vidno, že na všetkých 24 kapitolách test určil smer čítania textu správne. Okrem toho aj rozdiely medzi oboma smermi čítania sú väčšie než to bolo pri nerozdelených slovách (obr. 2.3).

Zvolený interval $\langle -5, 5 \rangle$ pre premennú δ a delenie riadkov podľa formuly $60 \pm \delta$ pritom nemá žiaden mimoriadny vplyv na výsledky testov. Všetky testy boli skúšané aj rôznymi inými intervalmi pre δ a zmeny vo výsledkoch testov boli zanedbateľné. Čo pre výsledky testov je podstatné je fakt, že sa na koncoch riadkov náhodne rozdelia slová. Počet zlomov riadkov v skúmanom texte má tiež vplyv na výsledky testov, t.j. väčší počet zlomov bude dávať lepšie výsledky, avšak nemá zmysel riadky príliš skracovať, pretože tým sa obmedzí počet legitímnych reťazcov v texte, čo výsledky testov zhorší. Takže formula $60 \pm \delta$ bola empiricky zvolená ako optimálna, pričom zväčšovanie intervalu $\langle -5, 5 \rangle$ pre δ má na výsledky len zanedbateľný vplyv.

2.4.3 Opakovanie testov

Ďalej sa pri testovaní ukázalo, že jeden test čítania zľava a jeden zprava na určenie smeru čítania textu stačiť nebudú. Aj pri náhodnom delení textu

[7]Pre lepšiu ilustráciu uvedeného faktu sú v tabuľke 2.1 (str. 63) uvedené porovnania relatívnych početností znakov románu *Demokrati* s relatívnymi početnosťami znakov na začiatkoch a koncoch slov. Všetky relatívne početnosti sú zaokrúhlené na desatiny percenta. Vidíme, že rozloženie znakov v prvom stĺpci zhruba zodpovedá rozloženiu znakov bežného slovenského textu. Rozloženia znakov na koncoch slov sa od prvého stĺpca veľmi výrazne líšia a rozloženie znakov na koncoch slov sa síce usporiadaním znakov na začiatku podobá bežnému usporiadaniu znakov v slovenčine, ale ak sa pozrieme na relatívne početnosti týchto znakov, tak vidíme výrazne rozdiely.

Relatívne početnosti znakov v románe *Demokrati*					
Znak	Všeobecne	Znak	Začiatky slov	Znak	Konce slov
A	11.93 %	**S**	12.26 %	**A**	19.47 %
O	9.00 %	**P**	9.56 %	**E**	12.75 %
E	8.77 %	**N**	9.24 %	**O**	10.91 %
I	7.08 %	**A**	7.59 %	**I**	9.99 %
N	5.94 %	**V**	6.66 %	**Y**	7.02 %
S	5.59 %	**Z**	6.08 %	**U**	6.94 %
T	5.27 %	**T**	6.06 %	**L**	5.92 %
L	5.22 %	**K**	5.14 %	**M**	5.26 %
R	4.45 %	**M**	4.81 %	**T**	3.69 %
K	4.27 %	**D**	4.58 %	**K**	2.75 %
V	3.99 %	**C**	4.12 %	**S**	2.05 %
D	3.84 %	**B**	3.83 %	**V**	2.03 %
U	3.38 %	**O**	3.64 %	**N**	2.02 %
M	3.38 %	**J**	3.05 %	**H**	1.97 %
C	3.28 %	**H**	2.99 %	**D**	1.90 %
P	2.87 %	**R**	2.62 %	**J**	1.73 %
Z	2.80 %	**L**	2.47 %	**Z**	1.58 %
Y	2.57 %	**U**	2.18 %	**C**	0.97 %
H	2.36 %	**I**	2.00 %	**R**	0.86 %
B	1.92 %	**E**	0.46 %	**B**	0.07 %
J	1.70 %	**F**	0.41 %	**F**	0.06 %
F	0.19 %	**G**	0.22 %	**P**	0.06 %
G	0.15 %	**Y**	0.00 %	**G**	0.03 %
X	0.01 %	**Q**	0.00 %	**X**	0.00 %
Q	0.00 %	**X**	0.00 %	**Q**	0.00 %
W	0.00 %	**W**	0.00 %	**W**	0.00 %

Tabuľka 2.1: Porovnanie relatívnych početností znakov v románe Janka Jesenského: *Demokrati* všeobecne, na začiatku a na konci slov

na riadky, popisovanom v predošlej časti, sa môže stať, že podstatná časť zlomov riadkov bude ukončená celým slovom. Prípadne sa môže jednať o text špeciálneho charakteru a už aj niekoľko málo „nevhodne“ rozdelených riadkov môže byť príčinou nesprávneho výsledku testu.

Z uvedeného dôvodu sme preto pristúpili ku mnohonásobnému opakovaniu testov v oboch smeroch čítania a ako výsledok testu sme brali aritmetický priemer výsledkov jednotlivých opakovaní. Tento postup testovania sa osvedčil. Ako vhodný počet opakovaní sme zvolili hodnotu 100, čo sa ukázalo byť optimum pomeru ceny a výkonu[8]. To znamená, že výsledky testov pri určovaní smeru čítania textu boli z nášho pohľadu veľmi dobré a časová náročnosť testov bola malá. Počet opakovaní testov zlepšuje hodnotu výsledkov, avšak už nie podstatným spôsobom a časová naročnosť narastá, aj keď len lineárne.

2.5 Testovacie texty

Na testovanie indexov sme vybrali texty v rôznych jazykoch. Vybrané jazyky boli: slovenčina, angličtina, latinčina, nemčina a hebrejčina. Hebrejčina bola pôvodne vybraná z toho dôvodu, aby sme si overili skúmané indexy aj na jazyku písanom zprava-doľava. Pri praktických testoch sa ale vyskytol problém s kódovaním a transkribciou hebrejských znakov na ACSII znaky, čo výrazne skreslilo výsledky testov. Preto sme pre testovanie použili aj texty v ostatných uvedených jazykoch, pričom sme v týchto textoch revertovali riadky, čím sme z nich spravili texty písané zprava-doľava. Teraz bližšie k výberu textov v jednotlivých jazykoch.

2.5.1 Slovenčina

Pre slovenčinu sme si vybrali text románu *Demokrati* od Janka Jesenského, ktorý je voľne dostupný v elektronickej podobe. Tento text má vyše 400 000 znakov, čize je dostatočne dlhý. Je písaný v pomerne modernej slovenčine a frekvencie znakov zodpovedajú súčasnej slovenčine. Text románu bol stiahnutý ako textový súbor a rozdelený na jednotlivé kapitoly, ktorých je 24. Celý text bol vyčistený od diakritických znamienok, čísel a prípadných špeciálnych znakov. Zostali teda len znaky TSA abecedy. Dĺžka textov jednotlivých kapitol kolísala medzi 1 710 až 20 452 znakmi. Na testovanie sme použili kompletné texty kapitol.

[8] V našm prípade sa pod pojmom cena myslí časová náročnosť a pod pojmom výkon rozumieme úspešnosť testov.

2.5.2 Angličtina

Pre testovanie angličtiny sme vybrali nasledovných päť textov:

1. **Lyndon Orr:** *Famous Affinities of History, The Romance of Devotion, The Story of Antony and Cleopatra*
2. **Arthur Train:** *Courts and Criminals*
3. **Rudyard Kipling:** *The Jungle Book*
4. **Robert Louis Stevenson:** *Treasure Island*
5. **Alexandre Dumas, Pere:** *Celebrated Crimes*

Všetky tieto texty sú voľne dostupné na stránkach projektu Gutenberg. Texty boli transkribované do TSA abecedy a z každého textu bol vybraný kompaktný úsek veľkosti 10 kB. To znamenalo, že dĺžky vybraných textov sa pohybovali medzi 10 030 až 10 058 znakmi. Drobné rozdiely v dĺžkach textov boli spôsobené rôznym počtom ich riadkov.

2.5.3 Latinčina

Pre testovanie latinčiny bolo rovnako ako v predchádzajúcom prípade zo stránok projektu Gutenberg vybraných týchto päť textov:

1. **Publi Vergili Maronis:** *Aeneidos*
2. **s. Aurelii Augustini:** *Confessiones*
3. **Gaius Iulius Caesar:** *de Bello Gallico*
4. **Gaius Iulius Caesar:** *Commentarii de Bello Gallico*
5. **Erasmus Roterodamus:** *Principally*

Texty boli transkribované do TSA abecedy a z každého textu bol vybraný kompaktný úsek veľkosti 10 kB. Dĺžky vybraných textov sa pohybovali medzi 9 973 až 10 066 znakmi.

2.5.4 Nemčina

Pre testovanie nemčiny bolo rovnako ako v predchádzajúcom prípade zo stránok projektu Gutenberg vybraných týchto päť textov:

1. **Generalfeldmarschall von Hindenburg:** *Aus meinem Leben*
2. **Georg Wilhelm Friedrich Hegel:** *Wissenschaft der Logik*
3. **Helene A. Guerber:** *Märchen und Erzählungen für Anfänger*
4. **Eduard Morike:** *Mozart auf der Reise nach Prag*
5. **Joseph M. Hägele:** *Zuchthausgeschichten von einem ehemaligen Züchtling*

Texty boli transkribované do TSA abecedy a z každého textu bol vybraný kompaktný úsek veľkosti 10 kB. To znamenalo, že dĺžky vybraných textov sa pohybovali medzi 10 026 až 10 066 znakmi.

2.5.5 Hebrejčina

Pre testovanie hebrejčiny bola použitá kniha:

- **Carl Ewald:** *Tales*

Táto kniha je voľne dostupná na stránkach projektu Gutenberg. Z nej bolo na testovanie vybraných 6 kapitol a z každej z týchto kapitol sa následne mal vybrať kompaktný úsek veľkosti 10 kB na testovanie.

Testovací program pracuje s TSA abecedou a problém pri hebrejských textoch nastal práve pri ich transkribcii do TSA abecedy. Hebrejčina má 22 znakov abecedy, spoluhlások, pričom 5 z nich môže nadobúdať dve rôzne podoby. Samohlásky sa bežne v hebrejských textoch nepíšu, avšak toto pravidlo má svoje výnimky. Okrem toho sa v textoch nachádzajú aj niektoré ďalšie znaky, ako sú interpunkcia, číslice, znaky z iných jazykov atď. Takéto znaky budeme nazývať „špeciálne“ a z testovaných textov boli pri konverzii vynechané.

Aj napriek tomu, že z uvedenej knihy sme sa snažili vybrať časti, ktoré mali čo najmenej špeciálnych znakov, tak počet rôznych znakov v týchto textoch kolísal od 41 po 93 a počet znakov v transkribovaných textoch sa pohyboval od 22 po 24. Došlo teda k výraznej strate znakov pôvodného textu, ktorá sa pohybovala od 46 % po 75 % a jej priemer bol 61 %.

Pri takmer dvojtretinovej strate počtu znakov textu sa nutne musí zmeniť aj štruktúra samotného textu a tento už potom nie je vhodný na testovanie, ktoré by sme potrebovali robiť. Žiaľ pri neznalosti hebrejčiny sa nám

nepodarilo nájsť žiadny dostatočne dlhý (t.j. okolo 10 kB) hebrejský text, ktorý by skutočne obsahoval len spoluhlásky a prípadne interpunkciu. Preto aj výsledky testov hebrejčiny sú horšie než výsledky testov iných jazykov. V podstate sme testy hebrejských textov robili len preto, aby sme si overili ako sa tieto budú správať na textoch, ktoré boli pomerne drastickým spôsobom pozmenené.

2.5.6 Revertované jazyky

Testovanie revertovaných textov, t.j. textov, ktorých smer čítania bol upravený na zprava-doľava, bolo nutné spraviť z toho dôvodu, aby sme si overili, že nami testované indexy skutočne dokážu odhaliť správny smer čítania textu. To znamená, že potrebujeme experimentálne overiť fakt, že hodnota týchto indexov je v prevažnej väčšine prípadov väčšia práve v správnom smere čítania textu a vylúčiť tým prípad, že by tieto indexy nejakou náhodou mali väčšiu hodnotu prevažne v smere čítania zľava-doprava.

Vzhľadom na problém s transkribciou hebrejčiny, sme indexy testovali aj na všetkých ostatných použitých jazykoch. V týchto textoch sme revertovali jednotlivé riadky, čiže smer čítania textu bol zmenený na zprava-doľava. Zo slovenských textov sme brali len 1., 3., 5., 9. a 10. kapitolu (5 textov) a texty sme skrátili na približne 10 kB, aby výsledky boli porovnateľné s ostatnými jazykmi. Dĺžky anglických, latinských a nemeckých textov ostali nezmenené.

Z podstaty našich testov vyplýva, že na overenie toho ako dobre alebo zle tieto testy budú určovať smer čítania textu pri jazykoch písaných zprava-doľava, nemusia byť nutne robené na hebrejčine alebo arabčine. Dá sa to overiť na akomkoľvek jazyku, ktorý je upravený tak, aby tok testovaného textu bol zprava-doľava.

2.6 Výsledky testov

2.6.1 Slovenčina

V tabuľkách na stranách 70–80 sú uvedené hodnoty indexov I_1 až I_{7V} na všetkých 24 kapitolách románu *Demokrati*. Testovanie slovenčiny sa od ďalších jazykov odlišovalo tým, že:

a. testovacích textov bolo viac (24 oproti 5),

b. testovali sa celé kapitoly a nie 10 kB časti textu.

Samotné hodnoty jednotlivých indexov uvedené v tabuľkách nie sú podstatné. Ako už bolo spomenuté skôr, s dĺžkou textu a počtom riadkov tieto hodnoty rastú a nepodarilo sa nájsť žiadnu invarianciu. Dôležitý je len pomer (porovnanie) indexov v jednom a druhom smere čítania textu. Úspešnosť jednotlivých testov bola nasledovná:

Slovenčina							
Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	16	8	66.7 %	5	14	10	58.3 %
2	19	5	79.2 %	5V	18	6	75.0 %
3	20	4	83.3 %	6	16	8	66.7 %
3V	19	5	79.2 %	6V	19	5	79.2 %
4	20	4	83.3 %	7	20	4	83.3 %
				7V	20	4	83.3 %

Z tabulky zobrazujúcej úspešnosť testov vidno, že všetky indexy dokázali pri väčšine textov správne určiť smer čítania. Úspešnosť sa pohybovala medzi 58.3 % a 83.3 %.

Zaujímavá je ale nielen samotná úspešnosť jednotlivých indexov, ale aj to ako sa tieto správali na jednotlivých kapitolách textu. Toto je vidno z tabuľky na strane 81. Z uvedenej tabuľky vidno, že na 3., 12., 20. a 22. kapitole väčšina testov zlyhala, t.j. určila smer čítania nesprávne. Môžme si teraz položiť otázku, čo bolo na týchto kapitolách iné?

1. Na prvý pohľad (z tabuľky na strane 81) je zrejmé, že kapitoly, na ktorých testy zlyhali, patrili medzi dlhšie, t.j. medzi kapitoly s nadpriemerným počtom znakov.

2. Pri podrobnejšom skúmaní zlomov riadkov a výpisov miest zlyhania testov sa ukázalo, že v týchto kapitolách sa, v porovnaní s ostatnými kapitolami, niekoľko slov a slovných spojení opakovalo s vysokou frekvenciou.

Predovšetkým druhý uvedený bod sa zdá byť kľúčovým faktorom zlyhania testov. Je pravdepodobné, že ak sa nejaké slová veľmi často vyskytujú v texte, tak aj pri nesprávnom smere čítania a náhodnom rozdelovaní textu na riadky, čo vlastne predstavuje náhodný výber dvoch miest textu, sa môžu tieto slová dostať k sebe. Takto potom dostávame falošné legitímne n-gramy.

Aby boli testy slovenských textov porovnateľné s ostatnými testovanými jazykmi, vytvorili sme z pôvodných kapitol tzv. „normované“ texty. Slovo „normované“ v tomto prípade znamená len to, že sme vypustili kapitoly, ktoré boli kratšie než 10 kB a zvyšné kapitoly sme skrátili na túto dĺžku. Text kapitol sme brali vždy od začiatku kapitoly po 10 kB a potom skrátili

tak, aby sme končili vždy celým slovom. Z tohoto dôvodu sa dĺžky kapitol môžu líšiť o niekoľko znakov. Takto sme dostali 17 normovaných kapitol, s ktorými budeme ďalej pracovať pri porovnávaní úspešnosti testov s inými jazykmi. Úspešnosť testov na normovaných kapitolách je v tabuľke 2.2:

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	12	5	70.6 %	5	8	9	47.1 %
2	12	5	70.6 %	5V	11	6	64.7 %
3	12	5	70.6 %	6	11	6	64.7 %
3V	11	6	64.7 %	6V	11	6	64.7 %
4	13	4	76.5 %	7	11	6	64.7 %
				7V	11	6	64.7 %

Tabuľka 2.2: Úspešnosť testov – Slovenčina – normované kapitoly

Výsledky testov na jednotlivých normovaných kapitolách sú na strane 82 v tabuľke 2.15. Ako vidno z tejto tabuľky, výsledky z veľkej časti zodpovedajú výsledkom na pôvodných kapitolách. Tento fakt nasvedčuje tomu, že zlyhanie väčšiny textov na 3., 12., 20. a 22. kapitole bolo spôsobené pravdepodobne tým, že sa tam vyskytovali slová a slovné spojenia s vysokou frekvenciou výskytu a nie dĺžkou samotného textu. Len v 17. a 21. kapitole sa zmenilo výsledné určenie smeru čítania textu nesprávnym spôsobom. Ale treba si všimnúť, že:

1. obe tieto kapitoly skrátením na 10 kB stratili približne 50 % svojho pôvodného textu,
2. v kapitole 17. sa v podstate zmenili len výsledky indexov I_4, I_7 a I_{7V}.

Jedine v 21. kapitole došlo k zmenám väčšiny indexov, ale tieto sa dajú vysvetliť jednak výrazním skrátením textu a jednak náhodným delením textu na riadky.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	107.07	93	123	108.03	92	123
2	46.71	36	55	46.35	37	57
3	95.0	83	109	94.72	79	112
4	64.89	56	73	63.96	51	76
5	119.81	103	137	119.37	103	141
6	67.74	57	82	66.92	53	78
7	52.18	42	64	52.92	41	63
8	89.29	78	104	88.55	75	102
9	117.71	97	135	119.06	100	136
10	141.24	118	157	141.89	122	162
11	82.15	65	95	81.91	65	99
12	130.15	112	148	129.56	109	149
13	138.78	116	162	138.24	123	158
14	146.1	126	167	146.73	125	168
15	110.57	99	127	111.5	94	127
16	115.39	97	138	113.49	97	127
17	160.64	141	182	161.04	136	182
18	117.81	96	141	117.77	95	138
19	131.7	112	148	132.91	114	150
20	112.68	98	128	108.96	92	125
21	146.52	121	167	144.43	123	164
22	139.08	119	168	139.04	117	162
23	104.54	89	118	104.45	89	125
24	20.95	14	27	20.53	15	26

Tabuľka 2.3: Test 1 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	556.34	443	654	543.64	448	652
2	119.73	75	197	116.34	63	183
3	429.71	358	518	431.93	325	519
4	191.99	135	252	183.05	125	245
5	646.47	521	772	639.32	491	730
6	175.97	113	241	168.08	112	225
7	164.67	106	240	148.29	96	219
8	343.16	251	429	334.72	245	399
9	701.03	547	849	681.57	506	805
10	776.66	658	928	755.99	629	889
11	395.42	286	496	367.47	270	465
12	632.39	531	732	636.91	507	769
13	626.81	528	746	613.86	494	709
14	893.19	692	1085	885.11	697	1017
15	585.95	473	686	573.57	458	711
16	637.58	507	737	632.89	547	755
17	1147.93	1006	1277	1137.03	960	1336
18	679.62	538	890	674.85	561	822
19	678.86	583	798	655.96	554	783
20	505.05	417	606	520.31	388	630
21	1175.66	980	1366	1176.51	1030	1369
22	726.34	512	849	735.98	602	869
23	503.54	427	611	491.71	366	594
24	24.87	4	52	23.15	3	52

Tabuľka 2.4: Test 2 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	1132.82	749	1770	1097.07	671	1475
2	187.53	93	387	180.3	69	337
3	763.63	519	1139	804.27	459	1125
4	327.55	161	521	292.12	154	464
5	1162.93	773	1558	1144.54	794	1412
6	274.76	165	409	246.7	129	388
7	260.88	145	468	228.48	123	372
8	618.65	395	941	585.17	345	848
9	1641.86	936	2605	1536.25	945	2127
10	1387.85	1000	1742	1346.73	951	1799
11	903.8	520	1459	730.98	486	1095
12	1240.48	756	1657	1263.61	861	1774
13	1032.31	768	1338	989.25	755	1261
14	2149.68	1187	3108	2083.43	1298	3077
15	1259.74	925	1717	1233.4	788	1615
16	1280.01	949	1669	1262.2	897	1733
17	2455.34	1821	3077	2431.45	1726	3054
18	1327.48	902	1955	1323.1	946	1700
19	1193.39	883	1584	1117.24	723	1533
20	945.06	626	1320	982.79	601	1448
21	2972.76	2198	3914	2915.55	2248	3954
22	1287.49	678	1697	1426.5	1053	2096
23	1166.57	668	1870	1107.12	629	1766
24	31.02	4	75	28.31	3	69

Tabuľka 2.5: Test 3 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	11748.55	4207	25618	11361.02	3707	22080
2	869.93	267	2171	807.37	151	2187
3	5435.74	1985	11528	6678.01	2139	11987
4	1787.81	464	3531	1394.96	523	3025
5	7546.09	3145	13674	7322.5	3472	11347
6	1480.15	443	3236	1142.15	325	2634
7	1164.79	488	2697	997.28	346	3820
8	4742.31	1473	10441	4241.84	1344	10527
9	30933.97	4680	91650	26088.34	6573	82819
10	8953.59	4956	13847	8795.28	4352	14834
11	11550.77	3486	25531	6078.76	2427	12666
12	12427.92	3346	21347	12566.45	6186	21707
13	5623.8	2721	8420	4938.82	3006	7710
14	49630.89	9312	112991	46747.84	7374	112548
15	13333.32	5420	22811	12951.97	4365	22229
16	11877.81	4637	19998	11527.76	4775	22893
17	23491.81	12134	36020	23804.87	11529	35275
18	9773.5	4341	15408	9664.93	5632	15097
19	8531.3	3344	15668	6813.3	3109	14412
20	8104.31	2998	17352	8528.12	2364	21785
21	44767.17	21341	74525	41727.58	20572	81367
22	8551.81	2704	21746	14260.34	5095	25709
23	26316.33	2894	66602	22312.76	2852	64395
24	76.57	14	218	64.28	3	197

Tabuľka 2.6: Test 3V na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú zvýraznené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	839.63	629	1215	814.36	550	1040
2	190.0	129	306	184.59	121	277
3	627.09	492	807	643.14	436	833
4	301.82	218	441	285.06	197	404
5	922.56	720	1161	911.26	750	1098
6	281.59	207	372	267.02	192	353
7	247.42	172	373	230.12	167	308
8	520.01	393	673	501.99	359	643
9	1094.35	768	1611	1037.38	715	1379
10	1099.21	894	1323	1079.95	856	1331
11	633.51	431	940	559.7	426	757
12	944.99	714	1166	955.74	748	1216
13	896.97	760	1104	879.22	734	1026
14	1435.33	1008	1988	1390.83	1012	1990
15	895.94	704	1173	880.78	683	1078
16	943.84	754	1170	937.74	740	1211
17	1687.95	1326	2019	1667.01	1276	1999
18	990.93	776	1313	985.66	783	1181
19	968.06	799	1193	935.64	688	1126
20	756.95	587	930	774.07	582	1025
21	1831.23	1423	2344	1802.9	1507	2384
22	1036.34	690	1259	1090.03	916	1500
23	810.86	592	1186	779.98	563	1141
24	49.35	31	84	46.86	27	74

