UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI

Şcoala Doctorală de Electronică, Telecomunicaţii şi Tehnologia Informaţiei

Decizie nr. 1046 **din** 10-07-2023

REZUMAT TEZĂ DE DOCTORAT

Ing. Alexandru DINU

MODELE ŞI METODE STATISTICE UTILIZATE ÎN STUDIUL ŞTIINŢELOR VIEŢII

STATISTICAL MODELS AND METHODS USED IN LIFE SCIENCES

COMISIA DE DOCTORAT

Prof. Dr. Ing. Gheorghe BREZEANU Univ. Politehnica din Bucureşti	Preşedinte
Prof. Dr. Ing. Adriana VLAD Univ. Politehnica din Bucureşti	Conducător de doctorat
Prof. Dr. Ing. Corneliu BURILEANU Univ. Politehnica din Bucureşti	Referent
Maître de conférences Mihai MITREA *Télécom* Sud-Paris	Referent
Prof. Dr. Ing. Victor Adrian GRIGORAŞ Univ. Tehnică "Gheorghe Asachi" Iaşi	Referent

BUCUREŞTI 2023

Cuprins

Capitolul 1

Introducere

Prezenta lucrare de doctorat abordează un domeniu de studiu fecund si insuficient explorat: științele vieții văzute și analizate dintr-o perspectivă statistică interdisciplinară. Prin "științe ale vieții" se înțelege orice domeniu de studiu, cercetare și cunoaștere care are legătura cu viața. În această categorie sunt incluse biologia, zoologia, medicina, antropologia, psihologia, sociologia și multe altele, mai noi sau combinații ale celor menționate mai sus [1–3]. Așa cum se poate vedea, domeniul științelor vieții poate fi foarte vast [4–6].

După cum se va putea lesne observa în capitolele următoare, subiectele abordate în această lucrare de doctorat au în comun abordarea statistică și matematică, însă în același timp diferă foarte mult, variind de la domeniul psihologiei combinat cu elemente de teoria informației, până la studiul sistemelor dinamice haotice dintr-o perspectivă ubicuă. Sunt de părere că progresul și evoluția noastră ca specie se vor produce doar considerând și utilizând beneficiile aduse de idei cum ar fi interdisciplinaritatea, multidisciplinaritatea sau transdisciplinaritatea, după cum tot mai mulți autori au început să observe recent [7, 8]. Este uimitor cum există atât de multe similitudini între soluțiile găsite de cercetători din diverse domenii ale cunoașterii aparent necorelate [9].

În timpul studiilor de licență de la Universitatea Politehnica București, sub îndrumarea doamnei profesor Adriana Vlad, am devenit pasionat de domeniul criptografiei și semnalelor haotice, așa că nu e deloc surprinzător faptul că rezultatele cele mai importante din lucrarea mea de licență au fost extinse și publicate în Buletinul UPB [10].

Odată cu începerea studiilor doctorale, domeniul de cercetare de interes a trecut prin ușoare modificări, migrând din zona semnalelor haotice către zona psiholingvisticii și a statisticii aplicate pe limba română scrisă. Acest lucru este aliniat și cu tema mai largă de cercetare aleasă pentru studiile doctorale: modele și metode statistice utilizate în studiul științelor vieții, și cu studiile complementare de psihologie finalizate în paralel cu studiile doctorale.

Capitolul 2 abordează subiectul limbii române scrise analizată dintr-o dublă perspectivă: psihologică și statistică. În secțiunea 2.1, care e bazată pe lucrarea [11] prezentată la CONSILR 2019, am revizitat noțiunea de independență statistică pentru limba română

scrisă, folosind un corpus literar existent în cadrul colectivului de cercetare. Obiectivul principal a constat în îmbunătățirea percepției și înțelegerii conceptului de independență statistică pentru limbaj natural și de a utiliza acest concept pentru a evalua din punct de vedere numeric proprietățile limbii scrise. Secțiunea 2.2, care e bazată pe lucrarea [12] prezentată la conferința COMM 2020, extinde rezultatele inițiale prezentate în secţiunea 2.1, de la m-grame la cuvinte. Am făcut apel la noțiunea de probabilitate pentru eroarea statistică de tipul II, care a fost amplu analizată în [13]. Secțiunea 2.3 continuă investigația legată de independența statistică pentru limba română scrisă văzută ca lanț de cuvinte. Studiul prezentat în 2.4 continuă analiza realizată de echipa de cercetare coordonată de Prof. Dr. Ing. Adriana Vlad privind structura de cuvinte în limba română. În cadrul acestei faze de cercetare au fost obţinute rezultate privind digramele (grupurile de 2 cuvinte succesive) aflate la sfârșitul frazelor / propoziţiilor, au fost îmbogățite cu explicații suplimentare rezultatele referitoare la digramele de cuvinte de început de frază și au fost investigate și perspective noi privind Legea Zipf si analiza pe subcorpus de autor. Secțiunea 2.5 revizitează noțiunea de corpus reprezentativ pentru un corpus lingvistic obținut prin alipiri de corpusuri literare distincte.

Capitolul 3 continuă investigaţia începută în capitolul 2, adăugând o nouă perspectivă: latura umană, psihologică a limbii scrise. Acest demers are drept scop investigarea legăturii dintre cunoștintele psiholingvistice dobândite de fiecare dintre noi de-a lungul vieții și validarea și corelarea acestora cu anumite concepte matematice și de teoria informatiei: entropia și probabilitățile condiționate.

Capitolul 4 abordează o direcţie diferită de cercetare - criptografia bazată pe sisteme haotice. Cercetarea inclusă în acest ultim capitol al tezei de doctorat încearcă să ofere o viziune unificatoare a mai multor perspective provenind de la un set foarte divers de discipline: biologie, genetică, economie şi criptografie, care lucrează aparent în paralel pentru a rezolva aceeaşi problemă. Toate aceste abordări transdisciplinare din aceste domenii au drept scop găsirea unei teorii a întregului, o teorie unificatoare care să facă ordine din haos, lumină unde e întuneric, să prezică cu acurateţe viitorul, pe baza faptelor trecute sau prezente.

Capitolul 5 prezintă concluziile lucrării sumarizând rezultatele originale principale obținute de-a lungul studiilor doctorale.

Capitolul 2

Psiholingvistica - studiul limbii române scrise dintr-o perspectivă statistică și psihologică

2.1 Independența statistică pentru limba română scrisă - caz de studiu - m-grame

2.1.1 Introducere

Noile rezultate experimentale obținute, referitoare la distanța minimă de independență statistică, sunt ilustrate pentru cazul m-gramelor de litere (litere, digrame, trigrame) într-un corpus lingvistic format din 49 de cărți scrise de 9 autori după cum urmează: Isaac Asimov – 9 cărți; Constantin Chiriţă – 5 cărți; Alexandre Dumas – 12 cărți; Colin Falconer – 1 carte; Frank Herbert – 8 cărți; Niven Larry – 1 carte; Orson Scott Card – 3 cărți; Michel Zevaco – 7 cărți; J. R. Tolkien – 3 cărți. Mărimea corpusului care include ortografie și punctuației este de 36 898 820 de caractere (peste 6 milioane de cuvinte). Cărțile au fost concatenate aleator pentru a forma corpusul analizat [14].

