SLOVENSKÁ TECHNICKÁ
UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY
A INFORMATIKY

Jozef Kollár

Autoreferát dizertačnej práce

Štatistická analýza textov pre potreby kryptoanalýzy

na získanie akademického titulu philosophiæ doctor (PhD.)
v odbore doktorandského štúdia

9.2.9 Aplikovaná informatika

Bratislava 2016

Dizertačná práca bola vypracovaná v externej forme doktorandského štúdia na Ústave Informatiky a Matematiky, Fakulty Elektrotechniky a Informatiky, Slovenskej Technickej Univerzity v Bratislave

Predkladateľ: Mgr. Jozef Kollár
Katedra Matematiky a Deskriptívnej Geometrie
SvF STU, Bratislava

Školiteľ: prof. RNDr. Otokar Grošek, PhD.
Ústav Informatiky a Matematiky
FEI STU, Bratislava

Oponenti: prof. RNDr. Gejza Wimmer, DrSc.
Katedra matematiky
Fakulta prírodných vied
Univerzita Mateja Bela, Banská Bystrica

doc. RNDr. Karol Nemoga, CSc.
riaditeľ Matematického ústavu SAV
Bratislava

Obhajoba dizertačnej práce sa koná ..... o ..... hod. pred komisiou pre obhajobu dizertačnej práce v odbore doktorandského štúdia vymenovanou predsedom odborovej komisie .....

Aplikovaná informatika – 9.2.9. aplikovaná informatika

na Ústave Informatiky a Matematiky, Fakulty Elektrotechniky a Informatiky, Slovenskej Technickej Univerzity v Bratislave
Ilkovičova 3, 812 19 Bratislava

prof. Dr. Ing. Miloš Oravec
dekan Fakulty Elektrotechniky a Informatiky
STU v Bratislave

Obsah

Úvod

Cieľom 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.

Súčasný stav problematiky

Vlastné výsledky sú obsiahnuté v 2. a 3. kapitole práce. Jedná sa o:

- Nájdenie metódy na určenie správneho smeru čítania textu.
- Lúštenie sovietskej šifry VIC.

Ani jeden z týchto dvoch problémov nebol doteraz riešený.

V prípade sovietskej šifry VIC existujú publikácie popisujúce túto šifru. Tieto sú uvedené v použitej literatúre a citované v texte. Väčšina z nich je však venovaná príbehu agenta Häyhänena a šifra VIC, ktorú používal, sa v nich spomína len okrajovo. Kryptoanalýzou alebo lúštením šifry VIC sa žiadna, nám známa, publikácia nezaoberá.

O probléme určovania správneho smeru čítania neznámeho textu doposiaľ, podľa našich vedomostí, 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 sa zaoberáme textom 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.

Popis kapitol práce

1. kapitola

Prvá kapitola práce je venovaná známym štatistickým indexom a metódam využívaným pri kryptoanalýze. Okrem historického úvodu a popisu problému rozpoznávania jazyka, je tu popis niektorých štatistických indexov a vlastností 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 vojne 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*. Vývoj od kappa indexu po index koincidencie je stručne popísaný práve v prvej kapitole.

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.

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

Okrem popisu textových štatistických indexov, obsahuje prvá kapitola aj dve časti venujúce sa vlastnostiam štruktúry jazyka. Sú to hĺbka závislosti znakov v texte a priemerná dĺžka slov. 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. Tento spôsob určovania miery závislosti je, podľa nám dostup-

ných informácií, originálny. Podľa výsledkov, uvedených v tabuľkách 1 a 2, 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

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 2: Hĺbka závislosti na vzorkách zhruba 5 miliónov znakov

v prvej tabuľke boli dosiahnuté na texte románu *20 000 míľ pod morom* v deviatich rôznych jazykoch[1] 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

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

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 %.

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 3: Priemerné dĺžky slov v rôznych jazykoch #1

V prvej kapitole sú ešte zaujímavé výsledky priemerných dĺžok slov v rôznych jazykoch, uvedené v tabuľkách 3 a 4. Na určovaní priemernej dĺžky slov nie je nič mimoriadneho a je to dobre známa štatistická 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.

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 4: Priemerné dĺžky slov v rôznych jazykoch #2

2. kapitola

Druhá kapitola práce sa venuje hľadaniu metódy určovania správneho smeru čítania neznámeho textu. Ako už bolo spomenuté, všetky uvádzané výsledky 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, sústredili 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:

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

Následne 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[2]. 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

[2]V zmysle, že v súlade s postavenou hypotézou, nesprávne určí smer čítania textu.

našich testov a ich výsledkov sme v závere 2. kapitoly 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.

3. kapitola

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.

	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 5: Ukážka substitučnej tabuľky VIC agenta Häyhänena

V prvej časti tretej kapitoly 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 odvolávamé 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. Po nej nasledovala prvá transpozícia. Bola to obyčajná tabuľková transpozícia s obdĺžnikovou tabuľkou. Po prvej transpozícii nasledovala druhá transpozícia. Táto bola tabuľková-obrazcová transpozícia, podobná československej šifre „Zubatka“ používanej počas 2. svetovej vojny.