Tabuľka 2.7: Test 4 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	0.03412890	0.028214	0.041950	0.03411160	0.026981	0.041703
2	0.02638954	0.018769	0.035451	0.02662540	0.019697	0.037301
3	0.03277611	0.027271	0.040040	0.03398587	0.027271	0.042728
4	0.03100297	0.024672	0.037250	0.03006673	0.024344	0.039862
5	0.03333753	0.027259	0.041384	0.03325094	0.027969	0.041385
6	0.03065780	0.022744	0.037902	0.03165708	0.023375	0.041849
7	0.02748464	0.019776	0.037783	0.02890956	0.023298	0.036610
8	0.03119958	0.024570	0.041418	0.03079950	0.025005	0.039352
9	0.03376366	0.028591	0.039895	0.03403115	0.026694	0.042227
10	0.03472189	0.027488	0.042241	0.03441778	0.028022	0.041349
11	0.03276163	0.027418	0.041862	0.03223912	0.024801	0.040384
12	0.03281508	0.027839	0.039936	0.03327695	0.025915	0.040417
13	0.03394340	0.028435	0.044617	0.03387369	0.027532	0.040685
14	0.03374993	0.028933	0.039428	0.03399325	0.028026	0.041299
15	0.03383821	0.028031	0.042854	0.03333820	0.025357	0.040505
16	0.03340719	0.027729	0.041254	0.03349753	0.027579	0.040283
17	0.03528569	0.028464	0.043623	0.03569882	0.029050	0.043427
18	0.03374528	0.027227	0.041308	0.03296211	0.026222	0.039582
19	0.03403278	0.027492	0.048141	0.03421521	0.028991	0.042083
20	0.03387338	0.027619	0.041754	0.03374921	0.028599	0.039957
21	0.03560597	0.028994	0.041388	0.03674421	0.029632	0.047558
22	0.03336477	0.028007	0.040988	0.03282005	0.027944	0.040184
23	0.03340029	0.026400	0.045556	0.03333138	0.026489	0.041849
24	0.01867981	0.010546	0.028112	0.01869104	0.012303	0.028116

Tabuľka 2.8: Test 5 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	0.53417577	0.455670	0.617388	0.53887722	0.448025	0.627012
2	0.18566291	0.130436	0.262766	0.18118477	0.132979	0.251864
3	0.45275618	0.376567	0.513682	0.45917407	0.369076	0.522909
4	0.24435624	0.196862	0.302900	0.23889175	0.198002	0.301093
5	0.59729638	0.519161	0.676819	0.59683200	0.519587	0.699533
6	0.24213592	0.203086	0.310787	0.23653638	0.183184	0.274940
7	0.22998965	0.184802	0.285052	0.21887706	0.161895	0.290333
8	0.37609133	0.312964	0.427106	0.36963113	0.309696	0.441678
9	0.68474553	0.559810	0.818630	0.66204208	0.501900	0.804118
10	0.65907069	0.590562	0.721153	0.64992358	0.559033	0.736622
11	0.44981058	0.381210	0.528108	0.42901596	0.332855	0.500884
12	0.60684179	0.486690	0.693404	0.61682610	0.533437	0.690370
13	0.58183808	0.509470	0.659246	0.57699139	0.511530	0.657181
14	0.70129052	0.606039	0.773853	0.70129419	0.611656	0.783318
15	0.62490668	0.526536	0.722458	0.61426192	0.481665	0.731439
16	0.60422107	0.523873	0.698469	0.59981759	0.487262	0.678589
17	1.00464798	0.822060	1.182730	0.99446549	0.822006	1.167086
18	0.75309932	0.596876	0.914778	0.73276186	0.551193	0.904437
19	0.66275620	0.591224	0.758705	0.66189607	0.560607	0.773630
20	0.52747161	0.451758	0.636510	0.53901330	0.461890	0.625975
21	0.98532189	0.848635	1.108412	0.98127670	0.824016	1.097461
22	0.59779070	0.530278	0.654969	0.59892656	0.533315	0.681127
23	0.48319567	0.422444	0.562521	0.47921112	0.394912	0.578668
24	0.07487771	0.040991	0.113049	0.06780806	0.031036	0.099566

Tabuľka 2.9: Test 5V na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	0.13167336	0.112294	0.156723	0.13103294	0.106289	0.157628
2	0.08630834	0.063741	0.113563	0.08504279	0.067665	0.104971
3	0.12174831	0.106210	0.135406	0.12508582	0.103525	0.146444
4	0.10235008	0.086282	0.125177	0.09841549	0.080885	0.118962
5	0.12982295	0.113023	0.147734	0.12915123	0.112808	0.153485
6	0.09708230	0.078186	0.114038	0.09770055	0.081665	0.114344
7	0.09384812	0.075992	0.112013	0.09656580	0.076955	0.119655
8	0.10850831	0.093505	0.128833	0.10690997	0.091653	0.120577
9	0.13904600	0.119850	0.164637	0.13770713	0.111458	0.162961
10	0.13513469	0.121140	0.158445	0.13161881	0.112334	0.151720
11	0.12596095	0.109008	0.148498	0.12124702	0.098080	0.142566
12	0.12585392	0.109919	0.150825	0.12656103	0.111845	0.144569
13	0.12414523	0.109039	0.150755	0.12267167	0.103815	0.139405
14	0.13404515	0.122381	0.149844	0.13540990	0.118809	0.159206
15	0.13424263	0.111891	0.159616	0.13054271	0.108416	0.153784
16	0.13127808	0.116677	0.148816	0.13036169	0.110173	0.152336
17	0.15179365	0.134303	0.174708	0.15226348	0.126871	0.182040
18	0.13918075	0.111577	0.171285	0.13438914	0.115533	0.158501
19	0.12994370	0.111994	0.157203	0.12992122	0.116945	0.151998
20	0.12715566	0.112614	0.151925	0.12874142	0.110482	0.150532
21	0.16342316	0.136894	0.183774	0.16461904	0.141793	0.187863
22	0.12816324	0.115682	0.144800	0.12635368	0.108694	0.142760
23	0.12528719	0.104462	0.152054	0.12439545	0.104106	0.151705
24	0.05245800	0.032816	0.071474	0.05168806	0.036309	0.079101

Tabuľka 2.10: Test 6 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	1.74769283	1.394510	2.079762	1.76217874	1.306570	2.119113
2	0.55768051	0.352657	0.887450	0.53670000	0.336651	0.800026
3	1.37322933	1.083642	1.669809	1.42468494	1.044239	1.675137
4	0.73532000	0.529000	0.953930	0.69105296	0.543254	0.914053
5	1.75826988	1.434952	2.102129	1.75358095	1.428989	2.122477
6	0.66269303	0.524622	0.836897	0.62968265	0.473636	0.768296
7	0.67666862	0.500220	0.897405	0.63030580	0.444440	0.937066
8	1.11548142	0.895941	1.367172	1.07834323	0.823591	1.323551
9	2.55388625	1.788265	3.478976	2.42864790	1.563793	3.465253
10	1.88778020	1.581565	2.129041	1.84577559	1.522088	2.191801
11	1.69078177	1.237380	2.179239	1.44251389	1.111242	1.770391
12	2.06577392	1.333562	2.650477	2.09863516	1.564167	2.510414
13	1.57607897	1.356905	1.831142	1.55870840	1.351412	1.808118
14	2.37043596	1.656971	2.912776	2.36281139	1.826249	2.762424
15	2.34477370	1.678824	2.933766	2.28359979	1.615787	2.845529
16	1.89372054	1.513206	2.304781	1.87644438	1.522091	2.245297
17	3.58446788	2.764056	4.414356	3.55064663	2.646305	4.386265
18	2.65227043	1.901922	3.330972	2.58237627	1.806878	3.395326
19	1.88389527	1.588185	2.192544	1.84881112	1.557856	2.296882
20	1.65932900	1.361933	2.158170	1.71214744	1.276517	2.192126
21	3.69354189	2.955734	4.198952	3.67108712	2.993140	4.368636
22	1.70974784	1.389420	1.981367	1.79766940	1.533251	2.067731
23	1.82376980	1.266446	2.377816	1.76861961	1.119518	2.367249
24	0.19088044	0.094875	0.314059	0.17169137	0.071448	0.264819

Tabuľka 2.11: Test 6V na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	0.12468559	0.079576	0.180998	0.11981149	0.064434	0.168878
2	0.05155814	0.023415	0.101136	0.04924468	0.016687	0.086532
3	0.09668920	0.061382	0.130264	0.10086864	0.052540	0.138631
4	0.06581246	0.040048	0.095807	0.05917395	0.033506	0.104467
5	0.10963014	0.073065	0.141590	0.10834529	0.073912	0.148981
6	0.05413097	0.029546	0.081383	0.04885976	0.025747	0.072855
7	0.06047115	0.034279	0.099155	0.05680583	0.031927	0.091103
8	0.07936442	0.050460	0.114998	0.07587161	0.044587	0.104013
9	0.15437724	0.095868	0.219321	0.14370710	0.087359	0.190268
10	0.11327794	0.080749	0.144487	0.10877988	0.078608	0.144138
11	0.13266504	0.073417	0.203511	0.10955732	0.070107	0.159973
12	0.11651183	0.064763	0.189485	0.11912250	0.082993	0.192238
13	0.08825899	0.065975	0.108607	0.08654529	0.061585	0.109246
14	0.14693893	0.090506	0.201618	0.14406244	0.099025	0.203149
15	0.13908154	0.096354	0.197851	0.13371689	0.081682	0.185606
16	0.12145965	0.081568	0.171523	0.12179846	0.087092	0.155206
17	0.16726386	0.118277	0.224103	0.16581726	0.128015	0.233342
18	0.13767207	0.098228	0.207831	0.13395539	0.094861	0.196482
19	0.10425416	0.081197	0.130262	0.10086608	0.065483	0.130450
20	0.10747718	0.075860	0.147726	0.11398780	0.066939	0.168638
21	0.21464386	0.170525	0.278316	0.21027863	0.162161	0.278318
22	0.10795627	0.065603	0.141311	0.11442690	0.084528	0.158832
23	0.12701678	0.077190	0.188834	0.12188811	0.069066	0.183935
24	0.01947010	0.004689	0.047515	0.01795546	0.001757	0.042230

Tabuľka 2.12: Test 7 na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Kapitola	Zľava	min_L	max_L	Zprava	min_P	max_P
1	1.34488900	0.555627	2.556827	1.31585755	0.543364	2.172575
2	0.27079073	0.087385	0.626365	0.25314709	0.049130	0.641071
3	0.78061473	0.312789	1.432402	0.91238263	0.301475	1.523070
4	0.37802741	0.151191	0.712509	0.30861935	0.141032	0.603897
5	0.88466200	0.409633	1.328272	0.88175812	0.425188	1.321792
6	0.28449128	0.100148	0.566723	0.22759832	0.074397	0.477977
7	0.30530245	0.131415	0.637009	0.26663868	0.109468	0.811501
8	0.64683884	0.282480	1.275912	0.58789729	0.248796	1.308948
9	2.98789235	0.632957	7.351727	2.56740794	0.841210	6.878124
10	0.88869599	0.517763	1.22491	0.85934647	0.489216	1.305427
11	1.87336818	0.693799	3.314570	1.06795856	0.455199	1.990747
12	1.86028884	0.419240	3.262043	1.90062741	0.800721	2.908021
13	0.58933147	0.366848	0.831531	0.56443179	0.358789	0.808064
14	3.14880515	0.750183	6.814558	2.98129784	0.693288	6.761130
15	2.24292187	0.919300	3.518630	2.13228781	0.886787	3.558157
16	1.25479560	0.561670	1.944521	1.23892284	0.642907	1.939983
17	2.79499337	1.634733	3.810541	2.75842385	1.713167	4.065297
18	1.87320072	0.798685	2.793273	1.83892630	0.933548	2.885678
19	0.88046783	0.428563	1.321284	0.77425977	0.372551	1.349215
20	1.10151503	0.553295	2.055909	1.13012534	0.341836	2.284476
21	3.75092816	2.395664	5.659665	3.60683167	2.409938	5.738056
22	0.84350379	0.322623	1.778339	1.22122244	0.576090	2.081897
23	2.81808599	0.538867	6.361750	2.46412070	0.431559	6.497068
24	0.04933816	0.008204	0.136683	0.04081031	0.001757	0.117354

Tabuľka 2.13: Test 7V na kapitolách románu *Demokrati*

Vysvetlivky:

- Skúmaný text sa rozdeľuje na riadky dĺžky $60 + \delta$, kde δ je náhodná hodnota z intervalu $\langle -5, 5 \rangle$.
- Test sa opakuje 100 krát na náhodne rozdelených vstupoch a v tabuľke je ako výsledok uvedený priemer z týchto 100 testov.
- Hodnoty min_L, max_L, min_P a max_P sú minimálne a maximálne hodnoty dosiahnuté v testoch pri čítaní textu zľava, resp. zprava.
- Prípady zlyhania testu sú vyznačené červenou farbou.

Test / Kapitola	Počet znakov	1	2	3	3V	4	5	5V	6	6V	7	7V
1	12161	Z	O	O	O	O	O	Z	O	Z	O	O
2	4319	O	O	O	O	O	Z	O	O	O	O	O
3	10418	O	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
4	6124	O	O	O	O	O	O	O	O	O	O	O
5	14091	O	O	O	O	O	O	O	O	O	O	O
6	6335	O	O	O	O	O	Z	O	Z	O	O	O
7	5111	Z	O	O	O	O	Z	O	Z	O	O	O
8	9202	O	O	O	O	O	O	O	O	O	O	O
9	13715	Z	O	O	O	O	O	O	O	O	O	O
10	16812	Z	O	O	O	O	O	O	O	O	O	O
11	8794	O	O	O	O	O	O	O	O	O	O	O
12	14552	O	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
13	15513	O	O	O	O	O	O	O	O	O	O	O
14	17631	Z	O	O	O	O	Z	Z	Z	O	O	O
15	12348	Z	O	O	O	O	O	O	O	O	O	O
16	13384	O	O	O	O	O	O	O	O	O	O	O
17	20452	Z	O	O	Z	O	Z	O	Z	O	O	O
18	13924	O	O	O	O	O	O	O	O	O	O	O
19	15354	Z	O	O	O	O	Z	O	O	O	O	O
20	12242	O	Z	Z	Z	Z	O	Z	Z	Z	Z	Z
21	18802	O	Z	O	O	O	Z	O	Z	O	O	O
22	16179	O	Z	Z	Z	Z	O	Z	O	Z	Z	Z
23	11330	O	O	O	O	O	O	O	O	O	O	O
24	1710	O	O	O	O	O	Z	O	O	O	O	O

Tabuľka 2.14: Zhrnutie testov na kapitolách románu *Demokrati*

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

Test / Kapitola	Počet znakov	1	2	3	3V	4	5	5V	6	6V	7	7V
1	10239	O	O	O	O	O	O	O	O	O	O	O
3	10240	O	O	Z	Z	O	Z	O	Z	Z	O	Z
5	10239	O	Z	O	O	O	O	Z	O	O	O	O
9	10238	O	Z	O	O	O	Z	Z	Z	O	O	O
10	10240	Z	O	O	O	O	Z	O	O	O	O	O
12	10239	Z	O	Z	Z	Z	O	O	O	Z	Z	Z
13	10241	O	O	O	O	O	Z	O	O	O	O	O
14	10241	O	Z	O	O	O	O	Z	Z	O	Z	O
15	10240	O	O	O	O	O	O	O	O	O	O	O
16	10241	O	O	O	O	O	Z	O	O	Z	O	O
17	10241	Z	O	O	Z	Z	Z	O	Z	O	Z	Z
18	10239	O	O	O	O	O	O	O	O	O	O	O
19	10237	Z	O	O	O	O	Z	O	O	O	O	O
20	10241	O	O	Z	Z	O	Z	O	Z	Z	Z	Z
21	10237	O	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
22	10240	O	Z	Z	Z	Z	O	Z	O	Z	Z	Z
23	10240	Z	O	O	O	O	O	Z	O	O	O	O

Tabuľka 2.15: Zhrnutie testov na normovaných kapitolách románu *Demokrati*

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

2.6.2 Ďalšie európske jazyky: angličtina, latinčina, nemčina

Na strane 84 je tabuľka zobrazujúca výsledky testov pre angličtinu, latinčinu a nemčinu. Z každého uvedeného jazyka bolo použitých päť vzoriek a všetky vzorky textov mali veľkosť približne 10 kB.

Nasledujúce štyri tabuľky zobrazujú uspešnoť jednotlivých indexov pre angličtinu, latinčinu a nemčinu zvlášť a potom tieto tri jazyky spolu:

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	4	1	80 %	5	3	2	60 %
2	4	1	80 %	5V	4	1	80 %
3	5	0	100 %	6	4	1	80 %
3V	4	1	80 %	6V	4	1	80 %
4	5	0	100 %	7	5	0	100 %
				7V	5	0	100 %

Tabuľka 2.16: Úspešnosť testov – Angličtina

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	1	4	20 %	5	3	2	60 %
2	5	0	100 %	5V	5	0	100 %
3	5	0	100 %	6	4	1	80 %
3V	4	1	80 %	6V	5	0	100 %
4	5	0	100 %	7	5	0	100 %
				7V	5	0	100 %

Tabuľka 2.17: Úspešnosť testov – Latinčina

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	0	5	0 %	5	0	5	0 %
2	5	0	100 %	5V	5	0	100 %
3	5	0	100 %	6	5	0	100 %
3V	5	0	100 %	6V	5	0	100 %
4	5	0	100 %	7	5	0	100 %
				7V	5	0	100 %

Tabuľka 2.18: Úspešnosť testov – nemčina

Angličtina, Latinčina, Nemčina							
Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	5	10	33.3 %	5	6	9	40.0 %
2	14	1	93.3 %	5V	14	1	93.3 %
3	15	0	100.0 %	6	13	2	86.7 %
3V	14	1	93.3 %	6V	14	1	93.3 %
4	15	0	100.0 %	7	15	0	100.0 %
				7V	15	0	100.0 %

Z tabuliek úspešnosti je vidno, že testy 1. a 5. indexu takmer úplne zlyhali a všetky ostatné indexy v prevažnej väčšine prípadov určili smer čítania textu správne. Úspešnosť testov, mimo 1. a 5. indexu, sa pohybovala medzi 86.7 % a 100 %, čiže bola výrazne vyššia než pri testovaní slovenských textov. Okrem toho sa, na rozdiel od slovenských textov, v testovaných vzorkách nevyskytol žiaden text, na ktorom by väčšina testov zlyhala. Tieto rozdiely oproti slovenským textom sú dané rozdielmi vo výbere vzoriek. Žiaden z testovaných textov neobsahoval slová a slovné spojenia s vysokou frekvenciou výskytu.

Test / Jazyk	1	2	3	3V	4	5	5V	6	6V	7	7V
angličtina 1	O	O	O	O	O	Z	O	Z	O	O	O
angličtina 2	Z	O	O	O	O	O	O	O	O	O	O
angličtina 3	O	O	O	O	O	O	O	O	O	O	O
angličtina 4	O	O	O	O	O	Z	O	O	O	O	O
angličtina 5	O	Z	O	O	O	O	Z	O	Z	O	O
latinčina 1	Z	O	O	O	O	Z	O	Z	O	O	O
latinčina 2	Z	O	O	Z	O	O	O	O	O	O	O
latinčina 3	O	O	O	O	O	O	O	O	O	O	O
latinčina 4	Z	O	O	O	O	Z	O	O	O	O	O
latinčina 5	Z	O	O	O	O	O	O	O	O	O	O
nemčina 1	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 2	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 3	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 4	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 5	Z	O	O	O	O	Z	O	O	O	O	O

Tabuľka 2.19: Zhrnutie testov na anglických, latinských a nemeckých textoch

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

Zaujímavé boli rozdiely medzi testovanými jazykmi. Výrazná bola predovšetkým nemčina, kde na všetkých piatich vzorkách dopadli testy rovnako. Testy 1. a 5. indexu kompletne zlyhali a testy ostatných indexov mali 100 % úspešnosť. Okrem toho pri nemčine boli rozdiely medzi hodnotami indexov v správnom a nesprávnom smere čítania výrazne vyššie než pri angličtine a latinčine.

Zaujímavo dopadol ešte test 1. indexu na anglických textoch. Tam 1. index, s jedinou výnimkou, správne určil smer čítania textu, pričom na nemeckých textoch úplne a na latinských textoch takmer úplne prepadol.

2.6.3 Texty písané zprava-doľava

Testy na textoch písaných v smere zprava-doľava bolo treba spraviť z toho dôvodu, aby sme ukázali, že nami navrhnuté indexy skutočne dokážu správne určiť smer čítania textu. Hypoteticky by totiž mohla nastať situácia, že sa nám podarí navrhnúť index, ktorý bez ohľadu na smer čítania textu bude v prevažnej väčšine prípadov dávať väčšie hodnoty vždy len v jednom smere.

Naviac, vzhľadom na metodiku akou sa naše testy vykonávajú, je veľmi nepravdepodobné, že ak len prevrátime smer skúmaného textu, tak dostaneme rovnaké hodnoty indexov v oboch smeroch, ale v prehodenom poradí. To by platilo len v prípade ak by sme na danom texte vykonali len jediný výpočet indexov v jednom aj druhom smere, bez toho aby sa daný text menil. My však na danom texte robíme 100 výpočtov indexov v každom smere a pre každý tento výpočet daný text najskôr náhodne rozdelíme na riadky podľa kľúča $60 \pm \delta$. Toto delenie na riadky sa samozrejme robí podľa predpokladaného smeru čítania. Preto, hoci sme testy pre smer čítania zprava-doľava, robili na rovnakých textoch, len s revertovanými riadkami, ako pre smer zľava-doprava, nie sú výsledky testov v jednom aj druhom smere, jednotlivých indexov, identické. Čo však je pre nás dôležité je fakt, že rovnako ako v smere zľava-doprava, nami navrhnuté testy v prevažnej väčšine prípadov správne určili smer čítania textu. Pri testovaní vzoriek anglických, latinských, nemeckých a slovenských textov došlo len v jedinom prípade[9] ku zlyhaniu testu a na tomto texte došlo ku zlyhaniu testu aj v smere zľava-doprava.

Na strane 88 je tabuľka 2.24 zobrazujúca výsledky testov pre jazyky písané zprava-doľava. Z každeho uvedeného jazyka bolo použitých päť vzoriek a všetky vzorky textov mali veľkosť približne 10 kB.

Ako vidno z tabuľky 2.24, výsledky testov zodpovedajú testom v smere zľava-doprava. Jedine v slovenských a hebrejských vzorkách textov sa vyskytli texty, kde zlyhala väčšina testovaných indexov. Hebrejské texty, vzhľa-

[9] 3. kapitola románu *Demokrati*

dom na úplnú neznalosť hebrejčiny, nie sme schopní posúdiť a preto sú aj uvedené oddelene len pre ilustráciu. Okrem toho, ako bolo spomenuté v časti o výbere textov, v prípade vzoriek hebrejských textov sa jedná o transkribciou veľmi skreslené vzorky, nevhodné pre naše účely.

V prípade slovenských textov, zlyhala tretia kapitola románu *Demokrati*, ktorá rovnako zlyhala aj pri testovaní v normálnom smere čítania a ktorej text má, z nášho pohľadu, nevhodnú štruktúru, čo už tiež bolo opakovane spomenuté skôr.

Tabuľky úspešnosti testovaných indexov pre jednotlivé jazyky vyzerajú nasledovne:

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	1	4	20 %	5	3	2	60 %
2	5	0	100 %	5V	4	1	80 %
3	5	0	100 %	6	5	0	100 %
3V	5	0	100 %	6V	5	0	100 %
4	5	0	100 %	7	5	0	100 %
				7V	4	1	80 %

Tabuľka 2.20: Úspešnosť testov – Angličtina reverzne

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	1	4	20 %	5	2	3	40 %
2	3	2	60 %	5V	4	1	80 %
3	5	0	100 %	6	3	2	60 %
3V	4	1	80 %	6V	4	1	80 %
4	4	1	80 %	7	5	0	100 %
				7V	5	0	100 %

Tabuľka 2.21: Úspešnosť testov – Latinčina reverzne

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	0	5	0 %	5	0	5	0 %
2	5	0	100 %	5V	5	0	100 %
3	5	0	100 %	6	4	1	80 %
3V	5	0	100 %	6V	5	0	100 %
4	5	0	100 %	7	5	0	100 %
				7V	5	0	100 %

Tabuľka 2.22: Úspešnosť testov – Nemčina reverzne

Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	2	3	40 %	5	2	3	40 %
2	4	1	80 %	5V	3	2	60 %
3	4	1	80 %	6	2	3	40 %
3V	3	2	60 %	6V	4	1	80 %
4	3	2	60 %	7	3	2	60 %
				7V	3	2	60 %

Tabuľka 2.23: Úspešnosť testov – Slovenčina reverzne

Zaujímavé sú predovšetkým výsledky testov pre nemčinu, ktoré majú v oboch smeroch čítania takmer identické výsledky. Testy indexov I_1 a I_5 mali 0 % úspešnosť a testy zvyšných indexov mali 100 % úspešnosť, s výnimkou testov v smere zprava-doľava, kde v piatej vzorke zlyhal naviac aj test pre index I_6.