În cercetări anterioare ale autorilor a fost presupus faptul că aproximativ 200 de caractere sunt suficiente pentru a asigura independența statistică pentru limba română scrisă. Această ipoteză nu a fost pe deplin testată până în momentul în care am decis sa revizităm acest subiect important, dar rezultate indirecte anterioare îi susțineau validitatea.

2.1.2 Colectarea datelor și considerente teoretice

Colectarea datelor

Principala întrebare care se pune este: **care este distanța/numărul de caractere dintre o m-gramă și următoarea, astfel încât aceste două entități lingvistice nu sunt**

dependente/corelate una cu alta? Cu alte cuvinte, știind $m-grama_1$, avem vreo informație despre m-grama care se află la d caractere distanță, $m-grama_2$? Daca răspunsul este NU, atunci independența statistică a fost atinsă. Notă: m-grama aflată la d caractere distanță înseamnă că diferența dintre indexul primei litere din $m-grama_1$ și indexul primei litere din $m-grama_2$ este exact d. Cele descrise mai sus pot fi exprimate matematic în (2.1) sau (2.2):

$$P(m-grama_2|m-grama_1) = P(m-grama_2) \tag{2.1}$$

$$P(m-grama_1, m-grama_2) = P(m-grama_1) \cdot P(m-grama_2) \tag{2.2}$$

Noțiunea de independență statistică poate fi ușor explicată și pe baza vizualizării din Tabelul 2.1. Dacă independența statistică ar fi atinsă pentru $d = 6$, atunci probabilitatea că litera E ar urma la 6 caractere după litera A (sau dupa oricare literă) ar fi identică cu probabilitatea literei E: $P(E|A) = P(E)$. Aceasta este de fapt transpunerea (2.1) pentru cazul unigramelor/literelor. În mod similar, pentru cazul digramelor și presupunând ca ne interesează tot cazul $d = 6$, $P(ER|AC) = P(ER)$ sau $P(ERG|ACU) = P(ERG)$.

Tabel 2.1 Exemple de digrame și trigrame în limbajul natural.

Text	**A**	**C**	**U**	M		M	**E**	**R**	**G**	E	M
Numar	**1**	**2**	**3**	4	5	6	**7**	**8**	**9**	10	11

Teoria informației

O abordare pe care am folosit-o pentru a evalua distanța minimă de independență statistică este derivată din domeniul teoriei informației. M-gramele sunt colectate din textul scris conform vizualizării din Tabelul 2.1, Figura 2.1 și Figura 2.2.

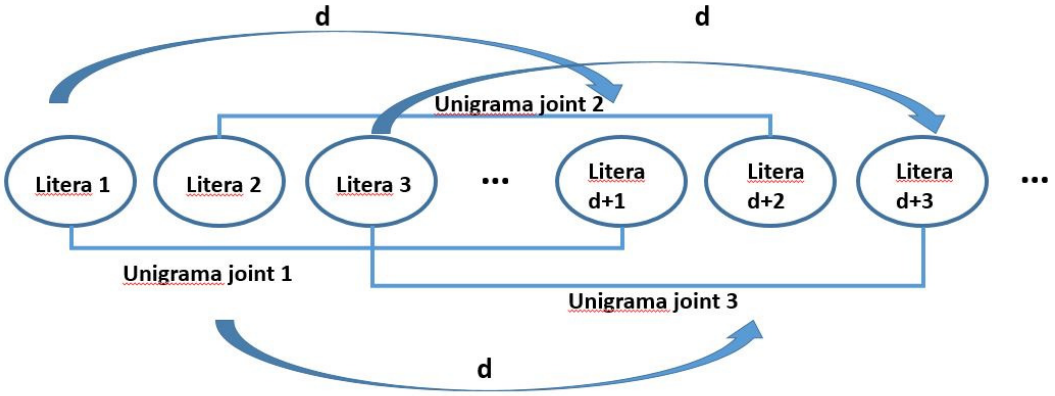

Figura 2.1 Colectarea literelor din corpus, fără salt

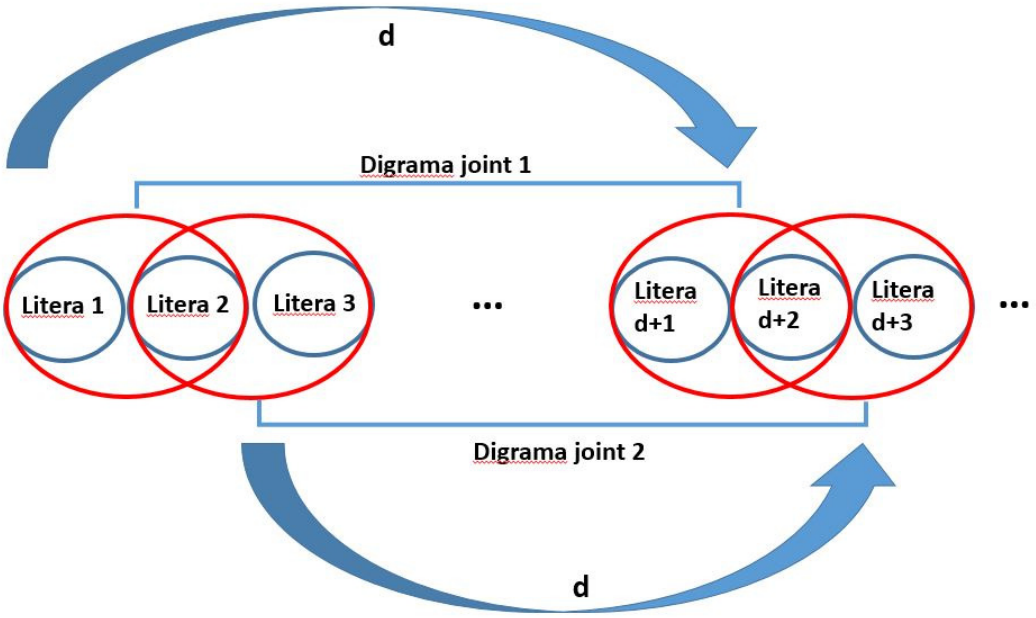

Figura 2.2 Colectarea digramelor din corpus, fără salt

Figura 2.3 prezintă vizualizarea tabelului de ocurențe/contingență atunci când literele sunt analizate, literele fiind ordonate de la cea mai frecventă în corpus până la cea mai puțin frecventă (sus-jos și stânga-dreapta).