Za popisom šifry VIC nasleduje časť, ktorá už obsahuje vlastné výsledky a je venovaná analýze a lúšteniu šifry VIC. Tieto výsledky už boli publikované[3] a celá druhá časť 3. kapitoly je napísaná na báze článku [26]. 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 zámenou. Keďže lúštenie jednoduchej zámeny je triviálna,

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

automatizovateľná a veľmi rýchlo realizovateľná záležitosť, je týmto šifra VIC rozlúštená.

V texte popísaný útok bol aj implementovaný a výsledky sú uvedené v závere druhej časti 3. kapitoly. 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 závere 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.

Fotografia Häyhänenovej depeše z mince:

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*

Publikačná a vedecká činnosť autora

Práce v kategórii A

- Emmanuel Faure, Thierry Savy, Barbara Rizzi1, Camilo Melani1, Oľga Stašová, Dimitri Fabrèges, Róbert Špir, Mark Hammons, Róbert Čunderlík, Gaëlle Recher, Benoît Lombardot, Louise Duloquin, Ingrid Colin, Jozef Kollár, Sophie Desnoulez, Pierre Affaticati, Benoît Maury, Adeline Boyreau, Jean-Yves Nief, Pascal Calvat, Philippe Vernier, Monique Frain, Georges Lutfalla, Yannick Kergosien, Pierre Suret, Mariana Remešíková, René Doursat, Alessandro Sarti, Karol Mikula, Nadine Peyriéras & Paul Bourgine, **A workflow to process 3D+time microscopy images of developing organisms and reconstruct their cell lineage**, *Nature Communications 7, Article number: 8674*

Práce v kategórii B

- Kollár Jozef, **Soviet VIC Cipher: No Respector of Kerckhoff's Principles**, *Cryptologia 40, 2015/2016, s. 33–48*

Práce v zahraničných nekarentovaných časopisoch

1. Kollár Jozef, **Tajné písmo Martina Kukučína**, *Crypto-World, ISSN 1801-2140, Roč. 12, č. 1 (2010), s. 12–16*

2. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Úvod k seriálu**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 1 (2011), s. 2*

3. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 1.: Šifra TTS**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 1 (2011), s. 3–11*

4. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 2.: Šifra „Rímska dva“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 2 (2011), s. 2–11*

5. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 3.: Šifra „Rímska osem“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 3 (2011), s. 2–12*

6. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 4.: Šifra „Rímska deväť“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 4 (2011), s. 2–16*

7. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 5.: Šifra „Rímska desať“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 5 (2011), s. 2–13*

8. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 6.: Šifra „Rímska trinásť“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 6 (2011), s. 2–11*

9. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 7.: Šifra „Eva“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 7-8 (2011), s. 2–9*

10. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny Diel 8.: Šifra „Marta“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č. 9 (2011), s. 2–8*

11. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny. Diel 9., Šifra „Růžena“**, *Crypto-World, ISSN 1801-2140, Roč. 13, č.10 (2011), s. 2–12*

12. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny. Diel 10., Šifra „Utility“**, *Crypto-World, ISSN 1801-2140, Roč. 14, č. 2 (2012), s. 2–10*

13. Kollár Jozef, **ČS Šifry z obdobia 2. svetovej vojny. Diel 11., Šifra „Palacký“**, *Crypto-World, ISSN 1801-2140, Roč. 14, č. 3-4 (2012), s. 2–12*

14. Kollár Jozef, **Má zmysel používať autokľúč?**, *Crypto-World, ISSN 1801-2140, Roč. 14, č. 3-4 (2012), s. 12–17*

15. Kollár Jozef, **Andreas Figl – rakúsky dôstojník a kryptológ**, *Crypto-World, ISSN 1801-2140, Roč. 15, č. 3-4 (2013), s. 15–23*

16. Kollár Jozef, **Häyhänen – sovietsky špión**, *Crypto-World, ISSN 1801-2140, Roč. 15, č. 7-8 (2013), s. 2–9*

17. Kollár Jozef, **Sovietska šifra VIC**, *Crypto-World, ISSN 1801-2140, Roč. 15, č. 9-10 (2013), s. 2–16*

Prezentácie na zahraničných vedeckých konferenciách

1. Kollár Jozef, **Czechoslovak WWII Ciphers – pozvaná prednáška**, *Mikulášská kryptobesídka 2011: Sborník příspěvků, Praha 1.–2. decembra 2011, Praha: Trusted Network Solutions, 2011, s. 55–63*

Prezentácie na domácich konferenciách

1. Kollár Jozef, **O písme Martina Kukučína**, *MAGIA 2009: Mathematics, Geometry and their Applications, Conference Proceedings, SvF STU v Bratislave, 2009, s. 149–152, ISBN 978-80-227-3207-9*

2. Kollár Jozef, **Ako správne zvoliť heslo**, *MAGIA 2007: Mathematics, Geometry and their Applications, Conference Proceedings, SvF STU v Bratislave, 2007, s. 111–116*

3. Kollár Jozef, **Centralizovaná správa hesiel**, *MAGIA 2007: Mathematics, Geometry and their Applications, Conference Proceedings, SvF STU v Bratislave, 2007, s. 117–125*