Spoločná tabuľka úspešnosti indexov vyzerá takto:

Angličtina, Latinčina, Nemčina, Slovenčina – reverzne							
Test	Správne	Nesprávne	Úspešnosť	Test	Správne	Nesprávne	Úspešnosť
1	4	16	20 %	5	7	13	35 %
2	17	3	85 %	5V	16	4	80 %
3	19	1	95 %	6	13	7	65 %
3V	17	3	85 %	6V	18	2	90 %
4	17	3	85 %	7	18	2	90 %
				7V	18	2	90 %

Aj táto spoločná úspešnosť pre všetky testované jazyky spolu, zodpovedá spoločnej úspešnosti testov pre smer zľava-doprava. Drobné odchýľky sa dajú vysvetliť metódou testovania.

Test / Jazyk	1	2	3	3V	4	5	5V	6	6V	7	7V
hebrejčina 1	Z	O	O	O	O	Z	O	Z	O	Z	Z
hebrejčina 2	O	Z	Z	Z	Z	Z	Z	Z	Z	Z	O
hebrejčina 3	Z	O	O	O	O	Z	O	Z	O	Z	O
hebrejčina 4	Z	O	O	O	O	Z	Z	Z	O	Z	O
hebrejčina 5	Z	O	O	O	O	Z	O	Z	O	O	O
angličtina 1	Z	O	O	O	O	Z	O	Z	O	O	O
angličtina 2	Z	O	O	O	O	O	O	O	O	O	O
angličtina 3	Z	O	O	O	O	O	O	O	O	O	O
angličtina 4	Z	O	O	O	O	Z	O	O	O	O	O
angličtina 5	O	O	O	O	O	O	Z	O	O	O	O
latinčina 1	Z	Z	O	O	O	Z	O	Z	O	O	O
latinčina 2	O	Z	O	Z	Z	O	Z	O	Z	O	O
latinčina 3	Z	O	O	O	O	Z	O	O	O	O	O
latinčina 4	Z	O	O	O	O	Z	O	Z	O	O	O
latinčina 5	Z	O	O	O	O	O	O	O	O	O	O
nemčina 1	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 2	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 3	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 4	Z	O	O	O	O	Z	O	O	O	O	O
nemčina 5	Z	O	O	O	O	Z	O	Z	O	O	O
slovenčina 1	Z	O	O	Z	O	Z	O	Z	O	O	O
slovenčina 3	O	Z	Z	Z	Z	O	Z	Z	Z	Z	Z
slovenčina 5	O	O	O	O	Z	Z	Z	O	O	O	Z
slovenčina 9	Z	O	O	O	O	Z	O	Z	O	Z	O
slovenčina 10	Z	O	O	O	O	O	O	O	O	O	O

Tabuľka 2.24: Zhrnutie testov na textoch písaných zprava-doľava

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

2.7 Záver

Testami nami navrhnutých indexov v oboch smeroch čítania sme ukázali, že principiálne je možné určiť smer čítania textu aj pri neznámom jazyku. My sme skúmali 11 indexov, pričom podobným spôsobom aký sme použili, je možné navrhnúť celú radu ďalších indexov. Teraz ide o to, ako na základe výsledkov testov navrhnutých jedenástich indexov určiť smer čítania tak, aby úspešnosť bola čo najvyššia. Jedna cesta je vyhodnocovať všetkých 11 indexov a na základe toho stanoviť pravdepodobnosť čítania textu jedným a druhým smerom. Druhá cesta je výber podmnožiny indexov, na základe ktorej potom stanovíme smer čítania textu.

Na základe tabuliek úspešnosti testov angličtiny, latinčiny, nemčiny a slovenčiny v oboch smeroch čítania (tabuľky 2.2, 2.16, 2.17, 2.18 a 2.20, 2.21, 2.22, 2.23) určíme poradie úspešnosti jednotlivých indexov. Následne z celej sady indexov vyberieme päť, pomocou ktorých budeme určovať smer čítania textu.

2.7.1 Určovanie poradia úspešnosti

Určiť poradie úspešnosti testovaných indexov sa dá viacerými spôsobmi. My budeme poradie úspešnosti určovať dvoma spôsobmi a to:

1. na základe poradia v tabuľke úspešnosti a
2. na základe percentuálnej úspešnosti.

Poradia a percentuálne úspešnosti jednotlivých testovaných indexov z tabuliek 2.2, 2.16, 2.17, 2.18 a 2.20, 2.21, 2.22, 2.23 sú zhrnuté v nasledovnej tabuľke:

Tabuľka	Poradie a percentuálna úspešnosť (%)										
2.2, str. 69	4 76.5	1 70.6	2 70.6	3 70.6	3V 64.7	5V 64.7	6 64.7	6V 64.7	7 64.7	7V 64.7	5 47.1
2.16, str. 83	3 100	4 100	7 100	7V 100	1 80	2 80	3V 80	5V 80	6 80	6V 80	5 60
2.17, str. 83	2 100	3 100	4 100	5V 100	6V 100	7 100	7V 100	3V 80	6 80	5 60	1 20
2.18, str. 83	2 100	3 100	3V 100	4 100	5V 100	6 100	6V 100	7 100	7V 100	1 0	5 0
2.20, str. 86	2 100	3 100	3V 100	4 100	6 100	6V 100	7 100	5V 80	7V 80	5 60	1 20
2.21, str. 86	3 100	7 100	7V 100	3V 80	4 80	5V 80	6V 80	2 60	6 60	5 40	1 20
2.22, str. 86	2 100	3 100	3V 100	4 100	5V 100	6V 100	7 100	7V 100	6 80	1 0	5 0
2.23, str. 87	2 80	3 80	6V 80	3V 60	4 60	5V 60	7 60	7V 60	1 40	5 40	6 40

2.7.2 Výber sady indexov podľa poradia

Ak budeme vychádzať len z poradia indexov v tabuľkách úspešnosti, tak indexy s rovnakou percentuálnou úspešnosťou budú mať rovnaké poradie. Napríklad v tabuľke 2.17 na strane 83 máme 7 indexov na prvom mieste, 3 indexy na druhom mieste atď.
Poradia jednotlivých indexov v tabuľkách 2.2, 2.16, 2.17, 2.18 a 2.20, 2.21, 2.22, 2.23 sú tieto:

Index	1	2	3	3V	4	5	5V	6	6V	7	7V
Umiestnenia	3x2. 2x3. 2x4. 1x5.	5x1. 2x2. 1x3.	7x1. 1x2.	3x1. 4x2. 1x3.	6x1. 2x2.	1x2. 5x3. 2x4.	3x1. 4x2. 1x3.	2x1. 3x2. 3x3.	5x1. 2x2. 1x3.	6x1. 1x2. 1x3.	5x1. 2x2. 1x3.
Poradie	8.	4.	1.	5.	2.	7.	5.	6.	4.	3.	4.

Ak by sme si teda chceli len na základe poradia vybrať 5 najúspešnejších indexov, tak dostávame sady indexov:

$$\text{Výber}_1 = \{I_3, I_4, I_7, I_2, I_{6V}\}$$

alebo

$$\text{Výber}_2 = \{I_3, I_4, I_7, I_2, I_{7V}\}$$

alebo

$$\text{Výber}_{2'} = \{I_3, I_4, I_7, I_{6V}, I_{7V}\}$$

Je intuitívne zrejmé, že výber najúspesnejších indexov len na základe ich poradia v tabuľkách úspešnosti nebude najobjektívnejší spôsob voľby. V tabuľkách máme indexy, ktoré sú na 1. mieste so 100 % úspešnosťou a rovnako indexy na 1. mieste so 76.5 % úspešnosťou alebo indexy na 2. mieste so 80 % úspešnosťou a na 2. mieste s 60 % úspešnosťou. Ďalej máme v tabuľkách indexy na poslednom mieste so 60 % úspešnosťou a indexy na poslednom mieste s 0 % úspešnosťou. Okrem toho máme tabuľku, v ktorej posledné miesto je zároveň druhé a máme tabuľku, v ktorej posledné miesto je piate. Takže výber len na základe poradia sa zdá byť principiálne neobjektívny a zrejme by bolo vhodné zohľadňovať aj percentuálnu úspešnosť indexov, čo spravíme v nasledujúcej časti.

2.7.3 Výber sady indexov podľa percentuálnej úspešnosti

Pri výbere na základe percentuálnej úspešnosti vypočítame priemernú úspešnosť každého indexu a na základe toho zostavíme ich poradie.

Priemerná percentuálna úspešnosť (%)										
1	2	3	3V	4	5	5V	6	6V	7	7V
31.33	86.33	93.83	83.09	89.56	38.39	83.09	75.59	88.09	90.59	88.09

Poradie úspešnosti indexov										
1.	2.	3.	4.	4.	5.	6.	6.	7.	8.	9.
3	7	4	6V	7V	2	3V	5V	6	5	1

Voľbou 5 najúspešnejších indexov na základe ich percentuálnej úspešnosti dostávame sadu indexov:

$$\text{Výber}_3 = \{I_3, I_7, I_4, I_{6V}, I_{7V}\}$$

Je to rovnaká sada indexov ako $\text{Výber}_{2'}$, takže tento už v ďalšom nebudeme viac používať.

2.7.4 Záverečné testy

Záverečné testy budeme robiť na vzorkách štrnástich európskych jazykov. Sú to texty, ktoré sa v pôvodných testovaných sadách nevyskytli:

1. Angličtina – Arthur Conan Doyle: *The Valley of Fear*
   `http://www.gutenberg.org/ebooks/3289`
2. Čeština – Karel Čapek: *Továrna na absolutno*
   `http://www.pdfknihy.maxzone.eu/books/capek_karel/tovarna_na_absolutno.pdf`
3. Dánčina – Knud Hjortø: *Hans Råskov*
   `http://www.gutenberg.org/ebooks/41956`
4. Francúzština – René Bazin: *Madame Corentine*
   `http://www.gutenberg.org/ebooks/43787`
5. Holandčina – Lode Baekelmans: *Mijnheer Snepvangers*
   `http://www.gutenberg.org/ebooks/15048`
6. Latinčina – Burton Egbert Stevenson: *Mysterium Arcae Boulé*
   `http://www.gutenberg.org/ebooks/46456`
7. Maďarčina – Miklós Surányi: *A mester*
   `http://www.gutenberg.org/ebooks/19744`
8. Nemčina – Leonhard Frank: *Der Mensch ist gut*
   `http://www.gutenberg.org/ebooks/35176`

9. Polština – Ludwik Bruner: *Jeden miesiac życia*
   `http://www.gutenberg.org/ebooks/28014`

10. Portugalčina – José Agostinho: *Jaime de Magalhães Lima*
    `http://www.gutenberg.org/ebooks/27689`

11. Slovenčina – Ján Igor Hamaliar: *Kritické torzo*
    `http://zlatyfond.sme.sk/dielo/1765/Hamaliar_Kriticke-torzo/1`

12. Španielčina – Emilio Aguinaldo: *Reseña Veridica de la Revolución Filipina*
    `http://www.gutenberg.org/ebooks/14307`

13. Švédčina – Verner von Heidenstam: *Folkungaträdet*
    `http://www.gutenberg.org/ebooks/13371`

14. Taliančina – Sibilla Aleramo: *Il passaggio*
    `http://www.gutenberg.org/ebooks/49626`

Jazyky boli vyberané spomedzi európskych jazykov tak, aby sa dala ľahko spraviť ich transkribcia do TSA, vynechaním diakritiky. Zároveň boli vyberané také jazyky, pri ktorých sme buď rozumeli vybranému textu, alebo sme aspoň boli schopní posúdiť jeho charakter. Z každého uvedeného jazyka a diela bol vybraný kompaktný úsek textu, väčšinou hneď zo začiatku diela, o veľkosti približne 10 kB.

Na takto vybraných štrnástich vzorkách boli spravené testy smeru čítania. Výsledok testov sa určoval po prvé tak, že sa na základe všetkých jedenástich indexov určila pravdepodobnosť jedného alebo druhého smeru čítania textu. Robilo sa to tak, že všetkých 11 indexov bolo braných ako 100 % a ich percentuálne úspešnosti, uvedené v tabuľke v časti 2.7.3, určili váhy jednotlivých indexov. Za správny smer čítania textu sa potom zobral ten, ktorý mal pravdepodobnosť väčšiu. Okrem toho sa testovali vybrané sady piatich indexov:

$$\begin{aligned} \text{Výber}_1 &= \{I_3, I_4, I_7, I_2, I_{6V}\} \\ \text{Výber}_2 &= \{I_3, I_4, I_7, I_2, I_{7V}\} \\ \text{Výber}_3 &= \{I_3, I_7, I_4, I_{6V}, I_{7V}\} \end{aligned}$$

a za správny smer čítania textu sa považoval ten smer, ktorý väčšina testovaných indexov označila za správny.

Výsledné testy všetkých indexov sú v tabuľke 2.25 na strane 93. V poslednom stĺpci tejto tabuľky je uvedená pravdepodobnosť, že testovaná vzorka textu sa píše v smere zľava-doprava, čo je v našom prípade ten správny smer. Pri všetkých textoch dopadol test úspešne a pravdepodobnosti sa pohybovali

Test / Jazyk	1	2	3	3V	4	5	5V	6	6V	7	7V	Úspešnosť L-P
Angličtina	O	O	O	O	O	Z	O	O	O	O	O	95.47 %
Čeština	O	Z	Z	Z	Z	O	O	O	O	O	O	58.39 %
Dánčina	Z	O	O	O	O	Z	Z	Z	O	O	O	73.07 %
Francúzština	Z	O	O	O	O	Z	O	Z	O	O	O	82.86 %
Holandčina	Z	O	O	O	O	Z	O	O	O	Z	O	91.78 %
Latinčina	O	O	O	O	O	O	O	O	O	O	O	100.0 %
Maďarčina	Z	O	O	O	O	Z	O	Z	O	O	O	82.86 %
Nemčina	Z	O	O	O	O	Z	O	Z	O	O	O	82.86 %
Polština	Z	O	O	O	O	Z	O	O	O	O	O	91.78 %
Portugalčina	O	O	O	O	O	Z	O	Z	O	O	O	86.56 %
Slovenčina	O	O	O	Z	O	Z	Z	O	O	O	Z	65.49 %
Španielčina	O	Z	Z	Z	Z	O	O	O	O	O	O	58.39 %
Švédčina	O	Z	O	O	O	O	O	O	O	Z	O	79.43 %
Taliančina	Z	O	O	O	O	O	O	O	O	O	O	96.31 %

Tabuľka 2.25: Finálne testy na štrnástich jazykoch

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

od 58.39 % po 100 %. Najhoršie pritom dopadli vzorky textov češtiny, španielčiny a slovenčiny a najlepšie dopadli vzorky textov latinčiny, taliančiny a angličtiny.

Výsledky týchto testov výberu piatich indexov, či už podľa poradia, alebo podľa percentuálnej úspešnosti, sú v tabuľkách 2.26, 2.27 a 2.28. Najúspešnejším sa ukázal výber piatich indexov podľa ich percentuálnej úspešnosti[10]. Pri tomto výbere bolo všetkých 14 vzoriek vyhodnotených správne. Pri prvých dvoch výberoch indexov len na základe ich poradia (Výber_1 a Výber_2) bolo správne vyhodnotených 12 textov zo 14, čo je približne 85.7 %.

2.8 Zhrnutie

Realizované testy nasvedčujú tomu, že hypotéza 1 (str. 52) je platná. V prevažnej väčšine prípadov testy správne dokázali určiť smer čítania textu, pri-

[10] Čo zhodou okolností bol aj jeden z výberov indexov podľa ich poradia.

Test / Jazyk	Počet znakov	3	4	7	2	6V
Angličtina	10245	O	O	O	O	O
Čeština	10241	Z	Z	O	Z	O
Dánčina	10240	O	O	O	O	O
Francúzština	10242	O	O	O	O	O
Holandčina	10239	O	O	Z	O	O
Latinčina	10241	O	O	O	O	O
Maďarčina	10239	O	O	O	O	O
Nemčina	10240	O	O	O	O	O
Polština	10242	O	O	O	O	O
Portugalčina	10242	O	O	O	O	O
Slovenčina	10243	O	O	O	O	O
Španielčina	10240	Z	Z	O	Z	O
Švédština	10240	O	O	Z	Z	O
Taliančina	10240	O	O	O	O	O

Tabuľka 2.26: Finálne testy – výber indexov Výber_1

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

čom sa skúmali len zlomy riadkov a nevyužívali sa žiadne ďalšie vedomosti o texte alebo použitom jazyku.

Z vyskúšaných metód sa ako najvhodnejší javí spôsob, keď sa smer čítania textu určil na základe vyhodnotenia percentuálnej úspešnosti všetkých 11 indexov (tabuľka 2.25). V prípade použitia tejto metódy, vieme povedať s akou percentuálnou úspešnosťou sa skúmaný text číta v jednom alebo druhom smere. Za správny smer sa považuje ten, ktorého percentuálna úspešnosť je vyššia. Podstatným faktom je to, že pri tejto metóde máme okrem informácie o smere čítania textu aj informáciu o pravdepodobnosti správnosti tohto údaja. Výsledky tejto metódy by sa ešte dali zlepšiť vypustením najmenej úspešných indexov (napr. I_1 a I_5) z testovacej sady.

V prípade redukcie testovaných indexov na päť najúspešnejších, sa z vyskúšaných troch možností javí najlepšie Výber_3, ktorý zodpovedá výberu piatich najúspešnejších indexov, na základe ich percentuálnej úspešnosti (tabuľka 2.28). Pri tomto spôsobe sa na skúmanom texte určil smer čítania

Test / Jazyk	Počet znakov	3	4	7	2	7V
Angličtina	10245	O	O	O	O	O
Čeština	10241	Z	Z	O	Z	O
Dánčina	10240	O	O	O	O	O
Francúzština	10242	O	O	O	O	O
Holandčina	10239	O	O	Z	O	O
Latinčina	10241	O	O	O	O	O
Maďarčina	10239	O	O	O	O	O
Nemčina	10240	O	O	O	O	O
Polština	10242	O	O	O	O	O
Portugalčina	10242	O	O	O	O	O
Slovenčina	10243	O	O	O	O	Z
Španielčina	10240	Z	Z	O	Z	O
Švédština	10240	O	O	Z	Z	O
Taliančina	10240	O	O	O	O	O

Tabuľka 2.27: Finálne testy – výber indexov Výber_2

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

podľa vybraných piatich indexov na základe väčšiny.

Žiadna z uvedených metód nie je a principiálne ani nemôže byť vo všeobecnosti[11] 100 % spoľahlivá. Vždy sa dá nájsť, prípadne aj umelo skonštruovať zmysluplný text, na ktorom testy všetkých nami navrhnutých indexov zlyhajú. Avšak uvedenými metódami sa dá pre väčšinu bežných textov s vysokou pravdepodobnosťou správne určiť smer čítania textu. Úspešnosť závisí na použitom jazyku a charaktere textu (próza, poézia, úradný dokument,...), avšak vo všetkých skúmaných jazykoch a textoch sa pohybovala od zhruba 60 % nahor a vo väčšine prípadov okolo 80–90 %.

[11] Týmto sa myslí ľubovoľný zmysluplný text, v ľubovoľnom jazyku.

Test / Jazyk	Počet znakov	3	4	7	6V	7V
Angličtina	10245	O	O	O	O	O
Čeština	10241	Z	Z	O	O	O
Dánčina	10240	O	O	O	O	O
Francúzština	10242	O	O	O	O	O
Holandčina	10239	O	O	Z	O	O
Latinčina	10241	O	O	O	O	O
Maďarčina	10239	O	O	O	O	O
Nemčina	10240	O	O	O	O	O
Polština	10242	O	O	O	O	O
Portugalčina	10242	O	O	O	O	O
Slovenčina	10243	O	O	O	O	Z
Španielčina	10240	Z	Z	O	O	O
Švédština	10240	O	O	Z	O	O
Taliančina	10240	O	O	O	O	O

Tabuľka 2.28: Finálne testy – výber indexov Výber_3

Vysvetlivky:

- „O“ znamená, že test určil smer textu správne.
- „Z“ znamená, že test určil smer čítania textu nesprávne.

2.9 Vplyv šifrovania na popísané metódy

Nami popisované metódy určovania smeru čítania textu boli navrhované tak, aby sa dali použiť aj na zašifrované texty v neznámom jazyku. Pritom predpokladáme, že použitá šifra bude buď jednoduchá zámena, prípadne jej rozšírenie typu n-gramovej substitúcie alebo kódu.

Prípad jednoduchej zámeny je triviálny. Tam sa znak substituuje za znak a to na nami používané metódy nemá vôbec žiaden vplyv. Takže voči jednoduchej zámene sú popisované metódy úplne imúnne.

Trochu komplikovanejšie sú prípady n-gramových substitúcií, kde sa n-gram otvoreného textu substituuje znakom zašifrovaného textu, prípadne kódy, kde sa celé časti otvoreného textu substituujú znakom kódu. V týchto prípadoch môže dôjsť k redukcii opakujúcich sa sekvencií znakov v texte. Stále však platí, že rovnakým častiam otvoreného textu, budú zodpovedať rovnaké časti zašifrovaného textu. Takže nami popisované metódy budú prin-

cipiálne fungovať, avšak v porovnaní s otvoreným textom budeme na analýzu potrebovať väčšiu vzorku, resp. pri rovnakej vzorke (rovnakom počte znakov) bude pravdepodobnosť správneho určenia smeru čítania nižšia.

V historických šifrovaných textoch, až do obdobia neskorej renesancie, sa iné šifry ako uvedené substitúcie takmer vôbec nepoužívali. Občas sa od začiatku renesancie vyskytovali homofónne substitúcie a výnimočne aj jednoduché formy transpozícií, ale väčšina historických kryptodokumentov používala niektoré z nami uvedených substitúcii.

Na začiatku textu sme spomínali ako príklady Voynichov manuskript a Rohonczi kódex. Nie je zatiaľ známe, či sa jedná o zmysluplné texty, alebo hoax, ale ak sa jedná o zmysluplné texty, tak podľa doby ich predpokladaného vzniku sa dá predpokladať, že použitá šifra bude práve niektorá z uvedených substitúcií. Napríklad v Rohonczi kódexe sa počet rôznych znakov pohybuje okolo 600, čo by mohlo nasvedčovať o bigramovej substitúcii. V takom prípade by sa nami popísané metódy dali použiť na určenie smeru čítania textu, ktorý je v Rohonczi kódexe zatiaľ neznámy a iba na základe rôznych indícii sa predpokladá, že asi je zprava-doľava.

2.10 Postup pri testovaní zašifrovaného textu

V prípade zašifrovaného textu v neznámom jazyku bude testovanie na smer čítania prebiehať podobne ako testy, ktoré sme robili na otvorenom texte. Na začiatku nevieme, ktorý smer čítania je správny. Pri testovaní nami navrhnutých indexov sme ukázali, že jeden test stačiť nebude, pretože pri nevhodnom charaktere textu, prípadne nevhodnom rozdelení riadkov môže hociktorý z použitých indexov zlyhať. Najmä ak riadky končia vždy celým slovom, čo sa u ručne písaných historických dokumentov dá predpokladať, je pravdepodobnosť zlyhania veľmi vysoká.

Takže najskôr budeme predpokladať smer čítania textu zľava-doprava, celú testovanú vzorku textu spojíme v predpokladanom smere čítania do jedného reťazca a potom opakovane budeme rozdelovať tento reťazec na riadky náhodnej dĺžky a na takto získanom texte budeme vykonávať testy. Za výsledné hodnoty indexov potom zoberieme priemer ich hodnôt z jednotlivých opakovaní.

Potom budeme predpokladať smer čítania textu zprava-doľava, opäť celú testovanú vzorku textu spojíme v predpokladanom smere čítania do jedného reťazca a ďalšie testovanie bude prebiehať rovnako ako v predošlom prípade.

Nakoniec výsledky testov v jednom a druhom smere čítania porovnáme a niektorou z nami popísaných metód určíme správny smer čítania textu, prípadne aj jeho pravdepodobnosť.

2.11 Softwarová realizácia

Softwarová realizácia zisťovania popisovaných indexov pozostávala z dvoch častí. Hlavnou časťou bol program v `C++`, ktorý na danom, pevne nariadkovanom texte, preskúmal zlomy riadkov a riadky, našiel všetky legitímne reťazce a vypočítal indexy tak, ako boli definované v texte. Tento program má len CLI interface a pomocou prepínača sa určuje smer čítania textu, v ktorom počíta indexy. Druhou, pomocnou časťou, bol skript v Pythone, ktorý daný text náhodne rozriadkoval v jednom alebo druhom smere a pre každé takéto rozriadkovanie na ňom spustil program na počítanie indexov. Toto zopakoval predpísaný počet krát (100) a následne vypočítal a vypísal priemerne dosiahnuté hodnoty. Pamäťová aj časová náročnosť oboch programov je zanedbateľná. Pre 10 kB text a 100 opakovaní testov, trvali tieto testy zhruba 2-3 sekundy.