Litera 2 / Litera 1	Blank	E	...	W	Q
Blank	n_{11}	n_{12}	...	$n_{1,k-1}$	n_{1k}
E	n_{21}	n_{22}	...	$n_{2,k-1}$	n_{2k}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
W	$n_{k-1,1}$	$n_{k-1,2}$	...	$n_{k-1,k-1}$	$n_{k-1,k}$
Q	n_{k1}	n_{k2}	...	$n_{k,k-1}$	n_{kk}

Figura 2.3 Numărul de apariții pentru perechi de litere separate de distanța d

Figura 2.3 corespunde canalului de informație/zgomot corespunzător distanței d, unde m-gramele de pe rânduri corespund intrărilor canalului, iar m-gramele de pe coloane ieșirilor [15]. Prin împărțirea dintre cele două cantități, se poate obține o estimație pentru probabilitatea condiționată a unei anumite perechi de m-grame separate de distanța d (vezi (2.3)).

$$P(m-grama_2|m-grama_1) \approx \frac{Numar\,de\,ocurente\,pentru\,perechea\,(m-grama_1, m-grama_2)}{Numar\,de\,ocurente\,pentru\,m-grama_1} \tag{2.3}$$

Atunci când ecuațiile (2.1) și (2.2) sunt îndeplinite, independența statistică este atinsă și valoarea transinformației devine 0. Un mod rapid de a evalua dacă independența statistică a fost atinsă este de a inspecta vizual tabelul de probabilități condiționate (matricea de zgomot asociată). Acesta are aceeași structură prezentată în Figura 2.3. Deviația fiecărei probabilități condiționate de la valoarea necondiționată este exprimată în (2.4). Atunci când independența se atinge, ne așteptăm ca matricea de eroare să fie foarte apropiată de 0.

$$Eroarea_{ij} = \frac{|P(m-gram_j|m-gram_i) - P(m-gram_j)}{P(m-gram_j)} \tag{2.4}$$

Tabelul 2.2 prezintă probabilitățile condiționate versus probabilitățile necondiționate pentru $d = 13$. De fapt, Tabelul 2.2 prezintă numeric ceea ce Figura 2.5 arată prin codul de culori asociat. Cele două tipuri de probabilități trebuie să fie egale atunci când se obține independența statistică. Se poate observa ușor că aproximarea între cele două tipuri de probabilități este mai vizibilă pentru simbolurile din colțul stânga-sus (simboluri frecvente), iar erorile cele mai mari sunt în colțul din dreapta-jos (cele mai rare litere).

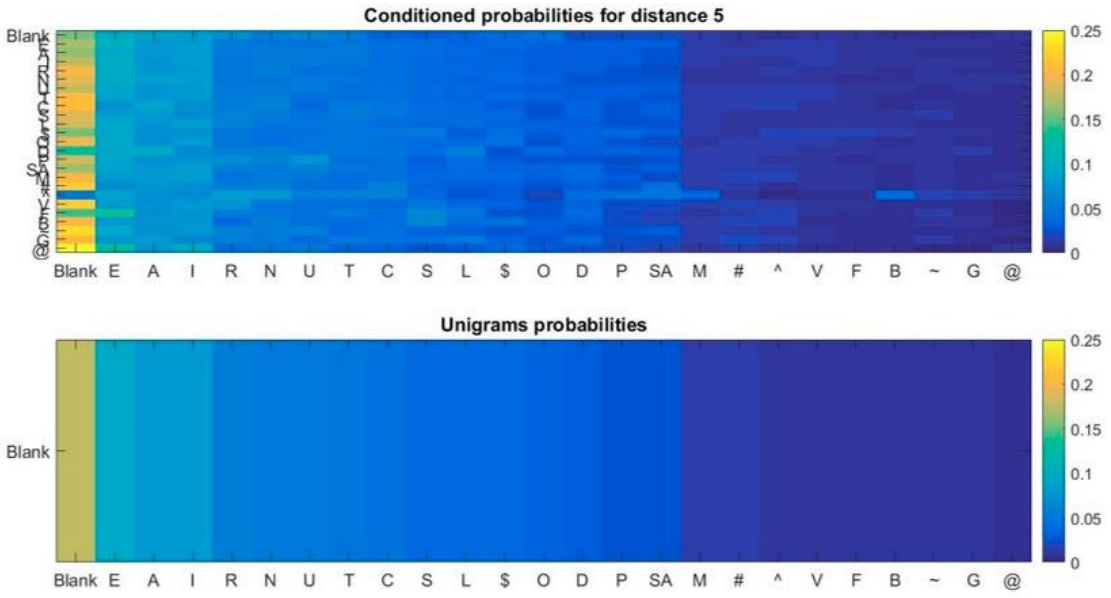

Figura 2.4 Probabilități condiționate pentru litere $P(Litera_2|Litera_1)$ pentru $d = 5$ (axa X - $Litera_2$; axa Y - $Litera_1$)

Comentariu asupra Tabelului 2.2: probabilitatea condiționată că litera A urmează la 13 litere dupaă litera E este 8.02% (am început analiza cu $d = 13$ pe baza rezultatelor din secțiunea cu abordarea bazată pe testul Hi-pătrat. SA este simbolul artificial obținut prin concatenarea celor mai puțin frecvente 8 caractere din alfabetul corpusului.

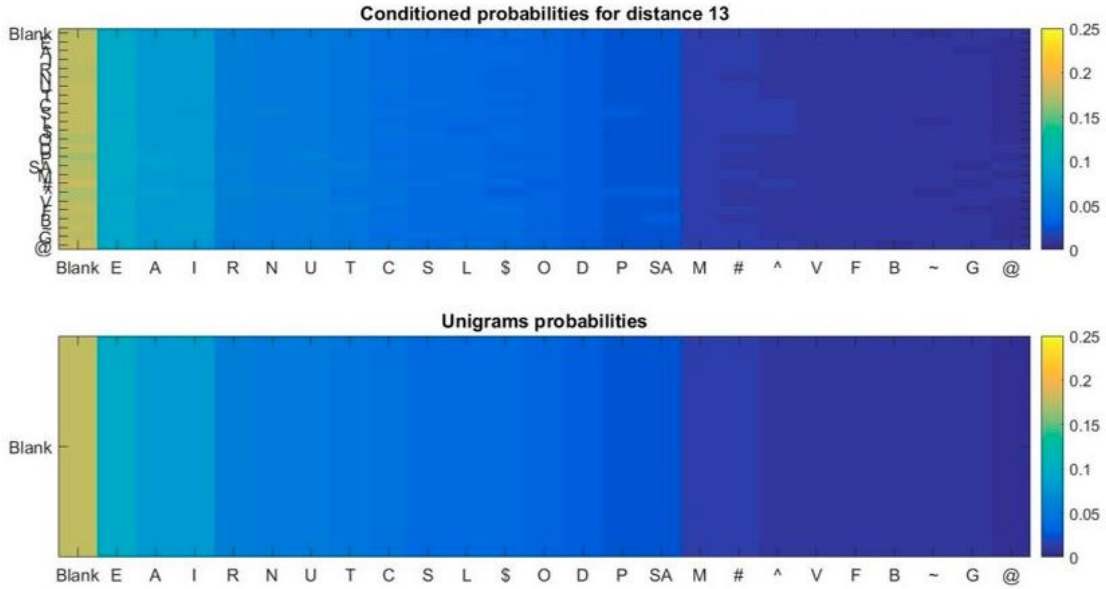

Figura 2.5 Probabilități condiționate pentru litere $P(Litera_2|Litera_1)$ pentru $d = 13$ (axa X - $Litera_2$; axa Y - $Litera_1$)