4. Kollár Jozef, **Klasické transpozičné šifry**, *MAGIA 2006: Mathematics, Geometry and their Applications, Conference Proceedings, SvF STU, Bratislava 2006, s. 81–89, ISBN 80-227-2583-8*

Skriptá a učebné texty

1. Kollár Jozef, **Matematika I. – Zbierka úloh ku cvičeniam**, *Bratislava: SvF STU, 2014, 167 s., ISBN 978-80-227-4244-3*

2. Kollár Jozef, **Riešené matematické úlohy z prijímacích skúšok FM UK 2008–2012**, *Bratislava: Univerzita Komenského v Bratislave, 2013, 383 s.*

3. Kollár Jozef, **Riešené matematické úlohy z prijímacích skúšok FM UK 2008–2011**, *Bratislava: Univerzita Komenského v Bratislave, 2012, 313 s., ISBN 978-80-223-3200-2*

4. Kollár Jozef, **Riešené matematické úlohy z prijímacích skúšok 2008–2009**, *Bratislava: Univerzita Komenského v Bratislave, 2010, 153 s., ISBN 978-80-223-2800-5*

5. Kollár Jozef, **Riešené matematické úlohy z prijímacích skúšok 2002–2007**, *Bratislava: Univerzita Komenského v Bratislave, 2008, 351 s., ISBN 978-80-223-2455-7*

6. Kollár Jozef, **Zbierka úloh z matematiky z prijímacích skúšok**, *Bratislava: FM UK v Bratislave, 2007., 264 s.*

7. Kollár Jozef, **Zbierka testov z matematiky na prijímacie skúšky**, *Bratislava: Univerzita Komenského v Bratislave, 2006*

Aktívna prezentácia výsledkov a prednášky

1. **ČS šifry z obdobia 2. svetovej vojny – 1.**
   Prednáška – Klasické šifry, FEI STU, Bratislava
   30. marec 2011

2. **ČS šifry z obdobia 2. svetovej vojny – 2.**
   Prednáška – Klasické šifry, FEI STU, Bratislava
   6. apríl 2011

3. **Czechoslovak WWII Ciphers – pozvaná prednáška**
   Konferencia MKB 2011, Praha
   2. december 2011

4. **Štatistické spracovanie textu pre kryptoanalýzu**
   Crypto-seminár, FEI STU, Bratislava
   11. apríl 2012

5. **Štatistické metódy spracovania textu – 1.**
   Štatistický seminár, SvF STU, Bratislava
   12. apríl 2012

6. **Štatistické metódy spracovania textu – 2.**
   Štatistický seminár, SvF STU, Bratislava
   19. apríl 2012

7. **Czechoslovak WWII Ciphers**
   Crypto History Experts Meeting, Heusenstamm
   7. jún 2012

8. **Matematika – tajná zbraň spojencov**
   FMFI UK, Bratislava, (pozvaná prednáška)
   13. september 2012

Recenzie

1. **Hybrid Cipher System**, *Computing and Informatics, 30. november 2009*

2. **On the Suitability of the Internet Multimedia ...**, *Kybernetika, 15. september 2011*

3. **Using Poly-Dragon Cryptosystem in a Pseudorandom...**, *Tatra Mountains, 21. október 2012*

Práca v riešiteľských kolektívoch

1.	**Označenie:**	VEGA 1/0489/08
	Názov:	Topologické metódy štúdia diskrétnych štruktúr a ich grúp symetrií
	Koordinátor:	Prof. RNDr. Jozef Širáň, DrSc.
2.	**Označenie:**	VEGA 1/0871/11
	Názov:	Algebraické a topologické metódy v štúdiu kombinatorických štruktúr s vysokým stupňom súmernosti
	Koordinátor:	Prof. RNDr. Jozef Širáň, DrSc.
3.	**Označenie:**	VEGA 1/3321/06
	Názov:	Moderné metódy matematického a počítačového modelovania v inžinierskych aplikáciách
	Koordinátor:	Prof. RNDr. Karol Mikula, DrSc.
4.	**Označenie:**	VEGA 1/0269/09
	Názov:	Vývoj efektívnych a spoľahlivých numerických metód pre inžinierske aplikácie
	Koordinátor:	Prof. RNDr. Karol Mikula, DrSc.
5.	**Označenie:**	Európsky projekt 6. rámcového programu
	Názov:	EMBRYOMICS – Reconstructing in space and time the cell lineage tree (2005–2008)
	Kontraktor:	Prof. RNDr. Karol Mikula, DrSc.
	Koordinátor:	Dr. Nadine Peyrieras, CNRS Paris

6.

Označenie:	Európsky projekt 6. rámcového programu
Názov:	BioEmergencies: „In what“ and „how much“ are individuals similar and different? Towards the measurement of the individual susceptibility to diseases or response to treatments (2005–2009)
Kontraktor:	Prof. RNDr. Karol Mikula, DrSc.
Koordinátor:	Prof. P.Bourgine, Ecole Polytechnique, Paris

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, the section about depth of letter dependency, and the section 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.