Kapitola 3

Sovietska šifra VIC a jej lúštenie

3.1 Úvod a známe publikácie o šifre VIC

VIC bola sovietska šifra z čias studenej vojny, ktorú používal Reino Häyhänen, alias Eugene Maki, alias Victor alebo Vic. Häyhänen pôsobil ako agent v USA v rokoch 1952 až 1957. V prípade šifry VIC sa jednalo o šifru typu STT, pričom S bola jedno a dvojmiestna zámena[1] a dve T predstavujú dve transpozície. Prvá z nich bola bežná tabuľková transpozícia a druhá bola špeciálna obrazcová transpozícia podobná československej šifre Zubatka z čias 2. svetovej vojny.
V pondelok 22. júna 1953 našiel 14-ročný chlapec Jimmy Bozart, predávajúci noviny na Foster Avenue v Brooklyne, preparovanú 5-centovú mincu (*nickel*), v ktorej bol schovaný mikrofilm so zašifrovanou správou. Mincu odovzdal polícii a tá ju posunula ďalej FBI. Táto sa neúspešne pokúšala správu rozlúštiť. To sa im nepodarilo až do mája 1957, keď Häyhänen, potom čo prebehol na americkú stranu, prezradil postup šifrovania a dešifrovania a použité heslá.

Potom čo bola história okolo agenta Häyhänena a šifry VIC v roku 1993 odtajnená, sa šifrou VIC zaoberalo viacero publikácií. Takmer všetky tieto publikácie sa šifrou VIC zaoberali len okrajovo a venovali sa predovšetkým príbehu Reina Häyhänena. Len v jedinom článku je podrobne popísaný postup šifrovania a lúštením sa nezaoberala žiadna nám známa publikácia.

Príbehom agenta Häyhänena sa podrobne zaoberali články [24], [37] a [28]. V Kahnovom článku [23] je podrobne popísaný postup šifrovania a je tam aj publikovaná Häyhänenova depeša, ktorú FBI našlo v preparovanej minci. Práve na tejto originálnej depeši Kahn vysvetľuje postup šifrovania.

[1] V anglickej terminológii *checkerboard cipher*.

Okrem toho sú príbeh agenta Häyhänena a šifra VIC okrajovo spomenuté aj v známej Kahnovej knihe *The Codebreakers* [22].

Na FEI STU v Bratislave sa šifrou VIC zaoberali dvaja študenti vo svojich bakalárskych prácach. Zoltán Mierka vo svojej práci *Šifra VIC a jej softvérová realizácia* [34] naprogramoval šifru VIC. Žiaľ autor tejto práce nemal k dispozícii článok [23] a vychádzal len z knihy [22] a webových zdrojov, kde popis nie je presný a kompletný. Oproti Häyhänenovej šifre VIC sú rozdiely v tom, že azbuka je zamenená za latinku, z čoho vyplýva zmena substitučnej tabuľky a okrem toho sa znaky do substitučnej tabuľky zapisovali iným (bezpečnejším) spôsobom.
V druhej a zaujímavejšej práci *Kryptoanalýza šifry VIC* [4], jej autor, Radoslav Čagala analyzoval šifru VIC a pokúsil sa o jej lúštenie. V tejto práci opäť autor nemal k dispozícii článok [23] a popis šifry VIC prebral z predošlej práce Zoltána Mierku [34]. Autor našiel najslabšie miesto šifry VIC a ukázal útok s využitím tejto slabiny, pričom ho aj softwarovo realizoval. Avšak, aj v dôsledku nepresnosti a nekompletnosti popisu, už nevyužil ďalšie slabiny šifry VIC, a preto bol tento útok časovo pomerne náročný (ale v praxi stále ešte realizovateľný).

V ďašej časti podrobne popíšeme šifru VIC, aby sme ukázali ako komplexná a komplikovaná bola táto šifra. Vo viacerých publikáciach sa VIC označuje za „najkomplexnejšiu a najsilnejšiu“ klasickú ručnú šifru. Celý, tu uvedený, popis je kompletne prevzatý z Kahnovho článku [23]. Uvedený článok je jediný, nám známy, podrobný a kompletný popis šifry VIC. Takisto všetky uvádzané texty a použité heslá sú prevzaté z uvedeného článku a jedná sa o autentické texty a heslá tak, ako boli na mikrofilme v minci. Na webe existujú mnohé ďalšie popisy šifry VIC, ale tie sú zväčša nekompletné, prípadne obsahujú chyby. A aby situácia okolo šifry VIC bola ešte zmätenejšia, existuje iná šifra VIC[2], ktorú používala východonemecká StaSi a ktorá, okrem mena, nemá s Häyhänenovou šifrou VIC nič spoločné.

3.2 Postup šifrovania

Pri šifrovaní správy pomocou šifry VIC používal agent heslo, ktoré pozostávalo z piatich častí. Prvé štyri časti boli pevne dané a zrejme[3] boli pridelované agentovi na dlhšiu dobu. Posledná piata časť bolo 5-ciferné číslo, ktoré sa volilo náhodne a malo byť pre každú depešu iné. V Häyhänenovom prípade boli pevné časti hesla nasledovné:

[2]Podľa pána Drobicka skúmajúceho archívy StaSi.
[3]Vyplýva to z popisu v článku [23].

1. Dátum sovietskeho víťazstva nad Japonskom: **3. september 1945**
2. Slovné heslo: **СНЕГОПАД**[4]
3. Slovná fráza: **Только слышно на улице одинокая бродит гармонь**[5]
4. Osobné číslo agenta: **13**

V prípade odchytenej depeše z mince bola piata časť hesla – inicializačný reťazec pozostávajúci z piatich cifier: **20818**.

Z uvedených častí hesla sa potom veľmi komplexným spôsobom vytvárali súradnice pre substitučnú tabuľku, permutácia stĺpcov prvej transpozičnej tabuľky a permutácia stĺpcov druhej transpozičnej tabuľky. Súradnice pre substitučnú tabuľku vždy pozostávali z nejakej permutácie cifier 0 až 9. Dĺžky transpozičných permutácií boli premenlivé a záviseli od toho ako sa zvolila piata, premenlivá časť hesla. Vytváranie substitučnej aj transpozičných permutácii si podrobne popíšeme v samostatnej časti, vzhľadom na jeho zložitosť. Na tomto mieste uvádzame len finálne permutácie použité v Häyhänenovej depeši z mince:

- Súradnice pre substitučnú tabuľku: `5 0 7 3 8 9 4 6 1 2`
- Permutácie stĺpcov prvej transpozičnej tabuľky:
  `14 8 16 2 3 1 13 4 9 10 5 11 15 17 6 12 7`
- Permutácie stĺpcov druhej transpozičnej tabuľky:
  `5 13 2 9 7 6 14 8 3 12 10 11 1 4`

Otvorený text depeše z mince bol nasledovný[6]:

1. Поздравляем с благополучным прибытием. Потверждаем получение вашего письма в адрес “В” и прочтение письма №1.
2. Для организации прикрытия мы дали указание передать вам три тысячи местных. Перед тем как их вложить в какое лубо дело посоветуйтесь с нами, сообщив характеристику этого дела.
3. По вашей просьбе рецептуру изготовления мягкой пленки и новостей передадим отдельно в месте с письмом матери.

[4]Ruské slovo **Снегопад** znamená *sneženie*.
[5]Jedná sa o prvú slohu (vtedy) v Rusku populárnej piesne *Osamelá harmonika*.
[6]Preklad tohto textu je v prílohe na strane 130.

4. Гаммы высылать вам рано. Короткие письма шифруйте, а побольше – [7] делайте со встабками. Все данные о себе, место работы, адрес и т.д. в одной шифровке передать нельзя. Вставки передавайте отдельно.

5. Посылку жене передали лично. С семьей все благополучно. Желаем успеха. Привет от товарищей.

№1/[8]03 Декабря

Teraz si ukážeme postup ako sa tento text šifroval. Popis bude rozdelený na jednotlivé časti, t.j. najskôr substitúcia, a potom prvá a následne druhá transpozícia.

3.2.1 Substitúcia

Použitá substitúcia sa v slovenskej aj českej terminológii nazýva *jedno a dvojmiestna zámena*. Je to číselná monoalfabetická substitúcia, v ktorej sa znaky otvoreného textu zamieňajú jedno a dvojcifernými číslami. Samotný fakt, že v zašifrovanom texte máme znaky rôznej dĺžky (jedno a dvojciferné) nijako zvlášť nekomplikuje lúštenie tejto šifry. Lúštenie jedno a dvojmiestnej zámeny je zhruba rovnako zložité ako lúštenie jednoduchej zámeny, v ktorej majú všetky znaky zašifrovaného textu rovnakú dĺžku.

Najskôr si musíme zostaviť substitučnú tabuľku. Táto bude mať 10 stĺpcov a 4 riadky. Stĺpce tabuľky budú očíslované substitučnou permutáciou, čiže v prípade depeše z mince je to: `5 0 7 3 8 9 4 6 1 2`. Prvý riadok tabuľky nebude očíslovaný[9] a súradnice druhého až štvrtého riadku tabuľky budú posledné tri cifry substitučnej permutácie, t.j. v tomto prípade: `6 1 2`.

Do prvého riadku tabuľky sa zľava zapísalo prvých 7 znakov slovného hesla. V prípade depeše z mince je to: **СНЕГОПА**. Posledné tri stĺpce prvého riadku zostávajú voľné. Vo zvyšných troch riadkoch sú potom zapísané ostatné písmená ruskej abecedy okrem znakov s diakritikou a tvrdého znaku (t.j. okrem **Ё, Й, Ъ**) a 7 špeciálnych znakov: **. , П/Л № Н/Ц НТ ПВТ**. Ostatné znaky ruskej abecedy sa do posledných 3 riadkov tabuľky zapisovali po stĺpcoch v poradí podľa abecedy tak, že tretí a piaty stĺpec[10] sa vynechávali. Do tretieho a piateho stĺpca tabuľky sa zapisovalo prvých 6 zo 7 špeciálnych znakov v poradí, v akom sú uvedené vyššie. Posledný špeciálny

[7] Na tomto mieste je v depeši slovo **тире**, čo po rusky znamená *pomlčka*.

[8] Namiesto znaku / je v depeši slovo **дробь**, ktoré znamená (aj) *lomítko*.

[9] Nebude mať súradnicu.

[10] Na tomto mieste je v článku [23] zrejme chybná informácia. Ak by bola substitučná tabuľka zostavovaná týmto spôsobom, výrazne by sa tým uľahčilo lúštenie šifry VIC. Ešte sa k tomuto vrátime v záverečných poznámkach na strane 110.

znak sa zapísal na koniec tabuľky. Takže výsledná substitučná tabuľka pre depešu z mince bola:

	5	0	7	3	8	9	4	6	1	2
—	**С**	**Н**	**Е**	**Г**	**О**	**П**	**А**			
6	Б	Ж	.	К	№	Р	Ф	Ч	Ы	Ю
1	В	З	,	Л	Н/Ц	Т	Х	Ш	Ь	Я
2	Д	И	П/Л	М	НТ	У	Ц	Щ	Э	ПВТ

Pokiaľ sa jedná o zmysel siedmych špeciálnych znakov, tak interpunkčné znamienka sú zrejmé. Zmysel znaku **П/Л** nie je známy, ale zrejme išlo o nejaký prepínač módov. V depeši z mince tento znak nie je použitý. Znak **№** je skratka pre číslo (č.). Znak **Н/Ц** je prepínač módov text/cifry. Zrejme je to skratka ruských slov Нормально/Цифра. Znak **НТ** označuje začiatok depeše a zrejme sa jedná o skratku ruských slov Начало Текста. Znak **ПВТ** znamená *opakovať* a je to skratka ruského slova Повторять.

Čísla, resp. cifry sa pri použitej substitúcii kódovali tak, že pomocou znaku **Н/Ц** sa prepol mód na cifry, potom sa príslušná cifra napísala trikrát po sebe a opäť sa znakom **Н/Ц** prepol mód na text. Takže napr. cifru 1 by sme zapísali ako: `18 111 18`, cifru 7 ako `18 777 18` atď. V depeši z mince sa nachádza viacero cifier. V jednom prípade je omylom cifra 0 zapísaná ako písmeno O. V podrobnom prepise depeše to je vyznačené.

Napokon, ešte predtým ako si uvedieme substituovaný text depeše, treba uviesť, že v šifre VIC sa text depeše pred substitúciou náhodne rozdelil na dve časti. Najskôr sa substituovala druhá časť depeše, potom sa uviedol znak **НТ**, označujúci začiatok textu a následne sa substituovala prvá časť depeše. Toto opatrenie malo sťažiť lúštenie tým, že na začiatku substituovanej depeše sa nenachádzali obvyklé frázy. Avšak týmto opatrením sa dosiahol pravý opak a lúštenie depeše sa týmto zjednodušilo.

Po substituovaní textu depeše za čísla sa ešte počet cifier doplnil na najbližší vyšší násobok 5 nulami (náhodnými ciframi). Depeša z mince má na svojom konci tri nuly (2 1 4).

Depeša z mince, po substituovaní znakov za jedno a dvojciferné čísla vyzerala nasledovne:

```
96920636961192012236125413202963410402079769725419111542319692
01961512662023751906114679769725197236346320141513860201911156
34638713206582571389858157192920197511504232017588652620151446
94631976920519206329211983825713467183331867981541672096985116
57697247919296929201038198151370201223123638209137063202008158
51972097697254252023819257131108152375197592051123823234197692
```

```
06718444186734232361156156113419111542369408676386981963207920
51123416206469292019717498658131116719206972571342019758155194
15634232067155725400617857657172375198694658196117425697520196
72567158250820162064698156379769725415419110713111012671551941
56320976972541542019781925713110867185551867985611363296070797
69725413201320660867557231172015576513438981329660867607134723
29597144679692015719819198154692026720681811118256986511818333
18257634656912281811186798102569415131272356513438981329660 61
23969206561192072367982519157696025472398132966702071541673892
05112341542569751717152215171720969866197020792051123468181111
86718222186725131286934020104242020214
```

Počet znakov (aj s nulami na konci) je 1030 a je to zapísané veľmi neprehľadne. Podrobne a prehľadne rozpísaná substitúcia depeše je uvedená v tabuľkách 3.7 a 3.8 na stranách 132, 133.

3.2.2 Prvá transpozícia

Prvá transpozícia bola obyčajná tabuľková transpozícia s obdĺžnikovou tabuľkou. Šírka tabuľky bola daná permutáciou stĺpcov. Substituovaná depeša sa do transpozičnej tabuľky zapisovala po riadkoch zľava doprava, zhora nadol a čítala sa po stĺpcoch zhora nadol. Poradie čítania stĺpcov určovala permutácia. To ako vypadala prvá transpozičná tabuľka pre depešu z mince možno vidieť v tabuľke 3.1 na strane 105. Použitá permutácia stĺpcov tabuľky je zapísaná tučným písmom v druhom riadku zhora. V riadku nad ňou je zapísaná číselná postupnosť, ktorej vyčíslením je táto permutácia. Tvorbu hesla si popíšeme neskôr, v samostatnej časti. Začiatok a koniec depeše po prvej transpozícii potom majú tvar:

```
6 5 7 3 0 9 4 3 3 7.....6 9 6 6 5 1 8 3 2 2
```

Takto zašifrovaná depeša sa ďalej šifrovala pomocou druhej transpozície.

3.2.3 Druhá transpozícia

Druhá transpozícia bola tiež tabuľková transpozícia s obdĺžnikovou tabuľkou, avšak obrazcová. Znamená to, že správa sa do tabuľky zapisovala v určitých obrazcoch. Šírku transpozičnej tabuľky určovala opäť permutácia stĺpcov, t.j. transpozičné heslo. Výstup z prvej transpozície sa do tabuľky druhej transpozície zapisoval po riadkoch zľava doprava a zhora nadol, ale nie po celej šírke riadkov. Druhú transpozičnú tabuľku pre depešu z mince môžme vidieť v tabuľke 3.2 na strane 107. V tejto tabuľke sú riadky rozdelené na biele a

9	6	0	3	3	1	8	3	6	6	4	6	9	0	4	7	5
14	**8**	**16**	**2**	**3**	**1**	**13**	**4**	**9**	**10**	**5**	**11**	**15**	**17**	**6**	**12**	**7**
9	6	9	2	0	6	3	6	9	6	1	1	9	2	0	1	2
2	3	6	1	2	5	4	1	3	2	0	2	9	6	3	4	1
0	4	0	2	0	7	9	7	6	9	7	2	5	4	1	9	1
1	1	5	4	2	3	1	9	6	9	2	0	1	9	6	1	5
1	2	6	6	2	0	2	3	7	5	1	9	0	6	1	1	4
6	7	9	7	6	9	7	2	5	1	9	7	2	3	6	3	4
6	3	2	0	1	4	1	5	1	3	8	6	0	2	0	1	9
1	1	1	5	6	3	4	6	3	8	7	1	3	2	0	6	5
8	2	5	7	1	3	8	9	8	5	8	1	5	7	1	9	2
9	2	0	1	9	7	5	1	1	5	0	4	2	3	2	0	1
7	5	8	8	6	5	2	6	2	0	1	5	1	4	4	6	9
4	6	3	1	9	7	6	9	2	0	5	1	9	2	0	6	3
2	9	2	1	1	9	8	3	8	2	5	7	1	3	4	6	7
1	8	3	3	3	1	8	6	7	9	8	1	5	4	1	6	7
2	0	9	6	9	8	5	1	1	6	5	7	6	9	7	2	4
7	9	1	9	2	9	6	9	2	9	2	0	1	0	3	8	1
9	8	1	5	1	3	7	0	2	0	1	2	2	3	1	2	3
6	3	8	2	0	9	1	3	7	0	6	3	2	0	2	0	0
8	1	5	8	5	1	9	7	2	0	9	7	6	9	7	2	5
4	2	5	2	0	2	3	8	1	9	2	5	7	1	3	1	1
0	8	1	5	2	3	7	5	1	9	7	5	9	2	0	5	1
1	2	3	8	2	3	2	3	4	1	9	7	6	9	2	0	6
7	1	8	4	4	4	1	8	6	7	3	4	2	3	2	3	6
1	1	5	6	1	5	6	1	1	3	4	1	9	1	1	1	5
4	2	3	6	9	4	0	8	6	7	6	3	8	6	9	8	1
9	6	3	2	0	7	9	2	0	5	1	1	2	3	4	1	6
2	0	6	4	6	9	2	9	2	0	1	9	7	1	7	4	9
8	6	5	8	1	3	1	1	1	6	7	1	9	2	0	6	9
7	2	5	7	1	3	4	2	0	1	9	7	5	8	1	5	5
1	9	4	1	5	6	3	4	2	3	2	0	6	7	1	5	5
7	2	5	4	0	0	6	1	7	8	5	7	6	5	7	1	7
2	3	7	5	1	9	8	6	9	4	6	5	8	1	9	6	1
1	7	4	2	5	6	9	7	5	2	0	1	9	6	7	2	5
6	7	1	5	8	2	5	0	8	2	0	1	6	2	0	6	4
6	9	8	1	5	6	3	7	9	7	6	9	7	2	5	4	1
5	4	1	9	1	1	0	7	1	3	1	1	1	0	1	2	6
7	1	5	5	1	9	4	1	5	6	3	2	0	9	7	6	9
7	2	5	4	1	5	4	2	0	1	9	7	8	1	9	2	5
7	1	3	1	1	0	8	6	7	1	8	5	5	5	1	8	6
7	9	8	5	6	1	1	3	6	3	2	9	6	0	7	0	7
9	7	6	9	7	2	5	4	1	3	2	0	1	3	2	0	6
6	0	8	6	7	5	5	7	2	3	1	1	7	2	0	1	5
5	7	6	5	1	3	4	3	8	9	8	1	3	2	9	6	6
0	8	6	7	6	0	7	1	3	4	7	2	3	2	9	5	9
7	1	4	4	6	7	9	6	9	2	0	1	5	7	1	9	8
1	9	1	9	8	1	5	4	6	9	2	0	2	6	7	2	0
6	8	1	8	1	1	1	1	8	2	5	6	9	8	6	5	1
1	8	1	8	3	3	3	1	8	2	5	7	6	3	4	6	5
6	9	1	2	2	8	1	8	1	1	1	1	8	6	7	9	8
1	0	2	5	6	9	4	1	5	1	3	1	2	7	2	3	5
6	5	1	3	4	3	8	9	8	1	3	2	9	6	6	0	6
1	2	3	9	6	9	2	0	6	5	6	1	1	9	2	0	7
2	3	6	7	9	8	2	5	1	9	1	5	7	6	9	6	0
2	5	4	7	2	3	9	8	1	3	2	9	6	6	7	0	2
0	7	1	5	4	1	6	7	3	8	9	2	0	5	1	1	2
3	4	1	5	4	2	5	6	9	7	5	1	7	1	7	1	5
2	2	1	5	1	7	1	7	2	0	9	6	9	8	6	6	1
9	7	0	2	0	7	9	2	0	5	1	1	2	3	4	6	8
1	8	1	1	1	1	8	6	7	1	8	2	2	2	1	8	6
7	2	5	1	3	1	2	8	6	9	3	4	0	2	0	1	0
4	2	4	2	0	2	0	2	1	4							

Tabuľka 3.1: Prvá transpozičná tabuľka pre depešu z mince

šedé časti, pričom šedé časti tvoria trojuholníky. Prvý z týchto trojuholníkov sa začína v prvom riadku v stĺpci s číslom 1[11]. Potom v každom ďalšom riadku posunieme začiatok šedej časti o jeden stĺpec doprava, až kým neprídeme na koniec riadku. Za tým nasleduje jeden biely riadok plnej šírky a potom sa v ďalšom riadku začína druhý šedý trojuholník v stĺpci s číslom 2. Opäť v nasledujúcich riadkoch posúvame začiatok šedej časti o jeden stĺpec doprava až po koniec riadku, potom spravíme jeden biely riadok plnej šírky atď. Počet riadkov tabuľky je daný šírkou tabuľky a dĺžkou šifrovanej správy.

Výstup z prvej transpozičnej tabuľky sa potom do druhej tabuľky zapisoval tak, že najskôr sa vypĺňala biela časť riadkov tabuľky. Až keď boli všetky biele časti riadkov vyplnené, tak sa zvyšná časť správy zapisovala do šedých častí riadkov, opäť začínajúc prvým riadkom zhora a píšuc v smere zľava doprava.

Transponovaná správa sa potom z druhej tabuľky čítala rovnakým spôsobom ako pri obyčajnej tabuľkovej transpozícii, t.j. po stĺpcoch zhora nadol a poradie čítania stĺpcov je určené permutačným heslom (permutáciou stĺpcov). Pri čítaní správy z druhej transpozičnej tabuľky sa už nerozlišovali biele a šedé časti riadkov. Kompletne zašifrovaná depeša z mince mala potom podobu:

```
14546 36056 64211 08919 18710 71187 71215 02906 66036 10922
11375 61238 65634 39175 37378 31013 22596 19291 17463 23551
88527 10130 01767 12366 16669 97846 76559 50062 91171 72332
19262 69849 90251 11576 46121 24666 05902 19229 56150 23521
51911 78912 32939 31966 12096 12060 89748 25362 43167 99841
76271 31154 26838 77221 58343 61164 14349 01241 26269 71578
31734 27562 51236 12982 18089 66218 22577 09454 81216 71953
26986 89779 54197 11990 23881 48884 22165 62992 36449 41742
30267 77614 31565 30902 85812 16112 93312 71220 60369 12872
12458 19081 97117 70107 06391 71114 19459 59586 80317 07522
76509 11111 36990 32666 04411 51532 91184 23162 82011 19185
56110 28876 76718 03563 28222 31674 39023 07623 93513 97175
29816 95761 69483 32951 97686 34992 61109 95090 24092 71008
90061 14790 15154 14655 29011 57206 77195 01256 69250 62901
39179 71229 23299 84164 45900 42227 65853 17591 60182 06315
65812 01378 14566 87719 92507 79517 99651 82155 58118 67197
30015 70687 36201 56531 56721 26306 87185 91796 51341 07796
76655 62716 33588 21932 16224 87721 85519 23191 20665 45140
66098 60959 71521 02334 21212 51110 85227 98768 11125 05321
53152 14191 12166 12715 03116 43041 74822 72759 29130 21947
15764 96851 22370 11391 83520 62297
```

[11] Číslo stĺpca určuje permutačné heslo.