Tabel 2.2 Probabilități condiționate pentru litere $P(Litera_2|Litera_1)$ pentru $d = 13$ (rânduri - $Litera_1$; coloane - $Litera_2$)

Litere/Probabilități	**Blank**	**E**	**A**	**...**	**SA**	**...**	**â**
Blank	17.56	9.40	8.13	...	2.61	...	0.80
E	17.75	9.70	8.02	...	2.54	...	0.73
A	17.79	9.54	8.08	...	2.52	...	0.72
...	...	...	...	...	...	...	...
SA (simbol artificial)	17.53	9.48	8.30	...	2.63	...	0.79
...	...	...	...	...	...	...	...
â	18.03	9.73	7.92	...	2.38	...	0.86
Probabilitatea literei cap de coloană	17.63	9.54	8.09	...	2.57	...	0.75

Datele din Tabelul 2.2 au fost utilizate pentru a calcula erorile dintre probabilitățile condiționate și cele necondiționate pe baza (2.4). Rezultatul poate fi vizualizat în Figura 2.6 pentru $d = 5$ și Figura 2.7 pentru $d = 13$. Atunci când independența statistică e atinsă, matricea de eroare ar trebui să devină 0 și acest lucru ar trebui să corespundă unei matrici complet albastre conform codului de culoare folosit.

Se poate observa cum dacă la distanța $d = 5$ mai apar câteva zone mai colorate (erori mari între probabilitățile condiționate și cele necondiționate - Figura 2.6), pe măsură ce distanța crește spre $d = 13$ și mai mult, erorile devin mult mai mici - Figura 2.7.

Testul Hi-pătrat în tabele de contingență

Acest test permite o viziune de ansamblu a situației și face apel la reprezentarea din Figura 2.3 (și implicit la matricea de zgomot asociată), iar împreună cu abordarea

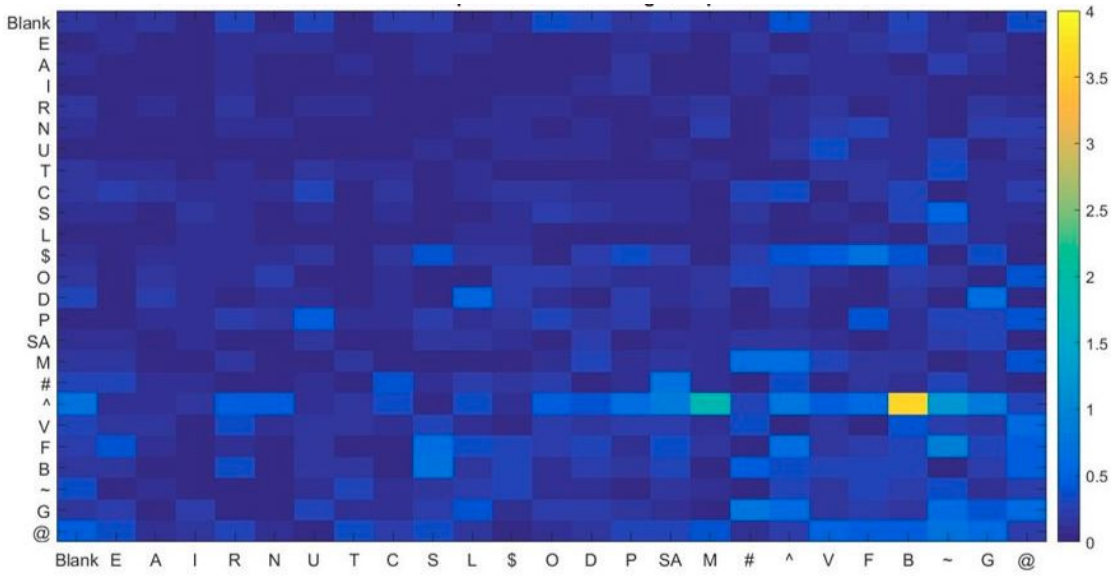

Figura 2.6 Matricea de erori (%) pentru litere pentru $d = 5$

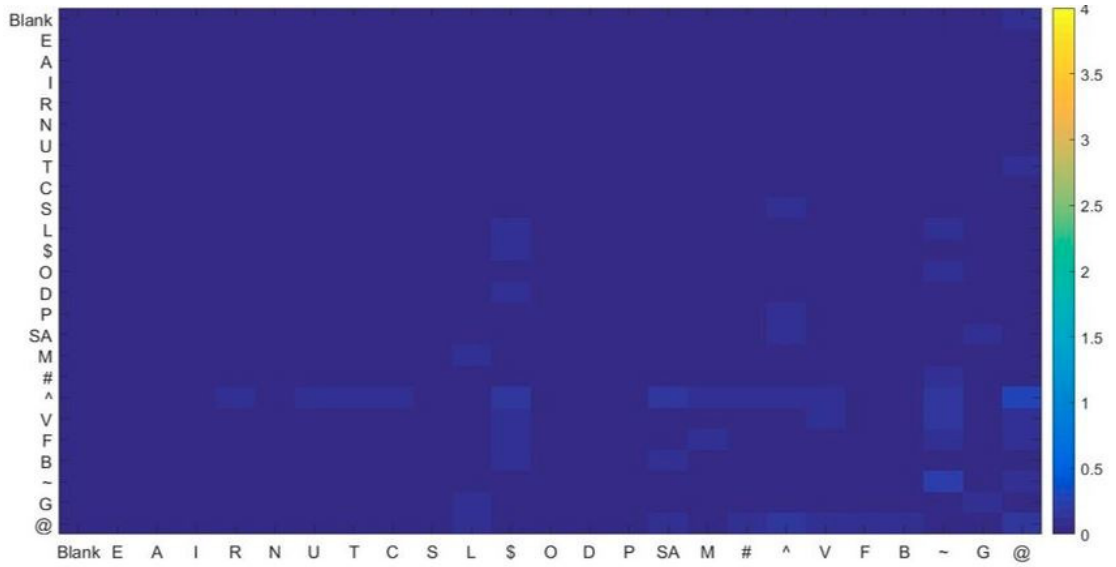

Figura 2.7 Matricea de erori (%) pentru litere pentru $d = 13$

bazată pe teoria informației conduce la un răspuns mai rapid referitor la problematica independenței statistice.

O cerință foarte importantă pentru testul Hi-pătrat este ca datele experimentale să provină din variabile aleatoare independente și identic distribuite, sensibilitatea testrului la date i.i.d. fiind un aspect cunoscut în literatura de specialitate [16]. Figura 2.8 prezintă procedura utilizată pentru a colecta datele i.i.d. din corpus necesare pentru testul Hi-pătrat.

Condiția referitoare la numărul minim de apariții al unei m-grame se aplică și în acest caz. Atunci când condiția DeMoivre-Laplace nu este îndeplinită, rândurile și coloanele corespunzătoare sunt concatenate într-un simbol artificial, astfel reducându-se mărimea tabelului de contingență. De exemplu, în cazul literelor, unde alfabetul corpusului

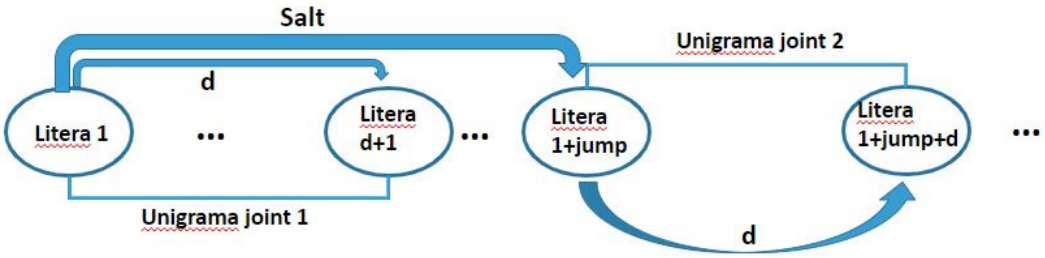

Figura 2.8 Colectarea datelor i.i.d. (cu salt) pentru testul Hi-pătrat în tabele de contingență

are mărimea 32, tabelul de contingență redus are un număr mai mic de simboluri, și anume 25. Algoritmul folosit pentru a reduce mărimea tabelului de contingență poate fi vizualizat în Figura 2.9.

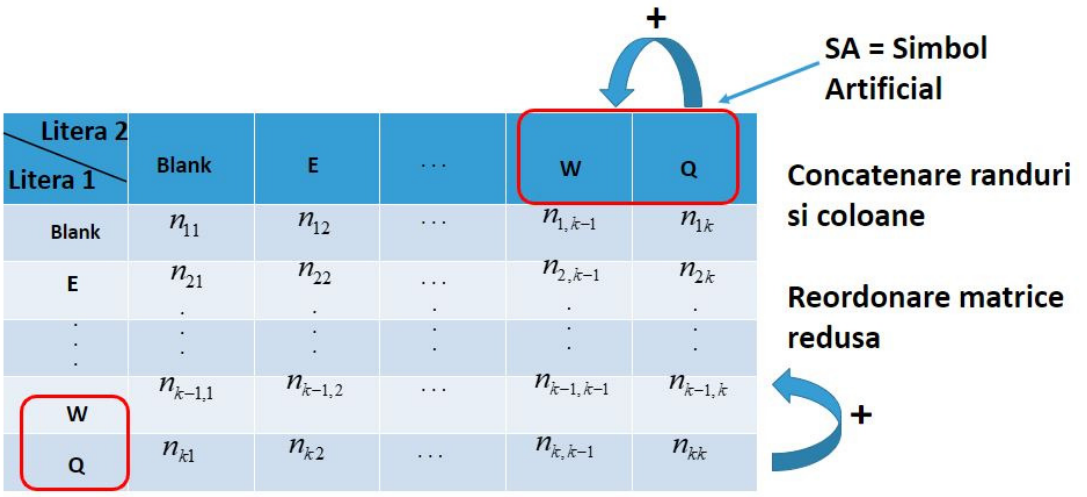

Figura 2.9 Reducerea tabelului de contingență pentru testul Hi-pătrat

Algoritmul este foarte simplu: pentru fiecare distanță d, este calculată o valoare de test conform (2.5) și aceasta este comparată cu o valoare admisă (alfa-cuantila unei distribuții Hi-pătrat cu $(k-1)\cdot(k-1)$ grade de libertate, notată z_α).

$$z = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(n_{ij} - \frac{n_i \cdot n_j}{n})^2}{\frac{n_i \cdot n_j}{n}} \tag{2.5}$$

Daca $z < z_\alpha$, spunem că testul Hi-pătrat trece și independența a fost atinsă; altfel, testul nu trece și următoarea distanță d este verificată [16–18].

2.1.3 Rezultate experimentale

Figurile 2.10 și 2.11 prezintă rezultatele testului Hi-pătrat atunci când literele sunt colectate din corpus cu un salt de 200 de caractere. Figura 2.10 arată evoluția valorii de test z comparativ cu z_α pentru diferite distanțe d. Figura 2.11 prezintă mărimea tabelului de contingență redus. Este interesant de observat că acum nu doar z variază cu distanța, dar și z_α este variabil, iar acest lucru se întâmplă datorită faptului că mărimea tabelului de contingență este variabilă de asemenea.

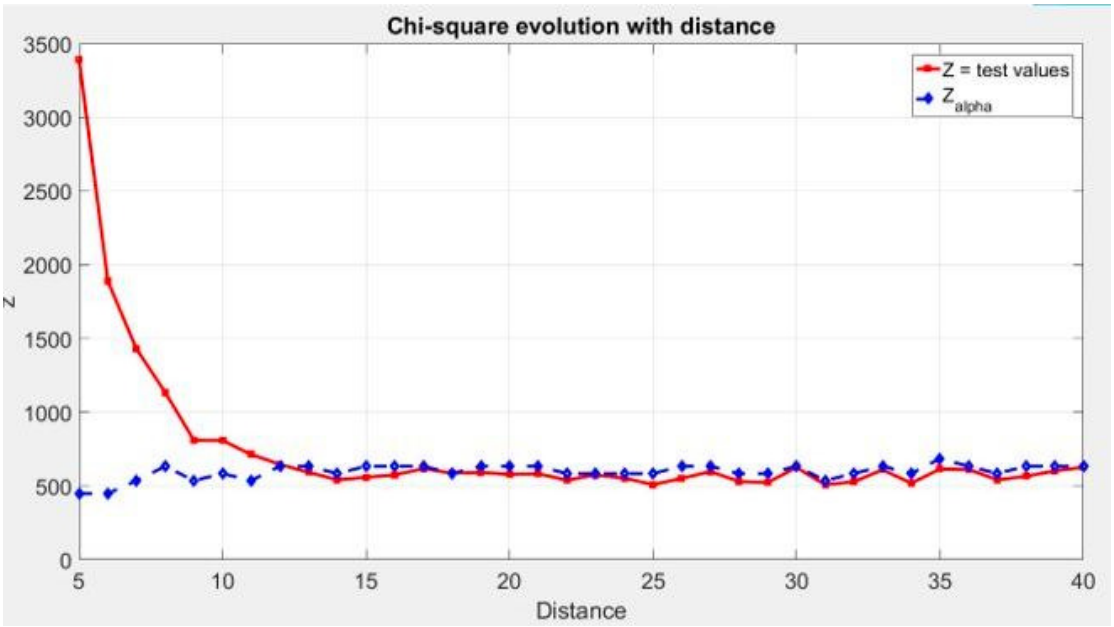

Figura 2.10 Rezultatele testului Hi-pătrat cu salt pentru litere (axa X = distanța d; axa Y - valoarea z de test - vezi (2.5))

Concluzia obținută după aplicarea testului Hi-pătrat este că independența statistică este atinsă pentru $d = 13$ în cazul literelor ($m = 1$).