3	0	2	7	4	3	0	4	2	8	7	7	1	2
5	**13**	**2**	**9**	**7**	**6**	**14**	**8**	**3**	**12**	**10**	**11**	**1**	**4**
6	5	7	3	0	9	4	3	3	7	5	7	1	1
9	1	8	9	3	9	1	2	3	3	4	5	4	2
7	9	3	3	6	0	9	6	2	6	1	9	5	0
1	2	1	5	9	2	1	6	1	2	4	1	4	9
5	3	0	1	1	3	1	6	9	0	6	6	6	6
7	1	1	3	2	8	2	0	2	1	5	0	3	1
8	9	3	9	8	8	1	4	6	5	5	1	6	2
3	1	2	7	7	1	6	4	2	6	2	8	0	0
1	2	2	1	2	4	6	1	6	5	9	2	5	6
7	0	5	7	1	8	1	1	9	3	0	0	6	0
3	6	9	5	2	8	2	5	8	1	1	6	6	8
4	6	6	2	4	8	7	1	4	5	1	3	4	9
2	5	1	9	5	4	1	5	9	6	5	1	2	7
7	4	9	8	8	2	5	3	9	7	7	5	1	4
5	5	2	1	1	2	0	2	0	2	2	6	1	8
6	1	9	6	9	1	3	9	2	1	0	5	0	2
2	4	1	9	0	6	1	1	5	2	6	8	8	5
5	0	1	5	8	5	1	1	1	6	7	1	9	3
1	6	7	7	1	6	6	8	1	3	7	2	1	6
2	6	4	6	9	2	4	4	1	0	1	0	9	2
3	0	6	1	7	9	3	2	5	6	9	1	1	4
6	9	3	6	1	9	0	3	7	8	5	3	8	3
1	8	2	9	1	2	4	1	6	7	0	7	7	1
2	6	3	4	7	3	1	6	4	1	1	8	1	6
9	0	5	8	7	6	7	2	6	8	2	1	0	7
8	9	5	3	0	4	4	8	1	5	5	4	7	9
2	5	1	3	1	4	8	2	2	9	6	5	1	9
1	9	8	2	0	9	2	0	1	1	6	6	1	8
8	7	8	9	7	4	2	1	2	7	9	6	8	4
0	1	5	5	0	1	7	1	4	9	2	8	7	1
8	5	2	1	6	7	2	1	6	6	5	7	7	7
9	2	7	9	3	4	7	9	6	5	0	7	1	6
6	1	1	7	9	2	5	1	6	1	6	1	2	2
6	0	0	6	1	3	9	8	0	3	2	9	1	7
2	2	1	8	7	0	2	5	5	4	9	9	5	1
1	3	3	6	1	2	9	5	9	1	0	2	0	3
8	3	0	3	1	6	1	6	0	0	1	5	2	1
2	4	0	4	1	7	3	1	2	7	3	0	9	1
2	2	1	9	4	7	0	1	1	7	9	7	0	5
5	1	7	9	1	7	2	0	9	9	1	7	6	4
7	2	6	2	9	6	1	2	2	6	7	9	6	2
7	1	7	6	4	1	9	8	2	7	9	5	6	6
0	2	1	1	5	4	4	8	9	6	7	1	0	8
9	5	2	1	9	3	7	7	5	6	1	7	3	3
4	1	3	0	5	1	1	6	6	5	2	9	6	8
5	1	6	9	9	5	5	7	1	5	2	9	1	7
4	1	6	9	5	6	7	6	5	6	9	6	0	7
8	0	1	5	8	5	6	7	0	2	2	5	9	2
1	8	6	0	6	3	4	1	2	7	3	1	2	2
2	5	6	9	8	0	9	8	3	1	2	8	2	1
1	2	6	0	0	9	6	0	5	6	9	2	1	5
6	2	9	2	3	0	8	3	2	3	9	1	1	8
7	7	9	4	1	2	5	5	1	3	8	5	3	3
1	9	7	0	7	8	1	6	5	5	4	5	7	4
9	8	8	9	0	5	2	3	1	8	1	5	5	3
5	7	4	2	7	8	2	2	9	8	6	8	6	6
3	6	6	7	5	1	3	8	1	2	4	1	1	1
2	8	7	1	2	2	7	2	1	1	4	1	2	1
6	1	6	0	2	1	0	2	7	9	5	8	3	6
9	1	5	0	7	6	1	2	8	3	9	6	8	4
8	1	5	8	6	1	1	3	9	2	0	7	6	1
6	2	9	9	5	1	3	1	1	1	0	1	5	4
8	5	5	0	0	2	9	6	2	6	4	9	6	3
9	0	0	0	9	9	1	7	3	2	2	7	3	4
7	5	0	6	1	3	8	4	2	2	2	3	4	9
7	3	6	1	1	3	3	3	9	4	2	0	3	0
9	2	2	1	1	1	5	9	3	8	7	0	9	1
5	1	9	4	1	2	2	0	9	7	6	1	1	2
4	5	1	7	1	7	0	2	3	7	5	5	7	4
1	3	1	9	3	1	6	3	1	2	8	7	5	1
9	1	7	0	6	2	2	0	9	1	5	0	3	2
7	5	1	1	9	2	2	7	6	8	3	6	7	6
1	2	7	5	9	0	9	6	6	5	1	8	3	2
1	1	2	1	0	6	7	2						

Tabuľka 3.2: Druhá transpozičná tabuľka pre depešu z mince

Ak si pozorne porovnáme zašifrovanú depešu a fotografiu depeše z mince v prílohe na strane 134, tak zistíme, že jediný rozdiel je v piatej pätici od konca depeše. Do skutočnej depeše sa na určenú pozíciu od konca depeše (v tomto prípade na piatu) zapisovala piata[12], premenlivá časť hesla. Bol to 5-ciferný inicializačný reťazec, ktorý mal v popisovanej depeši tvar: **20818**. Preto je depeša z mince o jednu päticu dlhšia než horeuvedená zašifrovaná depeša. Inicializačný reťazec sa mal voliť pre každú depešu náhodne a podľa možnosti pre každú depešu iný.

3.2.4 Tvorba permutácií z hesla

Ako sme už uviedli, heslo šifry VIC pozostávalo zo štyroch pevných a nemenných častí a jednej náhodne volenej časti – inicializačného reťazca. Jednotlivé časti hesla pre depešu z mince sú uvedené na strane 100. Teraz si popíšeme ako sa z týchto častí hesla zostavovali permutácie pre substitučnú a transpozičné tabuľky. V tejto časti sú nielen popis, ale aj označenia prevzaté z Kahnovho článku [23].

Tvorba hesla pri šifre VIC sa začínala dátumom, v našom prípade 3. septembrom 1945. Dátum sa zapísal v číselnej podobe, v tvare `dDmMYYYY`, čiže dostávame 391945. Dĺžka dátumu podľa uvedeného formátu môže byť 6 až 8 cifier.

Posledná cifra dátumu indikuje, na ktorej pozícii od konca depeše bude umiestnená pätica s inicializačným reťazcom. V našom prípade je to 5, čiže piata pätica od konca depeše bude inicializačný reťazec.

Ďalej si zoberieme inicializačný reťazec 20818 (ozn. ako riadok A) a odčítame od neho prvých 5 cifier dátumu 39194 (riadok B). Odčítanie sa robí modulo 10, t.j. bez prenosu desiatky:

```
  A:   2 0 8 1 8
−B:   3 9 1 9 4
  C:   9 1 7 2 4
```

Výsledné 5-ciferne číslo teraz reťazovým sčítaním modulo 10 natiahneme na 10 cifier. Pod reťazovým sčítaním sa myslí to, že 6. cifru dostaneme ako súčet 1. a 2. cifry, 7. cifru dostaneme ako súčet 2. a 3. cifry atď. Takto dostaneme 10-ciferné číslo:

```
9 1 7 2 4 0 8 9 6 4
```

V ďalších krokoch budeme „vyčíslovať“ textové aj číselné reťazce. To znamená, že budeme určovať poradie znakov, alebo cifier v týchto reťazcoch

[12] Zhoda v čísle 5 je v tomto prípade čisto náhodná.

podľa abecedy, prípadne podľa hodnoty cifier. Prí číselných reťazcoch si treba dať pozor na to, že nula sa berie ako číslo 10. Takže poradie cifier zoradených podľa hodnôt od najmenšej po najväčšiu je: 1 2 3 4 5 6 7 8 9 0.

V nasledujúcom kroku sa zobralo prvých 20 písmen slovnej frázy[13] (riadok D). Týchto 20 znakov sa rozdelí na dve časti po 10 znakov a znaky každej tejto časti sa vyčíslia podľa poradia písmen v ruskej abecede (riadok E). Pod čísla v ľavej časti napíšme 10-ciferné číslo, ktoré sme dostali reťazovým sčítaním riadku C (str. 108). Pod čísla v pravej časti zapíšeme cifry podľa hodnôt (1…0). Cifry v ľavej časti riadkov E a F sčítame modulo 10 a dostávame riadok G.

D:	т	о	л	ь	к	о	с	л	ы	ш	н	о	н	а	у	л	и	ц	е	г
E:	7	4	2	0	1	5	6	3	9	8	6	8	7	1	9	5	4	0	3	2
+F:	9	1	7	2	4	0	8	9	6	4	1	2	3	4	5	6	7	8	9	0
G:	6	5	9	2	5	5	4	2	5	2										

Každú cifru v riadku G teraz vyhľadáme v pravej časti riadku F a nahradíme ju cifrou, ktorá sa nachádza nad ňou v riadku E. Takto dostaneme riadok H a jeho vyčíslením riadok J:

H:	5	9	3	8	9	9	1	8	9	8
J:	3	7	2	4	8	9	1	5	0	6

Riadok J použijeme až v ďalšom kroku. Teraz 10-ciferné číslo z riadku H reťazovým sčítaním modulo 10 predĺžime na 60 cifier. Novovzniknuté cifry budeme zapisovať v riadkoch po 10. Takto dostávame riadky K až P:

H:	5	9	3	8	9	9	1	8	9	8
K:	4	2	1	7	8	0	9	7	7	2
L:	6	3	8	5	8	9	6	4	9	8
M:	9	1	3	3	7	5	0	3	7	7
N:	0	4	6	0	2	5	3	0	4	7
P:	4	0	6	2	7	8	3	**4**	**1**	1

Šírky prvej a druhej transpozičnej tabuľky určíme tak, že vezmeme posledné dve rôzne cifry v riadku P a pričítame k nim osobné číslo agenta. V horeuvedenej tabuľke sú tieto dve cifry 4 a 1 zvýraznené tučným písmom. Šírky transpozičných tabuliek budú:

$$\text{šírka 1. tabuľky:} \quad w_1 = 4 + 13 = 17$$
$$\text{šírka 2. tabuľky:} \quad w_2 = 1 + 13 = 14$$

[13] Tretia časť hesla na strane 100.

V článku [23], z ktorého je prebratý tento popis, je uvedené, že sa vezmú cifry na 8. a 9. pozícii v riadku P. Ako alternatívna varianta je uvedené, že sa vezmú posledné dve rôzne cifry (čo je v tomto prípade to isté ako 8. a 9. cifra). Ťažko povedať, ktorá z týchto dvoch možností je správna a zrejme ani sám Häyhänen si na to nevedel pri výsluchoch spomenúť. My sa z čisto psychologických dôvodov skôr prikláňame k variante posledných dvoch rôznych cifier. Z hľadiska bezpečnosti šifry je úplne jedno[14], či transpozičné tabuľky majú rovnakú alebo rôznu šírku. Ako autori šifry by sme ale radšej zvolili tabuľky rôznej šírky.

Takže máme už šírky oboch transpozičných tabuliek a teraz si musíme zostrojiť transpozičné heslá (permutácie). Tie dostaneme z riadkov K až P tak, že z nich vyberieme potrebný počet cifier (v našom prípade $w_1 + w_2 = 17 + 14$). Cifry budeme čítať po stĺpcoch zhora nadol a poradie stĺpcov nám určí permutácia z riadku J, ktorá je vyčíslením riadku H. Z takto vybratých cifier bude prvých 17 tvoriť riadok Q a ďalších 14 cifier bude tvoriť riadok R:

```
Q:   9 6 0 3 3 1 8 3 6 6 4 6 9 0 4 7 5
R:       3 0 2 7 4 3 0 4 2 8 7 7 1 2
```

Vyčíslením riadkov Q a R dostávame permutácie stĺpcov pre 1. a 2. transpozičnú tabuľku. Pod vyčíslením riadkov máme na mysli, rovnako ako predtým, určenie poradia cifier na riadkoch. Pritom opäť cifra nula má najvyššiu hodnotu.

```
1. transpozícia: 14 8 16 2 3 1 13 4 9 10 5 11 15 17 6 12 7
2. transpozícia:      5 13 2 9 7 6 14 8 3 12 10 11 1 4
```

Takže máme obe transpozičné heslá (permutácie). Zostáva nám ešte vygenerovať permutáciu pre substitučnú tabuľku. Tú dostaneme ako vyčíslenie cifier na riadku P a zapíšeme ju ako riadok S:

```
P:   4  0  6  2  7  8  3  4  1  1
S:   5  0  7  3  8  9  4  6  1  2
```

Týmto sme kompletne ukončili proces vytvárania permutácií pre substitučnú aj obe transpozičné tabuľky.

3.3 Záverečné poznámky k popisu šifry VIC

3.3.1 Substitučná tabuľka

Na strane 102 bolo spomenuté, že pri zápise abecedy do substitučnej tabuľky sa vynechávali tretí a piaty stĺpec. Takto to je uvedené v článku [23]. Ak by

[14] Samozrejme stále hovoríme o šifre VIC.

sa to skutočne robilo takto, t.j. ak by vynechávané stĺpce boli pevne dané, tak by sa tým zjednodušilo lúštenie šifry. Poznali by sme totiž pozície špeciálnych znakov v substitučnej tabuľke a vedeli by sme ich potom identifikovať v zašifrovanej depeši.

Pravdepodobnejšie je, že v Kahnovom článku je chyba. Zdrojom tejto chyby môže byť či už úmyselná, alebo neúmyselná chybná informácia od Häyhänena pri popise šifry VIC. To, ktoré stĺpce sa vynechávali, zrejme nejakým spôsobom záviselo od použitého hesla. V Häyhänenovom prípade to skutočne mohol byť 3. a 5. stĺpec. Čísla týchto stĺpcov mohli byť určené napríklad prvou a poslednou cifrou dátumu (**3**91945), prípadne nejakým zložitejším spôsobom.

Ak by sme pri lúštení nepoznali čísla vynechaných stĺpcov, museli by sme vyskúšať viacero možností. Pri výkone dnešných počítačov by sa tým lúštenie nijako dramaticky neskomplikovalo, ale v časoch, keď šifra VIC bola aktuálna, by táto komplikácia bola významná.

3.3.2 Druhá transpozičná tabuľka

V popise tvorby druhej šifrovacej tabuľky, v článku [23], nie je uvedené, čo robiť ak je šifrovaná správa dlhšia ako transpozičná tabuľka s toľkými šedými trojuholníkmi, koľko má stĺpcov. V prípade depeše z mince tento problém nenastal, pretože jej dĺžka je taká, že druhá transpozičná tabuľka má len 9 zo 14 možných šedých trojuholníkov a pritom deviaty šedý trojuholník ani nie je kompletný a kompletne vyplnený. Pre tento problém sa prirodzene ponúkajú dve možné riešenia:

1. Obmedziť dĺžku šifrovanej správy tak, aby sa zmestila do transpozičnej tabuľky s kompletnou sadou šedých trojuholníkov. Znamenalo by to, že dlhšie správy by bolo nutné rozdeľovať a šifrovať samostatne.

2. Po vyčerpaní kompletnej sady šedých trojuholníkov v druhej transpozičnej tabuľke by sa začali tvoriť ďalšie šedé trojuholníky, opäť začínajúc v stĺpci s číslom 1 a ďalej podľa už uvedeného postupu.

3.3.3 Osobné číslo

V snahe zvýšiť bezpečnosť šifry, zmenili Häyhänenoví nadriadení v roku 1956 jeho osobné číslo z 13 na 20. Tým sa v priemere zväčšili šírky oboch transpozičných tabuliek. Aby toto opatrenie bolo vždy realizovateľné, muselo sa aj reťazové sčítanie čísel z riadku H (str. 109) predĺžiť zo 60 na 70 cifier (o 1 riadok). Toto opatrenie však lúštenie šifry nijako nezmení, len ho môže mierne predĺžiť.

3.3.4 Úvahy o lúštení šifry VIC

V závere článku [23] Kahn píše, že pri znalosti šifrovacieho algoritmu a dostatočnom množstve odchytených depeší, je možné tieto analyzovať a šifru VIC lúštiť[15]. Toto tvrdenie by sme označili za nie celkom presné. Pri našom skúmaní šifry VIC sme zistili, že pri znalosti šifrovacieho algoritmu tak ako je popísaný v Kahnovom článku, a teda aj v tomto texte, je šifra VIC pomerne jednoducho lúštiteľná už pri jedinej odchytenej depeši. To, že kryptoanalytici FBI neuspeli, bolo spôsobené tým, že nevedeli s čím presne majú dočinenia. Ak by poznali postup šifrovania, tak by Häyhänenovu depešu určite rozlúštili.

[15]Toto je približné parafrázovanie podstatne dlhšej poznámky, ale jej zmysel je zachovaný.

3.4 Lúštenie šifry VIC

V tejto časti si ukážeme ako možno, pri znalosti šifrovacieho algoritmu, ľahko lúštiť šifru VIC. Tu uvádzané výsledky sme už publikovali v článku [26] a nasledujúci text je napísaný na báze tohto článku.
Podobne ako pri šifrách s periodickým heslom[16] dokážeme lúštiť zašifrovaný text bez znalosti jazyka otvoreného textu a jeho štatistických vlastností, aj pri lúštení šifry VIC využívame štatistické vlastnosti otvoreného textu depeše len minimálne a na veľmi elementárnej úrovni. Je to vďaka skutočnosti, že návrh šifry VIC porušuje Kerckhoffové zásady v tom, že bezpečnosť šifry nie je podmienená utajením šifrovacieho kľúča, ale je podmienená utajením šifrovacieho algoritmu. Naviac tento šifrovací algoritmus obsahuje vo svojom návrhu niekoľko závažných slabín, ktoré spomenieme v závere. Na druhej strane bez znalosti šifrovacieho algoritmu by bolo lúštenie šifry VIC extrémne obtiažné a prakticky zrejme nemožné. Keďže sa jedná o veľmi komplexnú šifru s vysokou difúziou, ani pri znalosti viacerých zašifrovaných textov, prípadne znalosti zašifrovaného textu a časti zodpovedajúceho otvoreného textu, nevidíme žiadnu prakticky realizovateľnú možnosť lúštenia šifry.

Pre náš útok, ktorý si popíšeme v ďalších častiach, je podstatný spôsob, akým sa zo šifrovacieho kľúča generujú permutácie pre substitučnú a transpozičné časti šifry. Postup generovania týchto permutácií sme popísali v časti 3.2.4.

3.4.1 Aké je teoretické množstvo kľúčov šifry VIC?

V tejto časti si vypočítame teoretické množstvo kľúčov šifry VIC. Tieto úvahy však budú čisto akademického charakteru vzhľadom na to, že:

a) Skutočné množstvo kľúčov je výrazne menšie než počet teoreticky možných kľúčov. Je to napríklad dôsledkom toho, že ako heslá sa v kľúčoch používali len zmysluplné slová a frázy (viď časť 3.2).

b) Pre bezpečnosť šifry VIC, pri znalosti šifrovacieho algoritmu, má počet šifrovacích kľúčov ešte menší význam, než počet možných šifrovacích kľúčov pre jednoduchú substitučnú šifru. Ako je známe, jednoduchá substitúcia na 26-znakovej TSA abecede, má teoreticky $26! \approx 2^{88.4}$ možných kľúčov a pritom jej lúštenie je triviálne.

Pristúpme teda k výpočtu množstva teoreticky možných šifrovacích kľúčov. Budeme postupovať podľa piatich častí šifrovacieho kľúča agenta tak, ako sú tieto uvedené v časti 3.2.

[16] Vigenèrova šífra a jej rôzne varianty.

Prvá časť kľúča: Dátum

Dátumy v kľúči šifry VIC sa zapisovali bežným európskym spôsobom, t.j. vo formáte `dDmMYYYY`. Ako príklad si zoberme Häyhänenov dátum: 3. september 1945. Tento sa zapisoval ako 391945. Takýmto spôsobom dostaneme 360 alebo 361 rôznych čísel pre daný rok, podľa toho či sa jedná o priestupný alebo nepriestupný rok. Existuje totiž päť dvojíc dní roka, ktoré pri európskom spôsobe zápisu dátumu, vedú k rovnakému číslu: 11. 1.–1. 11., 21. 1.–2. 11., 31. 1.–3. 11., 11. 2.–1. 12. a 21. 2.–2. 12.

Pre jednoduchosť predpokladajme, že dátumy kľúča sa vyberali len z rokov 20. storočia. Tým dostávame práve **36 019** rôznych čísel[17].

Druhá časť kľúča: Heslo

Heslo sa v šifre VIC používalo na „premiešanie“ abecedy v substitučnej tabuľke. V skutočnosti sa z hesla používalo len prvých 7 rôznych znakov. Šifra VIC používala azbuku s 30-timi znakmi[18]. Takže ak uvažujeme všetky kombinácie siedmich rôznych znakov, tak dostávame 30.29.28.27.26.25.24 možností, čo je presne **10 260 432 000** možných hesiel.

Tretia časť kľúča: Slovná fráza

Zo slovnej frázy sa používalo len prvých 20, nie nutne rôznych, znakov. Opäť, ak uvažujeme 20-znakový reťazec na azbuke s 30-timi znakmi, tak dostávame práve 30^{20}, čo je presne

348 678 440 100 000 000 000 000 000 000

rôznych slovných fráz.

Štvrtá časť kľúča: Osobné číslo agenta

Malé, 1-znakové, osobné čísla nie sú vhodné, pretože by viedli k transpozičným tabuľkám príliš malej šírky. Pre verziu šifry VIC, ktorú používal Häyhänen, má zmysel používať ako osobné čísla agenta, len čísla od 10 po 19. Häyhänen mal osobné číslo 13 a neskôr, v roku 1956, mu ho jeho nadriadení zmenili na 20. Táto zmena si však vyžiadala aj drobnú úpravu šifrovacieho procesu, aby bolo šifrovanie možné pre ľubovoľnú depešu[19]. Preto, pre Häyhänenovu verziu šifry VIC, máme **10** možných hodnôt osobných čísel agenta.

[17] V 20. storočí bolo 19 priestupných rokov.
[18] Vynechávali sa 3 znaky s diakritikou.
[19] Spomenuté v časti 3.3.3.

Piata časť kľúča: Náhodné 5-ciferné číslo

Piata časť šifrovacieho kľúča bolo náhodne zvolené 5-ciferné číslo a tých je 10^5, čize **100 000**, rôznych možných.
Ak teraz navzájom skombinujeme počet možností pre všetkých päť častí šifrovacieho kľúča, tak dostávame:

$$36\,019 \times 10\,260\,432\,000 \times 30^{20} \times 10 \times 10^5 =$$

$$= 128\,861\,265\,519\,502\,165\,540\,800\,000\,000\,000\,000\,000\,000\,000\,000 \doteq$$

$$\doteq 1.28 \times 10^{50} \doteq 2^{166}$$

teoreticky možných šifrovacích kľúčov. Toto číslo je astronomické. Ako už ale bolo spomenuté skôr, bezpečnosť šifry VIC s ním nesúvisí a celá táto časť bola venovaná len akademickým úvaham.

3.4.2 Slabiny šifry VIC

V časti 3.2.4 sme si ukázali ako sa z piatich častí šifrovacieho kľúča, generujú permutácie použité pre substitučnú tabuľku a obe transpozície. Z popisu tohto postupu je vidno, že „úzke hrdlo“ celého tohto procesu je v kroku, kde sa generuje „riadok H“. Riadok H, pozostávajúci z desiatich cifier, sa v ďalšom postupe reťazovým sčítaním predlžuje na 60 cifier. Z týchto 60-tich cifier sa potom priamo odvodzujú permutácie pre substitučnú aj transpozičné tabuľky. Týmto sa astronomický počet 1.28×10^{50} teoreticky možných kľúčov redukuje na 10^{10} možných hodnôt v riadku H.

Túto slabinu šifry VIC si vo svojej práci [4] všimol aj Radoslav Čagala. Tým sa časová náročnosť lúštenia hrubou silou výrazne zníži a útok sa stáva realizovateľným, avšak stále ešte zostáva časová náročnosť príliš vysoká na lúštenie šifry VIC v reálnom čase. Okrem toho autor uvedenej práce nemal k dispozícii kompletný a presný popis šifry VIC, takže analyzoval a lúštil len jej modifikovanú podobu.