2.1.4 Concluzii

Rezultatele menționate in 2.1 confirmă ideea că o distanță de 80-100 de caractere este mai mult decât suficientă pentru a asigura independența statistică în cazul m-gramelor extrase din texte naturale ale limbii române scrise. De asemenea, distanța de independență este analizată prin două metode: testul Hi-pătrat în tabele de contingență și teoria informației.

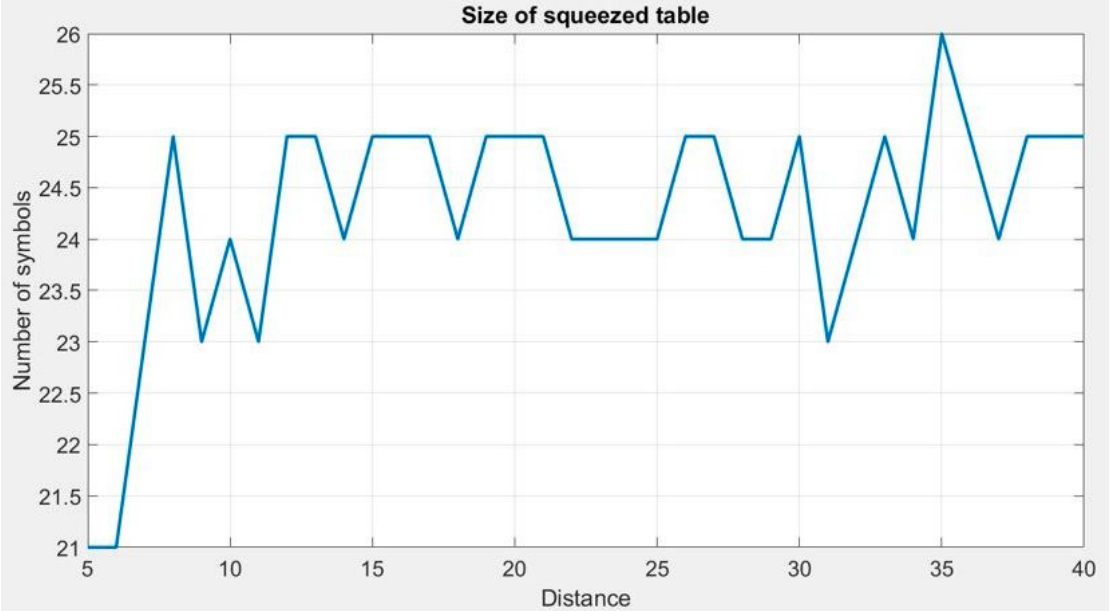

Figura 2.11 Rezultatele testului Hi-pătrat cu salt pentru litere (axa X = distanța d; axa Y - numărul de simboluri din tabelul de contingență redus)

2.2 Independența statistică pentru limba română scrisă - caz de studiu - cuvinte

2.2.1 Rezultate experimentale

Figurile 2.12 și 2.13 prezintă rezultatele testului Hi-pătrat atunci când cuvintele sunt colectate din corpus cu un salt de 100 de cuvinte. Figura 2.12 arată evoluția valorii de test z comparativ cu z_α pentru diferite distanțe d. Figura 2.13 prezintă mărimea tabelului de contingență redus. Acesta este de fapt tabelul folosit de testul Hi-pătrat pentru a decide legat de distanța de independență minimă. Este interesant de observat că acum nu doar z variază cu distanța, dar și z_α este variabil. Acest lucru este logic și se întâmplă datorită faptului că mărimea tabelului de contingență se modifică de asemenea cu distanța.

O primă concluzie obținută după aplicarea testului Hi-pătrat este că independența statistică este atinsă pentru $d = 9$ în cazul cuvintelor.

Teoria informației

Figurile 2.14, 2.15 și 2.16 arată cum pe măsură ce distanța d dintre cuvinte crește, probabilitățile condiționate (jumătatea superioară a figurilor) încep să arate foarte similar cu probabilitățile individuale/necondiționate (jumătatea inferioară a figurilor). Valori aproximativ constante pentru probabilitățile condiționate înseamnă că independența e atinsă.

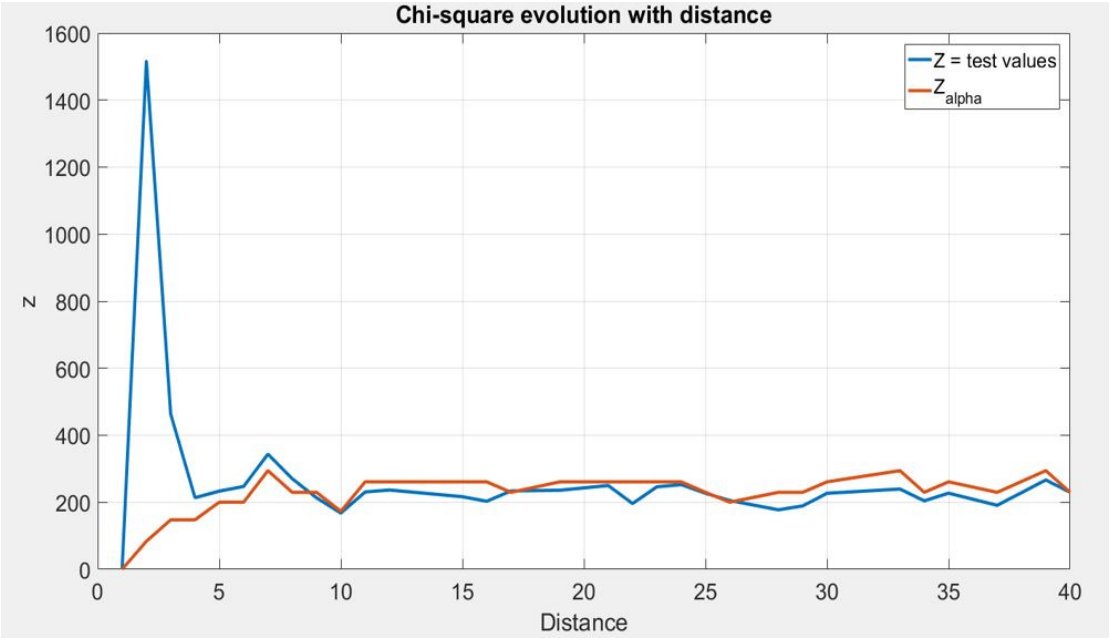

Figura 2.12 Rezultatele testului Hi-pătrat cu salt pentru cuvinte (axa X = distanța d; axa Y - valoarea z de test - vezi (2.5))

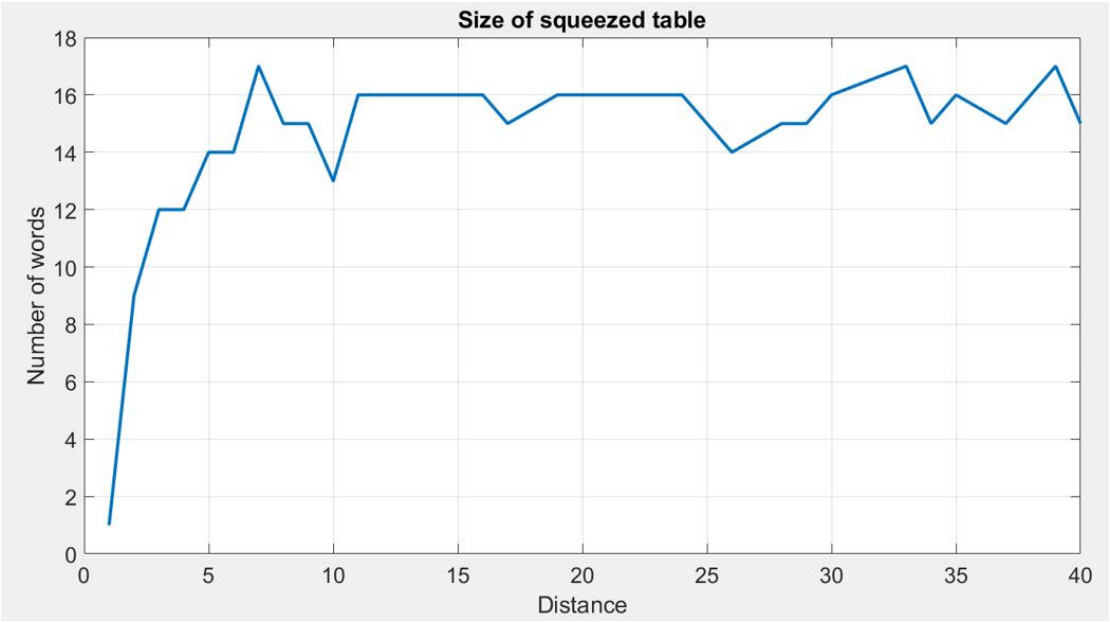

Figura 2.13 Rezultatele testului Hi-pătrat cu salt pentru cuvinte (axa X = distanța d; axa Y - numărul de cuvinte din tabelul de contingență redus)

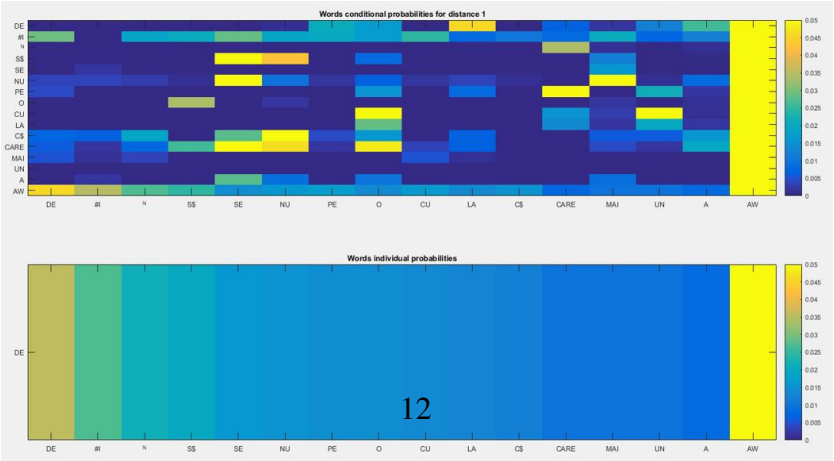

Figura 2.14 Probabilități condiționate pentru cuvinte $P(Cuvant_2|Cuvant_1)$ pentru $d = 1$

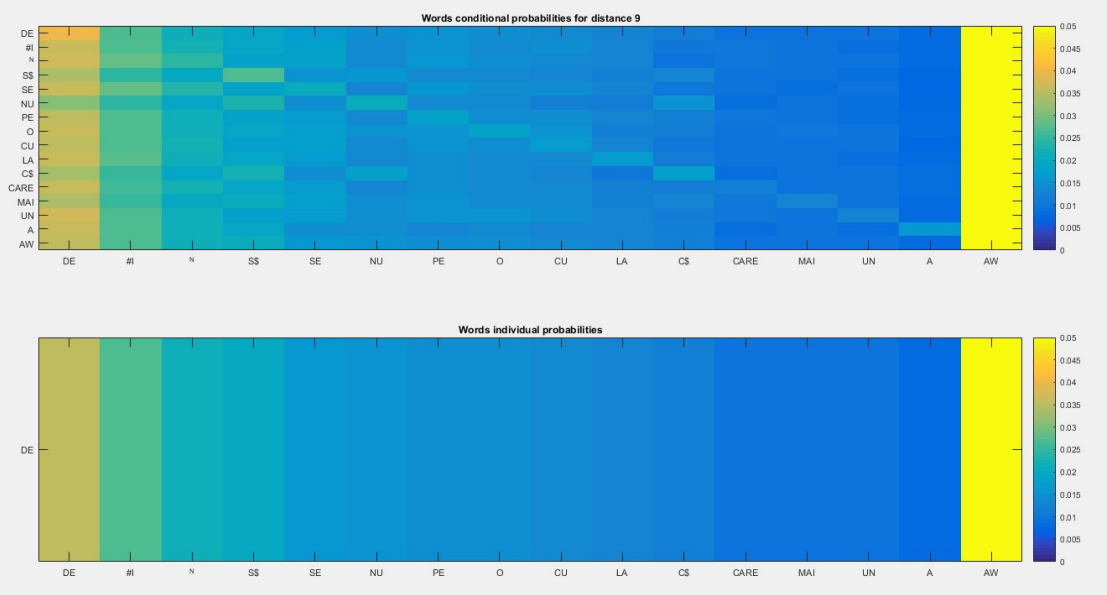

Figura 2.15 Probabilități condiționate pentru cuvinte $P(Cuvant_2|Cuvant_1)$ pentru $d = 9$ (axa X - $Cuvant_1$; axa Y - $Cuvant_2$)