Ak by teda útočník chcel hrubou silou vyskúšať všetky možné permutácie, tak mu stačí skúsiť len 10^{10} možností. Vôbec nepotrebuje skúšať všetky možné šifrovacie kľúče, pretože na samotné šifrovanie sa používajú permutácie a nie šifrovací kľúč agenta. Šifrovací kľúč agenta iba slúži na odvodenie týchto permutácií.

3.4.3 Útok na šifru VIC

Ako už bolo spomenuté, šifra VIC je zložená z jedno a dvojmiestnej zámeny a dvoch transpozícií, z ktorých prvá je bežná tabuľková transpozícia a druhá je

špeciálna obrazcová transpozícia. Útok na takúto šifru by bol vo všeobecnosti nesmierne náročný a na základe len jedného zašifrovaného textu prakticky nemožný.

V prípade šifry VIC sú ale substitúcia aj obe transpozície postavené na permutačných heslách, ktoré sa tvoria veľmi zložitým spôsobom zo šifrovacieho kľúča agenta. Je pritom zrejmé, že proces tvorby týchto permutácií nekontroloval žiaden kryptoanalytik, pretože útok na šifru VIC je pomerne jednoduchý.

3.4.4 Krok prvý – substitúcia

Vieme, že v šifre VIC sa používa jedno a dvojmiestna zámena. Veľmi dobre známou vlastnosťou takejto substitúcie je fakt, že pri zašifrovaní ľubovoľného zmysluplného a dostatočne dlhého textu, budú mať cifry riadkov substitučnej tabuľky najvyššie relatívne frekvencie výskytu. Jedine v prípade veľmi krátkych a/alebo umelo zvolených textov by toto pravidlo mohlo byť porušené. A skutočne v Häyhänenovej depeši, ktorú FBI našla v preparovanej minci, boli relatívne frekvencie výskytu cifier nasledovné:

cifra	1	2	6	9	5	7	0	3	8	4
# výskytov	191	132	118	107	97	96	81	80	70	58
frekvencia (%)	18.5	12.8	11.5	10.4	9.4	9.3	7.9	7.8	6.8	5.6

pričom použitá substitučná tabuľka mala podobu:

	5	0	7	3	8	9	4	6	1	2
—	**С**	**Н**	**Е**	**Г**	**О**	**П**	**А**			
6	Б	Ж	.	К	№	Р	Ф	Ч	Ы	Ю
1	В	З	,	Л	Н/Ц	Т	Х	Ш	Ь	Я
2	Д	И	П/Л	М	НТ	У	Ц	Щ	Э	ПВТ

Tabuľka 3.3: Häyhänenova substitučná tabuľka

Takže z analýzy relatívnych frekvencií výskytu cifier v zašifrovanom texte depeše vieme, že cifry označujúce riadky substitučnej tabuľky sú 1, 2 a 6 alebo 9. Okrem toho však vieme aj to, že cifry použité na označenie riadkov substitučnej tabuľky sú zároveň aj posledné tri cifry 10-ciferného čísla označujúceho stĺpce tejto tabuľky. Teraz sa pozrime na to, ako je toto 10-ciferné číslo generované, počnúc riadkom H v časti 3.2.4.

Začneme 10-ciferným číslom v riadku H, ktorý sme získali komplikovaným postupom. Toto 10-ciferné číslo reťazovým sčítaním predĺžime na 60-ciferné číslo. Potom vezmeme posledných 10 cifier vytvoreného 60-ciferného čísla, vyčíslime ich lexikograficky a tým dostaneme 10-cifernú permutáciu označujúcu stĺpce substitučnej tabuľky.

Stačí nám teda otestovať $10^{10} < 2^{34}$ možných „štartovacích čísel“. Reťazové sčítanie a lexikografické vyčíslenie posledných 10-tich cifier sú triviálne a časovo nenáročné operácie. Takto vygenerujeme všetky možné 10-ciferné permutácie označujúce stĺpce substitučnej tabuľky a pozrieme sa, či posledné 3 cifry týchto čísel sú najfrekventovanejšie cifry zašifrovanej depeše. Ak tomu tak je, tak použité „štartovacie číslo“ by mohlo byť správne a musíme si ho zapamätať pre ďalšie testovanie. Pokiaľ tomu tak nie je, tak „štarovacie číslo“ je nesprávne a môžme ho vylúčiť.

Použitím tohto jednoduchého testu na Häyhänenovej depeši, rýchlo vylúčime viac než 99.6% možných „štartovacích čísel“. Celý uvedený test prebehne na bežnom stolnom počítači s jedným jednojadrovým procesorom za približne 39 minút. Pritom už samotný princíp tohto testu je taký, že nie je problém úlohu rozdeliť na ľubovoľný počet častí a spustiť na ľubovoľnom dostupnom množstve jadier/procesorov/počítačov a tým úlohu vykonať v ľubovonom vopred stanovenom čase,.

Po tomto úvodnom teste nám zostalo už iba $51\,898\,770 < 2^{26}$ „štartovacích čísel“, ktoré prichádzajú do úvahy a je potrebné ich ďalej testovať. V celom doterajšom postupe sme samotnú depešu potrebovali len na určenie relatívnych frekvencií cifier jej zašifrovaného textu. To je triviálna operácia so zanedbateľnou časovou náročnosťou, ktorú sme spravili hneď na začiatku.

3.4.5 Krok druhý – eliminácia transpozícií

Po teste vykonanom v prvom kroku máme už len obmedzenú množinu „štartovacích čísel“, ktoré môžme použiť na generovanie permutácií. Tieto čísla teraz musíme otestovať jedno po druhom. Pre každé z týchto čísel vygenerujeme permutácie substitučnej, aj oboch transpozičných tabuliek. V skutočnosti, pri softwarovej realizácii, sa toto udialo už v prvom kroku. Vzhľadom na postup generovania permutácií z riadku H, popísaný v časti 3.2.4, je najefektívnejší postup vygenerovať permutácie pre transpozičné tabuľky ihneď po otestovaní permutácie pre substitúciu. K tomu potrebné 60-ciferné číslo, ktoré sme z riadku H dostali reťazovým sčítaním už totiž máme vygenerované a vytvoriť z neho potrebné permutácie vyžaduje len niekoľko triviálnych a časovo nenáročných operácií (viď časť 3.2.4). Takže akonáhle sa niektoré „štartovacie číslo“ ukáže byť použiteľné, zapamätáme si ho spolu so všetkými troma permutáciami.

Permutácie pre obe transpozičné tabuľky závisia nielen na 10-cifernom „štartovacom čísle“, ale aj na osobnom čísle agenta[20]. Preto pre každé použiteľné „štartovacie číslo“ musíme vyskúšať niekoľko možných osobných čísel agenta a detekovať všetky použiteľné dvojice „štartovacích čísel“ a osobných čísel agenta.

Akonáhle máme vytvorené permutácie transpozičných tabuliek, môžme invertovať obe transpozície a získame tak pôvodný text zašifrovaný už len jedno a dvojmiestnou zámenou. Toto je vlastne prvý krok celého lúštenia, kde pracujeme s textom zašifrovanej depeše. Takže pre každú použiteľnú dvojicu „štartovacieho čísla“ a osobného čísla agenta eleminujeme zo zašifrovaného textu transpozície a ďalej budeme testovať už iba substitúciu. Týmto dokážeme rýchlo vylúčiť veľkú časť doteraz použiteľných „štartovacích čísel“.

3.4.6 Pred tretím krokom

Jedno- dvojmiestna zámena, popísaná v Kahnovom článku [23], je definovaná príliš rigídnym spôsobom a triviálne lúštiteľná. Takto definovaná substitúcia je dokonca jednoduchšie lúštiteľná než jednoduchá zámena, a to vďaka pevne definovanému rozloženiu znakov v substitučnej tabuľke. Existujú len dve vysvetlenia tohto faktu. Prvé z nich je pre bezpečnosť šifry VIC zlé a druhé ešte horšie. V oboch prípadoch je bezpečnosť použitej substitúcie veľmi nízka.

Prvé vysvetlenie je, že popis zostavovania substitučnej tabuľky v článku [23] je nesprávny. Toto je s vysokou pravdepodobnosťou to správne vysvetlenie a zrejme to nie je chyba Dr. Kahna. Popis šifry VIC prezradil americkým vyšetrovateľom Häyhänen, po svojom prebehnutí v roku 1957. Podľa príbehu o Häyhänenovi, ktorý je podrobne popísaný v článkoch [24] a [37], bol Häyhänen ťažký alkoholik a jeho výpovede pred vyšetrovateľmi FBI neboli veľmi spoľahlivé ani dôveryhodné. Je preto celkom možné a pravdepodobné, že aj keď vyšetrovateľom správne popísal postup šifrovania pre svoju verziu šifry VIC, zabudol, alebo možno ani nevedel, že popisovaný postup sa týka výlučne jeho verzie šifry a jemu prideleného kľúča. Postup zostavovania substitučnej tabuľky je v článku [23] na strane A17. Podľa uvádzaného postupu sa špeciálne znaky[21] zapisovali len do tretieho a piateho stĺpca tabuľky. A nielen to, ale aj v pevne stanovenom poradi. Takýto postup by bol príliš rigídny a nesmierne uľahčuje a predovšetkým urýchľuje lúštenie. Ak totiž poznáme pozície špeciálnych znakov v substitučnej tabuľke, vieme jednoduchým a veľmi rýchlym testom eliminovať takmer všetky zostávajúce „štartovacie čísla“ bez toho, aby sme museli lúštiť samotnú depešu. Je preto veľmi pravdepodobné,

[20] Štvrtá časť kľúča.

[21] Prepínače, indikátor začiatku textu a pod.

že v skutočnosti sa stĺpce, do ktorých sa zapisovali špeciálne znaky, určovali na základe kľúča prideleného agentovi. Napríklad v Häyhänenovom prípade by sa mohlo jednať o dátum[22], ktorý bol 3. september 1945. Jeho európsky zápis je **3**91945, čiže prvá a posledná cifra dátumu by mohla určovať stĺpce pre zápis špeciálnych znakov do substitučnej tabuľky. Zmenou dátumu by sa potom zmenila aj pozícia špeciálnych znakov, čo by trochu predĺžilo lúštenie. Tieto úvahy sú ale len našími hypotézami a dnes už zrejme ťažko niekto overí ako to bolo v skutočnosti.

Druhé vysvetlenie, ktoré by pre bezpečnosť šifry VIC bolo podstatne horšie je to, že popis zostavovania substitučnej tabuľky v článku [23] je správny. Treba len dúfať, že toto nie je to správne vysvetlenie.

V treťom kroku lúštenia šifry VIC spravíme ďalší rýchly test použiteľnosti „štartovacieho čísla". Tento test bude založený na kontrole štruktúry depeše zašifrovanej už len jedno a dvojmiestnou zámenou. Takže stále sa nejedná o pokusy lúštiť depešu a získanie otvoreného textu.

Najskôr ukážeme prípad, kedy budeme predpokladať, že popis zostavovania substitučnej tabuľky v [23] je správny. Tento prípad je totiž jednoduchšie realizovateľný. Potom v ďalšej časti ukážeme ako by sa dala situácia zachrániť v prípade, že nepoznáme pozície špeciálnych znakov v substitučnej tabuľke. Tento druhý prípad je trochu náročnejší na realizáciu a vedie k vyššiemu počtu použiteľných „štartovacích čísel".

3.4.7 Krok tretí – testovanie substitúcie

Skôr ako sa pustíme do prvého testu substitúcie, musíme si pripomenúť jednu jej zvláštnosť. V substitúcii pri šifre VIC sa cifry 0 až 9 nesubstitujú, pretože cifry nie sú obsiahnuté v substitučnej tabuľke. Namiesto toho sa cifry 0 až 9 kódujú. V substitučnej tabuľke je špeciálny znak **Н/Ц**, ktorý slúži ako prepínač módu medzi textom a zápisom cifier. Ak teda chceme zapísať nejaké číslo, tak mu musí predchádzať tento prepínač, potom sa každá cifra opakuje trikrát a napokon opäť nasleduje uvedený prepínač. Prepínač **Н/Ц** je v substitučnej tabuľke v piatom stĺpci na treťom riadku[23]. Preto, keď sa budeme pokúšať rozdeliť depešu, zašifrovanú len substitúciou, na jedno a dvojmiestne čísla, musíme mať na pamäti, že depeše môžu obsahovať aj trojciferné číselné kódy.

1. **Prvý test:**

   V tomto teste sa pokúsime depešu, zašifrovanú už len jedno a dvojmiestnou zámenou, rozdeliť na jedno a dvojciferné čísla zodpovedajúce

[22] Prvá časť kľúča.

[23] V Häyhänenovej depeši má kód 18. Takže číslo 1 by sme zapísali ako: `18 111 18`.

jednotlivým znakom. Ak takéto rozdelenie nie je možné, tak použité „štartovacie číslo“ nie je prípustné.

Pre každé „štartovacie číslo“ poznáme permutáciu označujúcu stĺpce, a teda aj riadky, substitučnej tabuľky. Začneme prvou cifrou depeše. Ak je to cifra označujúca riadky, tak k nej vezmeme aj ďalšiu cifru a dostávame dvojciferný kód znaku. Inak túto cifru berieme len ako jednociferný kód znaku z prvého riadku tabuľky. Potom pokračujeme s ďalšou cifrou v poradí.

Ak sme v predošlom kroku dostali dvojciferný kód znaku, musíme skontrolovať, či sa nejdná o **Н/Ш** prepínač módu. Ak áno, tak za ním musia nasledovať trojice identických cifier a po nich musí nasledovať opäť **Н/Ш** prepínač. Inak je použité „štartovacie číslo“ nesprávne a testovanie môžme ukončiť. Jediná výnimka môže nastať na konci depeše, ktorý je doplnený nulami[24] tak, aby dĺžka depeše bola násobkom 5. Takže na posledných 4 cifrách depeše je zlyhanie tohto testu tolerované.

Len ak celú depešu dokážeme rozdeliť na jedno a dvojciferné[25] číselné kódy, podľa popísaných pravidiel, budeme pokračovať ďalším testom. Inak použité „štartovacie číslo“ zamietame.

2. **Druhý test:**

   Každá depeša zašifrovaná šifrou VIC je pred substitúciou náhodne rozdelená na dve časti. Najskôr sa substituje druhá časť, potom nasleduje znak **НТ**, označujúci začiatok depeše a za ním sa substituje prvá časť depeše. To znamená, že po substitúcii musí každá depeša obsahovať práve jeden znak **НТ**. My naviac vieme, že tento znak sa nachádza v piatom stĺpci a štvrtom riadku substitučnej tabuľky a poznáme aj jeho kód. Musíme teda potencionálnu depešu skontrolovať na tento znak. Kontrolujeme už, v predošlom kroku, rozdelenú depešu. Ak táto neobsahuje práve jeden výskyt znaku **НТ**, tak použité „štartovacie číslo“ je nesprávne a môžme ho zamietnuť.

Po zbehnutí týchto dvoch jednoduchých testov na Häyhänenovej depeši nám zostalo presne 11 392 použiteľných dvojíc „štartovacich čísel“ a osobných čísel agenta. Inými slovami, dostali sme 11 392 depeší zašifrovaných už len jednoduchou zámenou, ktoré treba skúsiť vylúštiť a overiť, či vedú k zmysluplnému ruskému textu. Lúštenie jednoduchej zámeny je triviálna a veľmi

[24] Nulami v kryptografickom význame tohto slova.
[25] V prípade čísel aj trojciferné.

rýchlo realizovateľná úloha, takže týmto môžme považovať lúštenie šifry VIC za ukončené.

3.4.8 Modifikovaný tretí krok

V tejto časti si ukážeme ako by sa dalo pri lúštení šifry VIC postupovať ak by sme nevedeli, v ktorých stĺpcoch substitučnej tabuľky sú zapísané špeciálne znaky, avšak ostatné časti zostavovania substitučnej tabuľky by zodpovedali popisu z článku [23].

Testy z tretieho kroku zostávajú v zásade rovnaké ako sme ich popísali v predošlej časti. Spravíme len jednu malú modifikáciu pridaním ďalšieho testu na začiatok. V dôsledku použitej heuristiky tohto úvodného testu dostaneme trochu viac použiteľných dvojíc „štartovacich čísel“ a osobných čísel agenta.

1. **Úvodný test:**

   S veľmi vysokou pravdepodobnosťou bude každá depeša, zašifrovaná šifrou VIC, obsahovať čísla[26]. Tento fakt využijeme v úvodnom teste potencionálnej depeše.

   V tomto teste sa pokúsime identofikovať čísla v substituovanej potencionálnej depeši. V tomto prípade nepoznáme stĺpec, v ktorom je umiestnený **Н/Ц** prepínač. Vieme len, že tento prepínač je niekde v treťom riadku tabuľky a jeho kód je xy, kde x je cifra tretieho riadku a y je neznáma cifra.

   V prípade šifry VIC existujú len tri prípady, kedy sa nejaká cifra, v už substituovanej, depeši opakuje trikrát po sebe. Po prvé, ak sa jedná o kód nejakej cifry. Po druhé, ak sa v depeši trikrát po sebe opakuje znak z prvého riadku substitučnej tabuľky (napr. www). A napokon po tretie, ak je v depeši znak zakódovaný dvoma rovnakými ciframi riadku a stĺpca a za ním nasleduje znak z toho istého riadku substitučnej tabuľky. Druhý uvedený prípad, aj keď je málo pravdepodobný, je tiež možný a musíme ho brať do úvahy. Tretí prípad je tiež vzácny, pretože by to znamenalo, že depeša obsahuje bigram zostavený z dvoch znakov s nízkou relatívnou frekvenciou výskytu[27]. Najvyššiu pravdepodobnosť má prvý prípad, t.j. najpravdepodobnejšie sa jedná o kód cifry.

[26] Napríklad dátumy, počty položiek a pod.

[27] Pri jedno a dvojmiestnej zámene platí kryptografická zásada, že znaky s vysokou relatívnou frekvenciou výskytu sa dávajú do prvého riadku tabuľky, t.j. kóduju sa jednocifernými číslami.

Začneme teda hľadať po sebe idúce trojice identických cifier v potenciálnej depeši. Ak takú trojicu nájdeme, pozrieme sa na dvojice cifier bezprostredne pred ňou a za ňou[28]. Ak sú tieto dvojice cifier pred a za nájdenými identickými trojicami v tvare xy a naviac s rovnakými číslami xy, kde x je cifra tretieho riadku, tak sa s vysokou pravdepodobnosťou jedná o kódovanie cifier. Inak by sa mohlo jednať buď o druhý alebo tretí uvedený prípad, ktoré sú vzácne. Takže budeme počítať takéto prípady a keď ich počet presiahne určitú stanovenú hranicu, tak použité „štartovacie číslo“ zamietneme.

Po realizácii tohto testu už poznáme aj pozíciu stĺpca špeciálnych znakov **H/Ц** a **HT** a môžme pokračovať s ďalšími testami, tak ako boli popísané v predošlej časti.

2. **Prvý test:**

   V úvodnom teste sme identifikovali kód **H/Ц** prepínača. Pokúsime sa teda rozdeliť potencionálnu depešu na jedno-, dvoj- a prípadne trojciferné kódy podľa substitučnej tabuľky. Pritom budeme rešpektovať už identifikované a rozdelené kódy cifier a zvyšný postup je úplne rovnaký ako v prvom teste popisovanom v predošlej časti. Ak rozdelenie depeše nie je možné, tak použité „štartovacie číslo“ zamietneme. Inak pokračujeme ďalším testom.

3. **Druhý test:**

   Kód špeciálneho znaku pre začiatok depeše (**HT**) sme identifikovali už v úvodnom teste, vzhľadom na to, že sa nachádza v rovnakom stĺpci substitučnej tabuľky ako **H/Ц** prepínač. So znalosťou tohto kódu je druhý test úplne rovnaký ako druhý test popísaný v predošlej časti.

3.5 Implementácia a výsledky

Implementovali sme prvú verziu útoku, predpokladajúcu, že popis zostavovania substitučnej tabuľky pre šifru VIC, uvedený v článku [23], je správny. Tento prípad je jednoduchšie realizovateľný, priamočiarejší a neobsahuje žiadne heuristiky. Všetky testy boli robené na originálnej Häyhänenovej depeši. Všetky uvádzané výsledky, s jedinou výnimkou spomenutou neskôr, boli dosiahnuté na bežnom notebooku s i7 procesorom a spúšťané ako jeden proces na jednom jadre. Spomenutou výnimkou bol experiment paralelizácie testov.

[28] Ak identických trojíc cifier ide v depeši viacero za sebou, tak kontrolujeme až dvojicu za poslednou z týchto identických trojíc.

Všetky uvedené testy sú navzájom nezávislé, a preto sa dajú paralelizovať veľmi triviálnym spôsobom, jednoduchým rozdelením vstupných dát na viacero častí a ich paralelným testovaním na viacerých jadrách, procesoroch alebo počítačoch. Použitý i7 procesor má štyri jadrá. Skúsili sme preto ako sa zmení časová náročnosť testu rozdelením vstupných dát na štyri rovnako veľké časti a ich paralelným testovaním na všetkých štyroch jadrách súčasne, v porovnaní s časovou náročnosťou pri behu len na jednom jadre. Pamäťová náročnosť všetkých testov je prakticky zanedbateľná a z toho dôvodu ani nie je uvádzaná.

3.5.1 Prvý test

Celá implementácia útoku na šifru VIC bola rozdelená na dve samostatné časti. V prvej časti sa testovala permutácia pre substitučnú tabuľku, popísaná ako prvý krok v časti 3.4.4. Pri tomto teste sa, okrem určenia relatívnych frekvencií cifier zašifrovaného textu, vôbec nepracuje s textom depeše. Preto má zmysel robiť tento test samostatne. Testovali sme 10^{10} čísel od 0000000000 po 9999999999. Z týchto „štartovacích čísel“ sme generovali permutácie substitučnej tabuľky a overovali sme, či ich posledné tri cifry sú cifry s najvyššími relatívnymi frekvenciami. Ak tomu tak bolo, tak sme si použité „štartovacie číslo“ zapamätali, v opačnom prípade sme ho zamietli. Výsledkom bol zoznam prípustných „štartovacích čísel“.

Počet vstupných čísel	**Počet použiteľných čísel**	**Čas testovania**
10^{10}	51 898 770	$38'38''$

Tabuľka 3.4: Výsledky prvého testu

Ako je z uvedených výsledkov zrejmé, prvým testom sme eliminovali viac než 99.6 % „štartovacích čísel“.

3.5.2 Druhý a tretí test

Druhý aj tretí test už boli robené len na množine použiteľných „štartovacích čísel“, ktoré prešli prvým testom. Tie dva ďalšie testy už pracujú aj so zašifrovanou depešou a berú do úvahy osobné číslo agenta.

Keďže druhý a tretí test spolu úzko súvisia, boli implementované spoločne. Na strane 114 sme uviedli, že pre Häyhänenovu verziu šifry VIC malo zmysel ako osobné čísla agenta používať len čísla 10–19. Pre testovacie účely sme však použili interval 1–20 v prvej implementácii a interval 10–20 v druhej implementácii.

Druhý test je popísaný na strane 119 ako prvý test v treťom kroku. Pre 51 898 770 použiteľných štartovacích čísel a osobných čísel agenta z intervalu 1–20, dostávame 1 037 975 400 dvojíc „štartovacích čísel“ a osobných čísel agenta, ktoré treba otestovať. Po zbehnutí druhého testu, ktorý sa len pokúšal rozdeliť zašifrovanú depešu na kódy znakov, sme z tohto počtu vylúčili 786 478 739 možností. Druhým testom sme teda eliminovali približne 76.8 % zostávajúcich možností.

Tretí test, popísaný na strane 120 ako druhý test v treťom kroku, už bežal len na zostávajúcich dvojiciach „štartovacích čísel“ a osobných čísel agenta. Po zbehnutí druhého testu sme eliminovali ďalších 248 471 874 možností, čo je zhruba 23.9 % pôvodných všetkých možností.

Ako už bolo spomenuté skôr, v Häyhänenovej verzii šifry VIC má zmysel používať len osobné čísla z intervalu 1–19. Menšie čísla môžu viesť k príliš malej šírke transpozičných tabuliek a osobné číslo 20 môže, pre niektoré „štartovacie čísla“, viesť k transpozičným tabuľkám, ktorých súčet šírok je väčší než 50 a teda nerealizovateľný. Preto sme do našej implementácie zahrnuli ešte doplňujúci test, ktorý predchádzal druhý a tretí test. Tento doplňujúci test len overoval, či pri danom „štartovacom čísle“ a 20 ako osobnom čísle agenta, je možné zostaviť transpozičné tabuľky. Pri zbehnutí tohto doplňujúceho testu sme našli 3 013 395 takých „štartovacích čísel“, ktoré sa spolu s osobným číslom 20 nedajú použiť. To je približne 0.3 % všetkých možných dvojíc.

3.5.3 Výsledky druhého a tretieho testu

Oba testy spolu eliminovali 1 037 964 008 dvojíc „štartovacích čísel“ a osobných čísel agenta. To je približne 99.999 % všetkých možností. Zostalo nám 11 392 možných dvojíc, čiže približne 0.001 % z celkového počtu. Inými slovami, lúštenie depeše, zašifrovanej šifrou VIC, vieme previesť na lúštenie 11 392 depeší zašifrovaných jednoduchou zámenou. Toto je jednoduchá a rýchlo realizovateľná úloha.

Pretože osobné čísla agenta z intervalu 1–9 nemá zmysel používať, skúsili sme našu implementáciu upraviť tak, aby sa testovali len osobné čísla z intervalu 10–20. V takomto prípade máme 570 886 470 všetkých možnosti, z ktorých nám po druhom a treťom teste zostane len 6 463.

Posledný experiment bola paralelizácia celého procesu. Množina prípustných „štartovacích čísel“ po prvom teste, bola rozdelená na štyri rovnako veľké časti. Potom druhý a tretí test bežali paralelne na týchto častiach, na štyroch jadrách jedného procesora. Nebola pritom robená žiadna optimalizácia pri prideľovaní procesov na jednotlivé jadrá, prípadne úprave priority procesov a pod. Všetky nastavenia boli ponechané na sheduler OS (Kubuntu

Procesy / jadrá	**1 / 1**	**1 / 1**	**4 / 4**
Počet možností po prvom teste	51 898 770	51 898 770	51 898 770
Rozsah osobných čísel agenta	1–20	10–20	1–20
Prípustné možnosti	11 392	6 463	11 392
Čas behu	7 h 17 m 50 s	4 h 17 m 18 s	2 h 16 m 0 s

Tabuľka 3.5: Výsledky druhého a tretieho testu

Linux). Počas behu testov bolo vidno, že jednotlivé procesy mnohokrát preskakovali medzi jadrami procesora. Preto aj celkový čas, pri takejto primitívnej forme paralelizácie, bol väčší než teoreticky očakávaná štvrtina pôvodného času. Prehľad všetkých výsledkov je v tabuľke 3.5.

Z uvedených výsledkov je zrejmé, že celý proces lúštenia šifry VIC vieme realizovať v ľubovoľne vopred stanovenom čase, jednoduchým rozdelením vstupných dát medzi dostatočné množstvo jadier/procesorov/počítačov. Pritom toto množstvo potrebných HW prostriedkov nieje v súčasnosti nijako prehnané. Na notebooku sme schopní celé lúštenie realizovať rádovo v hodinách, na stredne velkom serveri, s 32 jadrami, je to už len otázka minút.

V tabuľke 3.6 vidíme ako uskutočnené tri testy postupne redukovali počet možných dvojíc „štartovacích čísel“ a osobných čísel agenta. V tejto tabuľke sme pracovali s osobnými číslami z intervalu 1–20. Okrem toho musíme brať do úvahy, že prvý test s osobným číslom agenta vôbec nepracuje. Takže v prvom teste pracujeme s 10^{10} „štartovacími číslami“ a výsledky tohto testu sú platné pre všetky osobné čísla agenta.

	Štart	**Po 1. teste**	**Po 2. teste**	**Po 3. teste**
Počet	$10^{10} \times 20$	$51\,898\,770 \times 20$	248 483 266	11 392
Percento	100 %	$\approx 0.519\,\%$	$\approx 0.124\,\%$	$\approx 0.006\,\%$

Tabuľka 3.6: Ako realizované testy eliminujú prípustné možnosti

3.6 Poznámka na záver

Každá depeša zašifrovaná šifrou VIC obsahuje náhodné 5-ciferné číslo zapísané niekde na posledných desiatich pozíciach zašifrovanej správy. Je to piata, náhodná, časť kľúča, ktorej pozícia v depeši závisí od dátumu.

V našich experimentoch sme používali zašifrovaný text, ktorý už túto náhodnú päticu neobsahoval. Pri reálnom lúštení depeše zašifrovanej šifrou VIC, by sme uvedený proces museli opakovať 10 krát, pre každú možnú pozíciu tejto náhodnej pätice. Ak by sme však mali viacero depeší od toho istého agenta, tak toto by stačilo spraviť len pri lúštení prvej depeše, pretože pozícia tejto pätice bude vo všetkých depešiach jedného agenta rovnaká.

3.7 Protiopatrenia voči uvedenému útoku

V tejto časti si popíšeme protiopatrenia, ktoré by zabránili popísanému útoku. Náš útok bol založený na troch pilieroch:

1. Úzke hrdlo pri vytváraní permutácií z kľúča.
2. Disproporcie pri relatívnych frekvenciách znakov použitej substitúcie.
3. Rigidné zostavovanie substitučnej tabuľky.

Vďaka týmto trom vlastnostiam šifry VIC môže lúštiteľ úplne zabudnúť na komplikovanú tvorbu permutácií z kľúča. Stačí overiť pomerne malé množstvo[29] možností. Okrem toho aj z tohto malého množstva dokážeme väčšinu možností rýchlo eliminovať a zredukovať lúštenie komplikovanej šifry VIC na lúštenie viacerých inštancií jednoduchej zámeny.

3.7.1 Úzke hrdlo pri vytváraní permutácií z kľúča

Šifra VIC používala šifrovací kľúč, ktorý pozostával z piatich častí. Dve z nich boli znakové, tri číselné. Útok na kľúč VIC hrubou silou by ešte ani dnes nebol realizovateľný. Avšak pri samotnom šifrovaní sa v šifre VIC nepoužívali kľúče a ich časti priamo, ale pomocou týchto kľúčov a komplikovaného algoritmu sa generovali tri permutácie a tie slúžili na šifrovanie, resp. aj dešifrovanie. Algoritmus generovania permutácií bol síce komplikovaný ale vo svojom závere končil 10-ciferným číslom. Následne sa z tohto čísla odvodili všetky tri generované permutácie. Preto toto 10-ciferné číslo v texte nazývame „štartovacie číslo“. A práve toto „štartovacie číslo“ je úzkym hrdlom generovania

[29] V porovnaní so všetkými teoreticky možnými.

kľúčov. Útočník sa totiž nemusí starať o to aký je kľúč a o komplikované odvádzanie permutácií z kľúča. Útočník môže začať rovno od „štartovacieho čísla“, čím si zníži astronomický počet možností kľúča na iba 10^{10} možností pre toto číslo.
Existujú dva jednoduché spôsoby ako odstrániť toto úzke hrdlo:

- Môžme úplne od základov zmeniť proces spracovania kľúča a navrhnúť nejaký iný spôsob generovania permutácií na šifrovanie. To by však viedlo k úplnej novej šifre, čo je mimo nášho záujmu.

- Môžme zachovať väčšinu postupu generovania permutácií z kľúča, akurát v záverečnej fáze nebudeme generovať 10-ciferné číslo (riadok H v časti 3.2.4), ale dlhšie n-ciferné číslo, z ktorého potom odvodíme všetky tri permutácie. V takomto prípade by postačil malý zásah do šifrovacieho algoritmu šifry VIC.

3.7.2 Disproporcie pri relatívnych frekvenciách znakov použitej substitúcie

Jedno- a dvojmiestna zámena použitá v šifre VIC s tabuľkou, ktorá má 4 riadky a 10 stĺpcov pre 30 znakov azbuky a 7 špeciálnych znakov, nie je moc štastne zvolená. Rozloženie znakov v tabuľke je dané čiastočne heslom, ktoré je súčasťou kľúča a čiastočne je dané pevne. Takáto tabuľka a rozloženie znakov vedú k tomu, že v substituovanej depeši[30] budú mať najvyššie relatívne frekvencie cifry označujúce riadky. Toto platí dokonca aj v prípade optimálnej voľby hesla, t.j. ak najfrekventovanejšie znaky abecedy budú v prvom riadku tabuľky, čo bolo splnené aj v prípade Häyhänenovej depeše.

Aby sme aspoň čiastočne zabránili výrazným disproporciám v relatívnych frekvenciách cifier substituovanej depeše, musíme zväčšiť počet riadkov substitučnej tabuľky na 5 alebo 6. Čím viac riadkov zvolíme, tým vyrovnanejšie budú relatívne frekvencie cifier. Na druhej strane, pre jedno a dvojmiestnu zámenu viac než 6 riadkov nedáva zmysel. Preto sa tabuľka s piatimi riadkami a desiatimi stĺpcami javí ako optimálna. V takejto tabuľke máme miesto pre 46 znakov, čiže o 9 viac než je tomu pri šifre VIC. Buď by sme mohli nechať v tabuľke 9 pozícií prázdnych, alebo pridať 9 znakov. Z bezpečnostného hľadiska by asi bolo výhodnejšie nechať týchto 9 pozícií prázdnych a určiť ich miesta na základe kľúča.

[30] Pre dostatočne dlhý, zmysluplný a nie umelo zvolený otvorený text.

3.7.3 Rozloženie substitučnej tabuľky

Rozloženie znakov v substitučnej tabuľke by sa dalo výrazne zlepšiť viacerými spôsobmi:

- Po prvé: stĺpce, v ktorých sa nachádzajú špeciálne znaky, by mali byť určované na základe kľúča a nie pevne dané. V šifre VIC to tak pravdepodobne aj bolo a v článku [23] zrejme len prišlo k nedorozumeniu.
- Po druhé: špeciálne znaky **НТ**, **Н/Ц** a **ПВТ** sú úplne zbytočné a ich existenia je pre bezpečnosť šifry VIC fatálna. Tieto znaky by sa mali zo substitučnej tabuľky vypustiť a nahradiť nasledovnými opatreniami:
  - Znak pre začiatok depeše **НТ** môže byť v depeši nahradený náhodným ale legitímnym slovom, ktoré je evidentne mimo kontext depeše. Toto slovo by pre jednoduchšiu identifikáciu mohlo mať určenú dĺžku na základe kľúča, avšak takéto opatrenie vôbec nie je potrebné.
  - Prepínač módov text/cifry **Н/Ц** sa môže úplne vypustiť a cifry a čísla by sa vypisovali buď slovom, alebo pomocou preddefinovaných dvojznakových kódov. Iná možnosť, pre väčšiu substitučnú tabuľku 5×10, by bolo pridanie cifier 0 až 9 do substitučnej tabuľky. Ak by sa vyhodili spomínané tri špeciálne znaky, nebol by to žiaden problém, pretože tabuľka 5×10 má o 9 pozícií viac. Takže aj po pridaní cifier by ešte zostali dve pozície voľné.
  - Špeciálny znak „OPAKUJEM“ **ПВТ** je úplne zbytočný. Pokiaľ je to potrebné, môže sa to v texte depeše vypísať slovom. V texte Häyhänenovej depeše sa slovo „OPAKUJEM“ vyskytuje len raz a na to nie je potrebný špeciálny znak v substitučnej tabuľke.

Určite existuje ešte mnoho iných spôsobov ako by sa dala vylepšiť substitučná tabuľka, ale už iba spomenuté opatrenia by bezpečnosť sifry VIC výrazne zvýšili a zabránili by útoku, ktorý sme popisovali.

3.8 Bezpečnosť šifry VIC

Bezpečnosť šifry VIC je postavená na utajení šifrovacieho algoritmu. Akonáhle je tento šifrovací algoritmus známy, tak sa dajú depeše zašifrované šifrou VIC pomerne ľahko lúštiť. Toto je v zjavnom rozpore s druhou Kerckhoffovou zásadou, ktorá hovorí:

> *System must not be required to be secret, and it must be able to fall into the hands of the enemy without inconvenience.*[31]

Ale pokiaľ šifrovací algoritmus VIC nebol známy, bola táto šifra veľmi bezpečná. Jedná sa totiž o veľmi komplexnú šifru s vysokým stupňom difúzie. Čiže aj minimálna zmena v šifrovacom kľúči alebo malá zmena textu depeše vedie k veľkým zmenám v zašifrovanom texte. Ťažko si dokážeme predstaviť lúštenie šifry VIC len na základe jedného zašifrovaného textu. Jediné čo by sme analýzou jedného zašifrovaného textu mohli zistiť je skutočnosť, že v šifre je použitá nejaká substitúcia s obdĺžnikovou tabuľkou, ktorá má výrazne viac stĺpcov než riadkov, prípadne že je použitá jedno a dvojmiestna zámena. Toto vyplýva s disproporcií relatívnych frekvencií cifier zašifrovaného textu.

Dokázali by sme zistiť viac, ak by sme odchytili viacero depeší zašifrovaných rovnakým kľúčom? Aby sme na túto otázku mohli odpovedať, musíme si uvedomiť, že kľúč šifry VIC pozostáva z piatich častí, pričom jeho piata časť je náhodne zvolené 5-ciferné číslo a toto by sa pre každú depešu malo voliť iné. V modernej terminológii by sa táto časť kľúča nazývala *soľ*. Táto náhodná časť kľúča nám zaručí, že ak tú istú depešu zašifrujeme dvakrát po sebe a zmeníme len *soľ*'[32], tak zašifrovaný text bude takmer úplne odlišný od predchádzajúceho. Jediný spoločný parameter s predchádzajúcim zašifrovaným textom bude jeho dĺžka. Preto pri šifre VIC, pokiaľ sa používa správne, neexistuje čosi také ako dve depeše zašifrované rovnakým kľúčom.

Ak máme dva otvorené texty rovnakej dĺžky a zašifrujeme ich „rovnakým“[33] kľúčom, tak aj zašifrované texty budú mať rovnakú dĺžku. To je však všetko čo o nich vieme povedať. Kódy znakov v substitučnej tabuľke, permutácie transpozičných tabuliek a šírky transpozičných tabuliek budú s najväčšou pravdepodobnosťou rozdielne. Preto bez znalosti šifrovacieho algoritmu nevieme zo zašifrovaného textu toho veľa zistiť a to ani v prípade ak by sme tých zašifrovaných textov mali viacero.

V prípade, že by sme jeden otvorený text zašifrovali dvakrát kľúčmi, ktoré sa líšia len vo svojej poslednej, náhodnej, časti (*soli*), tak by sme mohli porovnať frekvencie znakov zašifrovaného textu a na ich základe by sme sa mohli pokúsiť uhádnuť permutáciu substitučnej tabuľky. Ale na základe len dvoch takýchto zašifrovaných textov by sa skutočne jednalo len o „hádanie“. A ak by sme aj správne uhádli permutáciu substitučnej tabuľky, tak táto nám nič nehovorí o permutáciach použitých v transpozičných tabuľkách.

Preto bez znalosti šifrovacieho algoritmu a len na základe zašifrovaného textu je útok na šifru VIC prakticky nemožný.

[31] `http://en.wikipedia.org/wiki/Kerckhoffs's_principle`

[32] Úplne postačí zmeniť jedinú cifru tohto čísla.

[33] Rovnakým okrem *soli*.

3.9 Príloha A

Slovenský preklad textu depeše z mince:

1. Gratulujeme vám k úspešnému príchodu. Týmto potvrdzujeme príjem vášho dopisu na adresu „V“ a prečítanie dopisu č. 1.

2. Na zabezpečenie vášho krytia sme dali príkaz doručiť vám 3000 dolárov. Predtým než ich investujete do akéhokoľvek podnikania, informujte nás o charaktere tohto podnikania a poraďte sa s nami.

3. Vami požadovanú receptúru na výrobu mäkkého filmu a novinky vám doručíme inou cestou, spolu s dopisom od matky.

4. Je skoro na to, aby sme vám posielali *gammy*[34]. Kratšie dopisy šifrujte a dlhšie robte po častiach[35]. Údaje o vás, ako sú miesto kde pracujete, adresa atď. nesmiete uvádzať spolu v jednej depeši. Jednotlivé časti[36] doručujte oddelene.

5. Balík sme vašej žene doručili osobne. Vaša rodina je v poriadku. Prajeme vám úspech. Súdruhovia vás pozdravujú.

č. 1 / 3. decembra

[34]Pojmom *gamma* sa v ruskej terminológii označovali OTP tabuľky.

[35]Na tomto mieste si nie som s prekladom celkom istý. Slovo **вставка** znamená v ruštine *vložka* (napr. medzera vložená do textu, obraz vložený do rámu, alebo tuha do pera), ale môže to znamenať aj *inzerát*. Zrejme sa tým myslelo delenie depeší na časti, ale nedá sa vylúčiť ani nejaká steganografická metóda, ako napr. skrývanie dopisov v dutých šróboch, preparovaných baterkách a pod., na čo bol Abel expert.

[36]Opäť je tu použité slovo **вставки**, viď predošlá poznámka pod čiarou.

3.10 Príloha B

Podrobný prepis substitúcie depeše z mince. Farebné označenia v tabuľkách 3.7 a 3.8 na nasledujúcich dvoch stranách sú nasledovné:

- **Zelenou** farbou je označený skutočný začiatok textu (symbol **HT**) depeše.
- **Žltá** farba označuje zakódované cifry.
- **Modrá** farba označuje symboly, ktoré boli vypísané slovom (**тире** = *pomlčka* a **дробь** = *lomítko*.
- **Fialová** farba označuje miesto, kde bola v depeši, zrejme omylom, zapísaná 0 (cifra) ako veľké O.
- **Šedá** farba označuje doplnené nuly na konci depeše.
- **Červená** farba označuje miesto, kde bol v článku [23] preklep (19 namiesto 69). Tento preklep sa vyskytuje len v prepise depeše v uvedenom článku. V skutočnej depeši je to uvedené správne.

П	Р	И	К	Р	Ы	Т	И	Я	М	Ы	Д	А	Л	И	У
9	69	20	63	69	61	19	20	12	23	61	25	4	13	20	29
К	А	З	А	Н	И	Е	П	Е	Р	Е	Д	А	Т	Ь	В
63	4	10	4	0	20	7	9	7	69	7	25	4	19	11	15
А	М	Т	Р	И	Т	Ы	С	Я	Ч	И	М	Е	С	Т	Н
4	23	19	69	20	19	61	5	12	66	20	23	7	5	19	0
Ы	Х	.	П	Е	Р	Е	Д	Т	Е	М	К	А	К	И	Х
61	14	67	9	7	69	7	25	19	7	23	63	4	63	20	14
В	Л	О	Ж	И	Т	Ь	В	К	А	К	О	Е	Л	И	Б
15	13	8	60	20	19	11	15	63	4	63	8	7	13	20	65
О	Д	Е	Л	О	П	О	С	О	В	Е	Т	У	И	Т	Е
8	25	7	13	8	9	8	5	8	15	7	19	29	20	19	7
С	Ь	С	Н	А	М	И	,	С	О	О	Б	Щ	И	В	Х
5	11	5	0	4	23	20	17	5	8	8	65	26	20	15	14
А	Р	А	К	Т	Е	Р	И	С	Т	И	К	У	Э	Т	О
4	69	4	63	19	7	69	20	5	19	20	63	29	21	19	8
Г	О	Д	Е	Л	А	.	Н/Ц	333	Н/Ц	.	П	О	В	А	Ш
3	8	25	7	13	4	67	18	333	18	67	9	8	15	4	16
Е	И	П	Р	О	С	Ь	Б	Е	Р	Е	Ц	Е	П	Т	У
7	20	9	69	8	5	11	65	7	69	7	24	7	9	19	29
Р	У	И	З	Г	О	Т	О	В	Л	Е	Н	И	Я	М	Я
69	29	20	10	3	8	19	8	15	13	7	0	20	12	23	12
Г	К	О	И	П	Л	Е	Н	К	И	И	Н	О	В	О	С
3	63	8	20	9	13	7	0	63	20	20	0	8	15	8	5
Т	Е	И	П	Е	Р	Е	Д	А	Д	И	М	О	Т	Д	Е
19	7	20	9	7	69	7	25	4	25	20	23	8	19	25	7
Л	Ь	Н	О	В	М	Е	С	Т	Е	С	П	И	С	Ь	М
13	11	0	8	15	23	7	5	19	7	5	9	20	5	11	23
О	М	М	А	Т	Е	Р	И	.	Н/Ц	444	Н/Ц	.	Г	А	М
8	23	23	4	19	7	69	20	67	18	444	18	67	3	4	23
М	Ы	В	Ы	С	Ы	Л	А	Т	Ь	В	А	М	Р	А	Н
23	61	15	61	5	61	13	4	19	11	15	4	23	69	4	0
О	.	К	О	Р	О	Т	К	И	Е	П	И	С	Ь	М	А
8	67	63	8	69	8	19	63	20	7	9	20	5	11	23	4

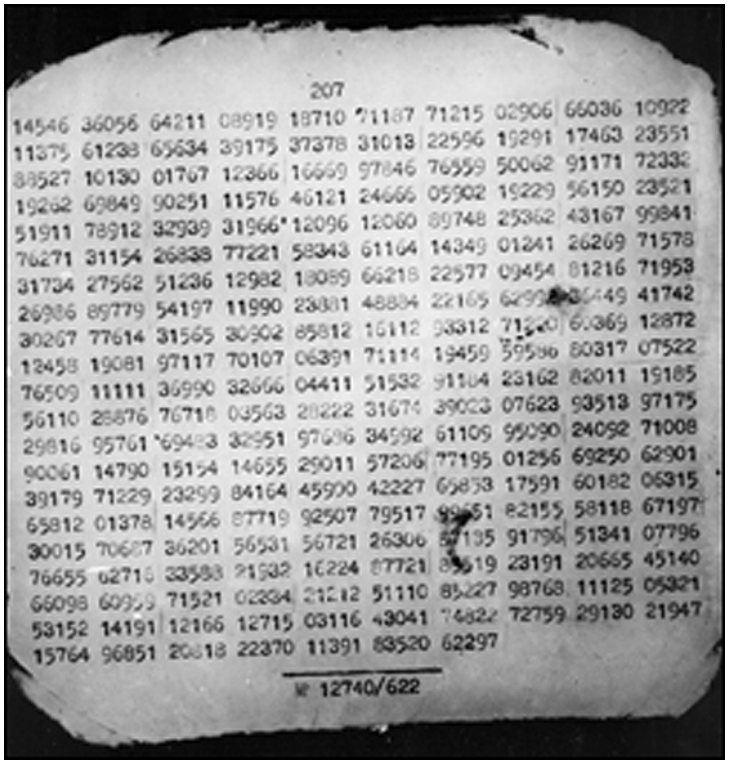
Tabuľka 3.7: Häyhänenova depeša po substitúcii, 1. časť

Ш	И	Ф	Р	У	И	Т	Е	,	А	П	О	Б	О	Л	Ь
16	20	64	69	29	20	19	7	17	4	9	8	65	8	13	11
Ш	Е	Т	И	Р	Е	Д	Е	Л	А	И	Т	Е	С	О	В
16	7	19	20	69	7	25	7	13	4	20	19	7	5	8	15
С	Т	А	В	К	А	М	И	.	В	С	Е	Д	А	Н	Н
5	19	4	15	63	4	23	20	67	15	5	7	25	4	0	0
Ы	Е	О	С	Е	Б	Е	,	М	Е	С	Т	О	Р	А	Б
61	7	8	5	7	65	7	17	23	7	5	19	8	69	4	65
О	Т	Ы	,	А	Д	Р	Е	С	И	Т	.	Д	.	В	О
8	19	61	17	4	25	69	7	5	20	19	67	25	67	15	8
Д	Н	О	И	Ш	И	Ф	Р	О	В	К	Е	П	Е	Р	Е
25	0	8	20	16	20	64	69	8	15	63	7	9	7	69	7
Д	А	В	А	Т	Ь	Н	Е	Л	Ь	З	Я	.	В	С	Т
25	4	15	4	19	11	0	7	13	11	10	12	67	15	5	19
А	В	К	И	П	Е	Р	Е	Д	А	В	А	И	Т	Е	О
4	15	63	20	9	7	69	7	25	4	15	4	20	19	7	8
Т	Д	Е	Л	Ь	Н	О	.	Н/Ц	555	Н/Ц	.	П	О	С	Ы
19	25	7	13	11	0	8	67	18	555	18	67	9	8	5	61
Л	К	У	Ж	Е	Н	Е	П	Е	Р	Е	Д	А	Л	И	Л
13	63	29	60	7	0	7	9	7	69	7	25	4	13	20	13
И	Ч	Н	О	.	С	С	Е	М	Ь	Е	И	В	С	Е	Б
20	66	0	8	67	5	5	7	23	11	7	20	15	5	7	65
Л	А	Г	О	П	О	Л	У	Ч	Н	О	.	Ж	Е	Л	А
13	4	3	8	9	8	13	29	66	0	8	67	60	7	13	4
Е	М	У	С	П	Е	Х	А	.	П	Р	И	В	Е	Т	О
7	23	29	5	9	7	14	4	67	9	69	20	15	7	19	8
Т	Т	О	В	А	Р	И	Щ	Е	И	№	Н/Ц	111	Н/Ц	Д	Р
19	19	8	15	4	69	20	26	7	20	68	18	111	18	25	69
О	Б	Ь	О	Н/Ц	333	Н/Ц	Д	Е	К	А	Б	Р	Я	НТ	Н/Ц
8	65	11	8	18	333	18	25	7	63	4	65	69	12	28	18
111	Н/Ц	.	П	О	З	Д	Р	А	В	Л	Я	Е	М	С	Б
111	18	67	9	8	10	25	69	4	15	13	12	7	23	5	65
Л	А	Г	О	П	О	Л	У	Ч	Н	Ы	М	П	Р	И	Б
13	4	3	8	9	8	13	29	66	0	61	23	9	69	20	65
Ы	Т	И	Е	М	.	П	О	Д	Т	В	Е	Р	Ж	Д	А
61	19	20	7	23	67	9	8	25	19	15	7	69	60	25	4
Е	М	П	О	Л	У	Ч	Е	Н	И	Е	В	А	Ш	Е	Г
7	23	9	8	13	29	66	7	0	20	7	15	4	16	7	3
О	П	И	С	Ь	М	А	В	А	Д	Р	Е	С	,	,	В
8	9	20	5	11	23	4	15	4	25	69	7	5	17	17	15
ПВТ	В	,	,	И	П	Р	О	Ч	Т	Е	Н	И	Е	П	И
22	15	17	17	20	9	69	8	66	19	7	0	20	7	9	20
С	Ь	М	А	№	Н/Ц	111	Н/Ц	.	Н/Ц	222	Н/Ц	.	Д	Л	Я
5	11	23	4	68	18	111	18	67	18	222	18	67	25	13	12
О	Р	Г	А	Н	И	З	А	Ц	И	И	*nuly*				
8	69	3	4	0	20	10	4	24	20	20	2 1 4				

Tabuľka 3.8: Häyhänenova depeša po substitúcii, 2. časť

3.11 Príloha C

Fotografia skutočnej depeše z mince:



Záverečné zhrnutie výsledkov dizertačnej práce

Cieľom tejto dizertačnej práce bolo hľadať a skúmať nové metódy kvantitatívnej lingvistiky, ich aplikácie v kryptoanalýze a predviesť praktickú ukážku kryptoanalýzy doteraz nerozlúštenej klasickej šifry. Zadanie práce obsahovalo tri úlohy a v súlade s nimi je práca členená na tri kapitoly:

1. Analyzujte textové štatistiky využiteľné v kryptoanalýze.
2. Pokúste sa o nájdenie lingvistických štatistík charakterizujúcich smer čítania.
3. Praktická ukážka kryptoanalýzy doteraz nerozlúštenej klasickej šifry.

Textové štatistiky používané v kryptoanalýze

Prvá kapitola je venovaná už známym štatistickým indexom a metódam využívaným pri kryptoanalýze. Okrem historického úvodu a popisu problému rozpoznávania jazyka, uvedeným na začiatku kapitoly, sú ďalej popisované niektoré štatistické indexy a vlastnosti štruktúry jazyka. Popis indexov začína tými najjednoduchšími, ako je abeceda jazyka, frekvencie znakov a n-gramov a končí pomerne komplikovanými ako sú index koincidencie a entropia. Index koncidencie je príklad indexu, ktorý sa vyvinul práve v súvislosti s kryptoanalýzou klasických ručných, polyalfabetických šifier. Tieto šifry boli veľmi dlhý čas považované za nerozlúštiteľné. Až v druhej polovici 19. stor. pruský dôstojník Kasisky objavil metódu, pomocou ktorej sa tieto šifry dali lúštiť, avšak ešte stále sa jednalo o pomerne komplikovanú záležitosť. Prelom nastal v období 1. svetovej vojny, keď viacerí kryptológovia, nezávisle na sebe, objavili štatistické vlastnosti jazyka, ktoré výrazne zjednodušili lúštenie polyalfabetických šifier s periodickým heslom. Po skončení vojny Friedman zhrnul tieto poznatky vo svojich kryptologických príručkach pre armádu ([5] až [8])

a v knihe [32] potom Kullback zaviedol index, ktorý dnes poznáme a používame pod menom *index koincidencie*. V časti 1.3.6 je stručne zhrnutý vývoj od kappa indexu po index koincidencie.

Koniec prvej kapitoly je venovaný pojmu *entropie*, pričom v celom texte týmto pojmom máme na mysli *Shannonovu entropiu*, zavedenú v súvislosti s kódovaním textu. V kryptológii sa, okrem iného, entropia využíva na určovanie, či sa v prípade neznámeho textu jedná o zmysluplný text, alebo len náhodnú postupnosť znakov.

Okrem popisu textových štatistických indexov, obsahuje prvá kapitola aj dve časti venujúce sa vlastnostiam štruktúry jazyka. Sú to hĺbka závislosti znakov v texte (1.3.4) a priemerná dĺžka slov (1.3.5). Obe tieto vlastnosti jazyka majú svoje využitie aj v kryptoanalýze. Výsledky experimentov, uvádzané v oboch spomenutých častiach, sú vlastné a boli robené na veľkých vzorkách textov. V prípade hĺbky závislosti dvoch znakov v texte stojí za povšimnutie spôsob určovania miery tejto závislosti pomocou koeficientu kontingencie (str. 20). Tento spôsob určovania miery závislosti je, podľa nám dostupných informácií, originálny. Podľa výsledkov, uvedených v tabuľkách 1.3 (str. 22) a 1.4 (str. 24), nám vyšla hĺbka závislosti znakov v intervale $\langle 15, 20\rangle$, resp. $\langle 9, 18\rangle$. Rozdiely medzi týmito údajmi vyplývajú z faktu, že výsledky v prvej tabuľke boli dosiahnuté na texte románu *20 000 míľ pod morom* v deviatich rôznych jazykoch[37] a výsledky z druhej tabuľky boli dosiahnuté na vzorkách textov v rozsahu zhruba 5 miliónov znakov v desiatich rôznych jazykoch. V literatúre, predovšetkým jazykovednej, sa bežne uvádza maximálna hĺbka závislosti znakov v rozsahu 20 až 50. Napríklad v Jaglomovej knihe [20] (str. 251) sa uvádza, že hĺbka závislosti znakov hovorových jazykov sa pohybuje okolo hodnoty 30. Pritom údaje o hĺbke závislosti znakov kolíšu podľa zdroja a metodika zisťovania týchto hodnôt v prevažnej väčšine prípadov nie je uvedená. V kryptologickej literatúre sa dajú nájsť aj nižsie hodnoty pre hĺbku závislosti znakov, pohybujúce sa v rozsahu okolo 6 až 8. Toto môže súvisieť so spôsobom stanovenia hranice, kedy ešte dva znaky textu sú závislé a kedy už sú nezávislé. My sme si stanovili hranicu koeficientu kontingencie na úrovni 0.005 % a najmenšiu vzdialenosť, pri ktorej koeficient kontingencie klesne pod túto hranicu definujeme ako hĺbku závislosti znakov. Samozrejme, že ak by sme si túto hranicu stanovili vyššiu, tak hĺbka závislosti znakov by bola menšia. V našich testoch na románe *20 000 míľ pod morom* by sme hĺbku závislosti znakov 6 dosiahli pre hranicu 0.12 %.

V prvej kapitole sú ešte zaujímavé výsledky priemerných dĺžok slov v rôznych jazykoch, uvedené v tabuľkách 1.5 (str. 25) a 1.6 (str. 26). Na určovaní priemernej dĺžky slov nie je nič mimoriadneho a je to dobre známa štatistická

[37] Počet znakov týchto vzoriek sa pohyboval v rozsahu približne 400 000 – 800 000 znakov.

vlastnosť jazyka. Avšak často sa traduje mýtus o tom, že nemčina má výrazne väčšiu dĺžku slov než napríklad angličtina a že slovanské jazyky majú väčšiu priemernú dĺžku slov než napr. románske alebo anglosaské. Ako je zrejmé z uvedených tabuliek, jedná sa skutočne len o mýtus. Je síce pravdou, že angličtina patrí medzi jazyky s najkratšou priemernou dĺžkou slov a nemčina zasa medzi jazyky s najväčšou priemernou dĺžkou slov. Ale takmer pri všetkých testovaných jazykoch nám vyšla priemerná dĺžka slov v rozsahu 5 až 6 znakov a rozdiely medzi jazykmi sú až za desatinnou čiarkou. Takže napríklad rozdiel medzi angličtinou a nemčinou je približne 1 znak. Okrem toho nám vyšli priemerné dĺžky slov slovanských jazykov (slovenčina, čeština, poľština) približne rovnaké ako priemerné dĺžky slov románskych jazykov. Výnimkami sú len poľština, pri ktorej nám, na vzorke zhruba 5 miliónov znakov, vyšla najmenšia priemerná dĺžka slov a latinčina, pri ktorej nám priemerná dĺžka slov vyšla najväčšia spomedzi všetkých testovaných jazykov.

Určovanie smeru čítania textu

Všetky výsledky uvádzané v druhej kapitole sú vlastné a podľa našich vedomostí sa týmto problémom doposiaľ nikto nezaoberal. Našou úlohou bolo nájsť metódu, na základe ktorej by bolo možné určiť správny smer čítania neznámeho textu. To znamená, že máme k dispozícii len skúmaný text a nemôžme využiť žiadne štatistické vlastnosti jazyka, ako sú napr. štatistické indexy a vlastnosti uvádzané v prvej kapitole. V skutočnosti ani nevieme o aký jazyk sa v prípade daného textu jedná. Tento problém má svoj pôvod v skúmaní Rohonczi kódexu. Rohonczi kódex je historický dokument v rozsahu približne 400 strán, ktorý obsahuje text, o ktorom sa nič nevie. Nepoznáme jazyk, nevieme či a akým spôsobom je text zašifrovaný a dokonca ani nevieme, či sa jedná o zmysluplný text, alebo len nejaký historický podvrh. Takže je to ukážkový príklad neznámeho textu spomínaného na začiatku tohto odstavca. Iným príkladom takéhoto textu je Voynichov manuskript.

Predpokladajme, že sa, v prípade napr. Rohonczi kódexu, jedná o zmysluplný text. Ak by sme sa chceli pokúsiť o jeho lúštenie, potrebujeme predovšetkým vedieť, ktorým smerom je tento text písaný, aby sme mali správnu nádväznosť riadkov a mohli text ďalej analyzovať. Z uvedeného vyplýva potreba nájdenia metódy určovania správneho smeru čítania textu, bez využitia štruktúry a/alebo akýchkoľvek vlastností použitého jazyka.

Spôsob, akým sme k spomenutému problému pristúpili bol ten, že sme na vzorkách otvorených textov, rôznych jazykov, experimentálne hľadali indexy, ktorých hodnoty by sa líšili v závislosti od smeru čítania textu. Keďže sme nemohli využívať žiadne štruktúry a vlastnosti daného jazyka, sústre-

dili sme sa len na skúmané vzorky textu. Pretože nepoznáme správny smer čítania textu, nemôžme riadky spájať a získavať tak dlhšie textové úseky. Informácie o štruktúre použitého jazyka sme teda získavali len z jednotlivých riadkov textu. Hľadali sme opakujúce sa n-gramy na riadkoch a tieto sme nazvali *legitímne reťazce*. Následne sme skúmali zlomy riadkov v predpokladanom smere čítania textu. Zisťovali sme to, či sa legitímne reťazce nachádzajú aj na zlomoch riadkov. Postavili sme si pracovnú hypotézu (str. 52):

> **Hypotéza:** *Pri čítaní textu v nesprávnom smere, bude na zlomoch riadkov „menej“*[38] *reťazcov, ktoré sa vyskytujú aj na riadkoch textu a sú teda pre príslušný text legitímne, než pri čítaní textu v správnom smere.*

V časti 2.3 sme si empiricky zadefinovali 11 indexov, ktoré nám vyjadrovali „množstvo“ legitímnych reťazcov na zlomoch riadkov v predpokladanom smere čítania textu. Potom sme na vzorkách otvoreného textu testovali vytvorené indexy. Skúmali sme, či sa hodnoty týchto indexov líšia v závislosti od smeru čítania textu a či sme pomocou nich schopní určiť správny smer čítania textu. Zistili sme, že každý z týchto indexov pri jednom vykonanom teste môže, pri vhodne zvolenom texte, zlyhať[39]. Avšak pri mnohonásobnom opakovaní testov a spriemerovaní ich výsledkov, navrhnuté indexy správne určovali smer čítania textu s vysokou pravdepodobnosťou. Dokonca sa nám aj podarilo detekovať niektoré okolnosti spôsobujúce zlyhanie indexov. Sú to napríklad slová a výrazy s vysokou relatívnou početnosťou v texte, alebo text rozdelený na riadky tak, že tieto sa končia vždy celým slovom. Na základe našich testov a ich výsledkov sme v závere (časť 2.8) navrhli metódy určovania správneho smeru čítania textu. Jedna z metód je postavená na vyhodnocovaní výsledkov všetkých jedenástich indexov a určenia percentuálnej úspešnosti čítania textu v jednom alebo druhom smere. Druhá metóda pracuje len s výberom piatich testovacích indexov a správny smer čítania textu určuje väčšinovým spôsobom.
Výsledky uvedené v druhej kapitole sa pripravujú na publikáciu.
Ďalšie navrhované smery výskumu v tejto oblasti sú:

- Navrhovanie a testovanie nových indexov, t.j. spôsobu bonifikácie legitímnych reťazcov na zlomoch riadkov, ktoré by lepšie dokázali odlíšiť smery čítania textu.
- Normalizácia navrhnutých indexov a skúmanie ich štatistických vlastností.

[38] Spôsob určovania tejto miery je uvedený v časti 2.3 (str. 53).
[39] V zmysle, že v súlade s postavenou hypotézou, nesprávne určí smer čítania textu.

Lúštenie sovietskej šifry VIC

Tretia kapitola je popisom úspešného útoku na sovietsku šifru VIC. Analýzou a lúštením tejto šifry sa, podľa dostupných informácií, doteraz nikto nezaoberal. V literatúre je šifra VIC označovaná za veľmi komplexnú a „neprelomiteľnú“ šifru ([22], str. 670–671). Existuje o nej viacero publikácií. Tieto sa však venujú len popisu šifry VIC a príbehu sovietskeho agenta Häyhänena, ktorý túto šifru používal pri kontakte so známym sovietskym agentom Abelom.

V prvej časti tretej kapitoly (časť 3.2) je uvedený podrobný popis šifry VIC. Tento popis je kompletne prevzatý z Kahnovho článku [23] a je robený na konkrétnej depeši agenta Häyhänena, ktorú našla a nebola schopná rozlúštiť americká FBI. Podrobný popis šifry VIC v texte uvádzame po prvé preto, aby čitateľ získal predstavu o komplexite tejto šifry a po druhé preto, že v časti o lúštení sa veľakrát odvolávame na mnohé veci obsiahnuté v popise šifry. VIC bola šifra typu STT, čiže pozostávala zo substitúcie a dvoch transpozícií. Substitúcia bola jedno a dvojmiestna zámena a je popísaná v časti 3.2.1. Po nej nasledovala prvá transpozícia, ktorá je popísaná v časti 3.2.2. Prvá transpozícia bola obyčajná tabuľková transpozícia s obdĺžnikovou tabuľkou. Po prvej transpozícii nasledovala druhá transpozícia, ktorej popis je v časti 3.2.3. Táto bola tabuľková-obrazcová transpozícia, podobná československej šifre „Zubatka“ používanej počas 2. svetovej vojny.

Časť 3.4 obsahuje už vlastné výsledky a je venovaná analýze a lúšteniu šifry VIC. Výsledky uvedené v tejto časti boli publikované v článku [26][40]. Ukázali sme, že napriek svojej komplexite, šifra VIC je jednoducho lúštiteľná, pokiaľ poznáme šifrovací algoritmus. To znamená, že VIC porušuje druhú Kerckhoffovu zásadu hovoriacu o tom, že bezpečnosť šifry musí byť postavená výlučne na šifrovacom kľúči a nie na utajení šifrovacieho algoritmu. Pri porušení tejto zásady nedokáže samotná komplexita šifry zabrániť jej úspešnému lúšteniu.
V texte sme detekovali niekoľko slabín šifrovacieho algoritmu VIC-u. Sú to:

1. Generovanie permutácií zo šifrovacieho kľúča.
2. Nevhodne vybraný typ substitúcie.
3. Zostavovanie substitučnej tabuľky.

Všetky uvedené slabiny sme využili na sériu testov, pomocou ktorých sme text zašifrovaný VIC-om previedli na sadu textov zašifrovaných jednoduchou

[40] Online v roku 2015 a tlačou v roku 2016.

zámenou. Keďže lúštenie jednoduchej zámeny je triviálna, automatizovateľná a veľmi rýchlo realizovateľná záležitosť, je týmto šifra VIC rozlúštená.

V texte popísaný útok bol aj implementovaný a výsledky sú uvedené v časti 3.5. Lúštenie bolo prakticky otestované na Häyhänenovej depeši. Časová náročnosť lúštenia je veľmi nízka. Na bežnom notebooku sa jedná o hodiny pri použití len jedného jadra jedného procesora. Vzhľadom na to, že aplikované testy sú nezávislé, sa celý postup lúštenia dá veľmi jednoducho paralelizovať rozdelením vstupnej sady dát. V dôsledku toho sa dá lúštenie výrazne urýchliť a je možné realizovať ho v podstate ľubovoľnom, vopred stanovenom, čase, jednoduchým zvýšením výpočtových zdrojov.

V časti 3.7 navrhujeme spôsoby, ktorými by sa dalo predísť nami navrhnutému útoku na šifru VIC. Tieto úvahy sú ale čisto akademického charakteru, pretože po prvé, klasické ručné šifry sú dnes už neaktuálne a po druhé, ani nami navrhnuté protiopatrenia neriešia hlavný problém šifry VIC, ktorým je porušenie druhej Kerckhoffovej zásady.

Literatúra

[1] Anděl, J.: Statistické metody, *MatfyzPress, Praha 2007*

[2] Bauer, F. L.: Entzifferte Geheimnisse, *Springer, 2000*

[3] Cover, T. M., King, R.: A convergent gambling estimate of the entropy of English, *IEEE Trans. on Inform. Theory, 24 (4), pp. 413–421, 1978*

[4] Čagala, R.: Kryptoanalýza šifry VIC,
bakalárska práca, FEI STU, školiteľ: Doc. Ing. Pavol Zajac, PhD, 2009

[5] Friedman, W. F.: Military Cryptanalysis - Part I.,
Washington/Aegean Park Press, 1938/1984
`http://www.nsa.gov/public_info/declass/military_cryptanalysis.shtml`

[6] Friedman, W. F.: Military Cryptanalysis - Part II.,
Washington/Aegean Park Press, 1938/1984
`http://www.nsa.gov/public_info/declass/military_cryptanalysis.shtml`

[7] Friedman, W. F.: Military Cryptanalysis - Part III.,
Washington/Aegean Park Press, 1938/1984
`http://www.nsa.gov/public_info/declass/military_cryptanalysis.shtml`

[8] Friedman, W. F.: Military Cryptanalysis - Part IV.,
Washington/Aegean Park Press, 1942/1984
`http://www.nsa.gov/public_info/declass/military_cryptanalysis.shtml`

[9] Friedman, W. F.: The Index of Coincidence and its Applications in Cryptanalysis, *Washington/Aegean Park Press, ????/1987*

[10] Gaines, H. F.: Cryptanalysis, a study of ciphers and their solution,
Dover Publication Inc., New York, 1939

[11] Ganesan, R., Sherman, A.: Statistical Techniques for Language Recognition: An Introduction and Guide for Cryptanalysis,
`http://web.cecs.pdx.edu/ bart/decrypter/`*, 1993*

[12] Grošek, O., Vojvoda, M., Zajac, P.: Klasické šifry,
STU v Bratislave, 2007

[13] Grošek, O., Zajac, P.: Automated Cryptanalysis,
Encyclopedia of Artificial Intelligence, 2008, pp. 179–185

[14] Grošek, O., Zajac, P.: Automated Cryptanalysis of Classical Ciphers,
Encyclopedia of Artificial Intelligence, 2008, pp. 186–191

[15] Grošek, O.: On a Reconstruction of a Markov Chain,
Journal of Combinatorics, Vol. 20, Nos. 1–4, 1995, pp. 85–93

[16] Grošek, O.: Entropia na algebraických štruktúrach,
Kandidátska dizertačná práca, Bratislava 1977

[17] Grošek, O.: Энтропия на алгебраических структурах,
Mathematica Slovaca 29, No. 4, 1979, pp. 411-424

[18] Grošek, O., Vojvoda, M., Zajac, P., Zanechal, M.: Základy kryptografie,
STU v Bratislave, 2006

[19] Guerrero Fabio G.: On the Entropy of Written Spanish,
IEEE Transactions on Information Theory

[20] Яглом А. М., Яглом И. М.: Вероятность и информация,
Издательство »НАУКА«, Москва 1973

[21] Janeček, J.: Odhalená tajemství šifrovacích klíčů minulosti,
Naše vojsko, 1994

[22] Kahn, D.: The Codebreakers, *Scribner, 1996*

[23] Kahn David: Number One from Moscow,
CIA Historical Review Program – declassified 1993
https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol5no4/pdf/v05i4a09p.pdf
https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol5no4/html/v05i4a09p_0001.htm

[24] Kahn Jeffrey: The Case of Colonel Abel,
Journal of National Security, Law & Policy, June 2010

[25] Kalina, M., Bacigál, T., Schiesslová, A.: Základy pravdepodobnosti a matematickej štatistiky, *Vydavateľstvo STU, Bratislava 2010*

[26] Kollár, J.: Soviet VIC Cipher: No Respector of Kerckhoff's Principles,
Cryptologia 40, 2016, pp. 33–48

[27] Kollár, J.: Sovietska šifra VIC, *Crypto-World 9–10, 2013, pp. 2–16*

[28] Kollár, J.: Reino Häyhänen – sovietsky špión, *Crypto-World 7–8, 2013, pp. 2–9*

[29] Konheim, A.: Cryptography: A Primer, *John Wiley & Sons, 1981*

[30] Kontoyiannis, I.: The Complexity and Entropy of Literary Styles, *NSF Technical Report No. 97, Stanford University, 1996/97*

[31] Kubáček, Ľ.: Confidence Limits for Proportions of Linguistic Entities, *Journal of Quantitative Linguistics, Vol. 1, No. 1, 1994, pp. 56–61*

[32] Kullback, S.: Statistical Methods In Cryptanalysis, *Aegean Park Press, ????/1976*

[33] Lennert, J.: Heuristic Language Analysis: Techniques and Applications, `http://web.cecs.pdx.edu/ bart/decrypter/`, *2001*

[34] Mierka, Z.: Šifra VIC a jej softwérová realizácia, *bakalárska práca, FEI STU, školiteľ: Prof. RNDr. Otokar Grošek, PhD, 2007*

[35] Oravec, J., Laca, V.: Príručka slovenského pravopisu, *SPN Bratislava, 1973*

[36] Ospanova Bikesh Revovna: Calculating Information Entropy of Language Texts, *World Applied Sciences Journal 22 2013, pp. 41–45*

[37] Rocafort W. W.: Colonel Abel's Assistant *CIA Studies in Intelligence, Vol. 3, Issue: Fall, 1959* – declassified 1994

[38] Shannon, C. E.: A Mathematical Theory of Communication, *Bell System Technical Journal Vol. 27, No. 3, [379–423], 1948*

[39] Shannon, C. E.: Prediction and Entropy of Printed English, *Bell System Technical Journal Vol. 30, [50–64], 1951*

[40] Shannon, C. E.: Communication Theory of Secrecy Systems, *Bell System Technical Journal, 1948*

[41] Solomon, M.: Algebraické modely v lingvistice, *Academia, Praha 1969*

[42] Torres, S., Gelbukh, A.: Comparing Similarity Measures for Original WSD Lesk Algorithm, *Research in Computing Science 43, 2009, pp. 155–166*

[43] Vajda, I.: Teória informácie a štatistického rozhodovania, *Alfa, edícia Epsilon, Bratislava 1982*

[44] Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S.: Úvod do analýzy textov, *Veda, 2003*

[45] Zborník prác: Teorie informace a jazykověda, *Academia, Praha 1964*

[46] Zvára, K., Štěpán, J.: Pravděpodobnost a matematická statistika, *MatfyzPress, Praha 2006*