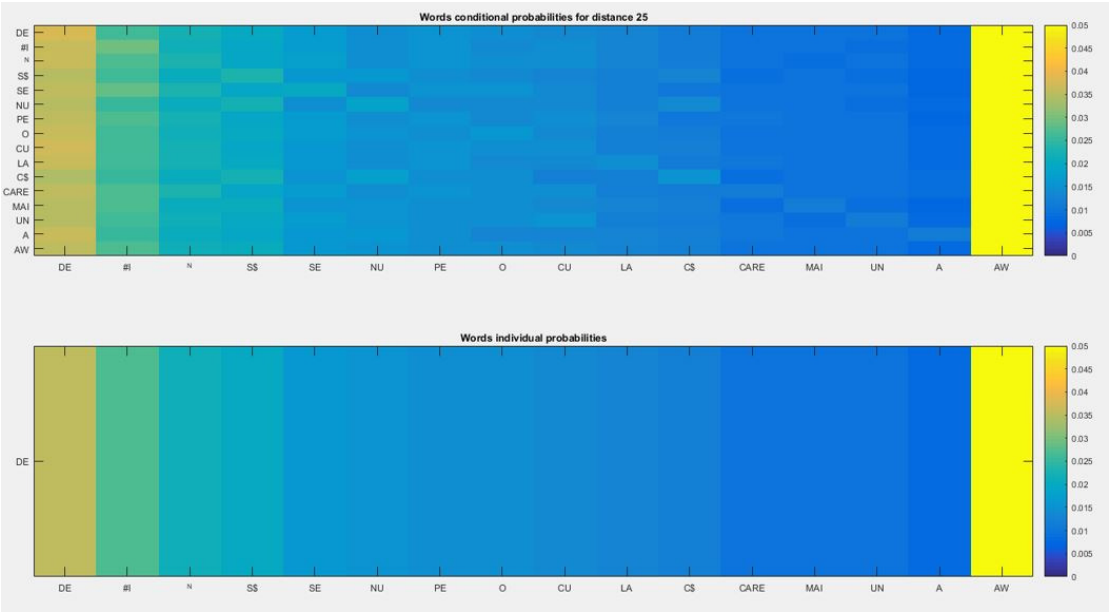

Figura 2.16 Probabilități condiționate pentru cuvinte $P(Cuvant_2|Cuvant_1)$ pentru $d = 25$ (axa X - $Cuvant_1$; axa Y - $Cuvant_2$)

Se poate observa ușor o coloană galben intens în dreapta fiecărei figuri (vezi Figura 2.16): aceasta corespunde CA (cuvântului artificial) care înglobează cuvinte ce acoperă aproximativ 76% din corpus. Această probabilitate condiționată este foarte bine aproximată și se apropie foarte mult de valoarea necondiționată datorită valorii sale foarte mari.

2.2.2 Concluzii

Am reușit să demonstrăm că aproximativ 100 de cuvinte sunt suficiente pentru a garanta independența statistică atât pentru corpusul cu un alfabet de 32 de simboluri, cât și pentru corpusul cu un alfabet de 47 de simboluri.

Secţiunile următoare din capitolul 2 construiesc si extind rezultatele din secţiunile 2.1 şi 2.2, aşa că ne vom rezuma la a prezenta doar concluziile, detaliile fiind incluse pe larg în teza de doctorat.

2.3 Independența statistică și probabilitatea de eroare statistică de tipul II (β)

Secțiunea 2.3 revizitează conceptul de independență statistică pentru limba română scrisă. Aici, focusul este eroarea statistică de tipul II, care joacă un rol important în discuția despre independență, iar rezultatele anterioare ale colectivului de autori [19, 20] sunt corelate cu rezultatele prezentate în [11, 12] din acest punct de vedere.

Pentru analiza care implică cuvintele și probabilitățile acestora, corpusul folosit nu este suficient de extins pentru a considera toate cuvintele distincte în investigația de independență statistică. Anumite aspecte, cum ar fi o anumită lungime a corpusului (cum este cel folosit în această lucrare, de peste 36 de milioane de caractere lungime) și considerarea erorii statistice de tipul II în evaluarea probabilităților cuvintelor, conduc la setarea unei valori limită (minime) a probabilității cuvintelor implicate în analiză de probabilitate, pentru a obține rezultate semnificative statistic.

Investigațiile de independență bazate pe testul Hi-pătrat în tabele de contingență au găsit aceste limite, exprimate prin reducerea dimensiunii tabelului de contingență asociat și introducerea Cuvintelor Artificiale obținute prin combinarea mai multor cuvinte distincte mai puțin frecvente în corpus (CA create pentru a în deplini cerințele impuse de β). În continuare, mai multe scenarii de creare a CA au fost analizate. În toate cele 4 scenarii rezultatele de independență obținute au fost similare (1 CA, 3 CA egal probabile, 10 CA egal probabile, 10 CA cu probabilități diferite). Distanța minimă de independență nu este dependentă de modul de creare a CA (rezultat așteptat pe baza ergodicității limbajului natural), singura limitare care trebuie îndeplinită este legată de valoare minimă de probabilitate a CA, valoare strâns legată de nivelul acceptat al probabilității erorii statistice de tipul II (β).

2.4 Analiză privind grupurile de două cuvinte succesive de început și respectiv de final de frază / propoziție, pe corpus literar de limbă română scrisă cu ortografie și punctuație

Corpusul de ansamblu pe care s-a realizat studiul (#NCO) are lungimea de 6.377.720 cuvinte. Corpusul este format din 49 de cărţi (romane şi nuvele de autori români şi traduceri) scrise de 9 autori şi a fost construit anterior de echipa de cercetare. Studiul prezintă şi o analiză comparativă cu un subcorpus format din cărţile a 3 autori, anume #CHIRIŢĂ, #HERBERT şi #DUMAS (în total, 20 de cărți din totalul de 49 de cărți ce formează corpusul de ansamblu).

S-au considerat în prezenta analiză patru delimitatori / semne de punctuaţie care marchează sfârşitul / începutul de frază / propoziţie:

1. Punct
2. Întrebare
3. Exclamare
4. Puncte de suspensie

Analiza a fost realizată în mai multe etape etape:

1. S-a realizat o extragere a digramelor de cuvinte de final de frază / propoziţie indiferent de semnul cu care se încheie fraza / propoziţia precedentă. De menţionat că acest lucru se poate face doar pentru acele fraze / propoziţii cu lungime de cel puţin 2 cuvinte. Aceste digrame de început au fost numărate şi sortate, lucrarea prezentând tabelar cele mai importante digrame de cuvinte de final de frază / propoziţie. În corpusul de ansamblu există peste 522 000 de fraze de lungime egală sau mai mare decât 2 şi un număr de digrame distincte de final de propoziţie / frază de 289 894.
2. Analiza de la punctul 1 a fost reluată separat pentru fiecare din cei 4 delimitatori (punct / semnul întrebării / semnul exclamării / puncte de suspensie).
3. În lucrări anterioare ale autorilor au fost prezentate mai multe rezultate privind repartizarea cuvintelor ce compun corpusul in 3 regiuni diferite de-a lungul graficului Legii lui Zipf, în functie de numărul de apariții pentru fiecare cuvant distinct. Regiunea 1 corespunde cuvintelor cu mai mult de 200 de apariții. Acestea sunt primele 2762 cuvinte și acoperă aproximativ 72% din corpus. Ne interesează în cate dintre digramele de sfârșit de frază apar cuvinte din zona 1 Zipf. Rezultatele sunt următoarele: peste 19% dintre digramele de final de frază conțin ambele

cuvinte din zona 1 Zipf, aproximativ 64.4% dintre digrame au in componență un singur cuvânt din prima regiune Zipf și 16.5% dintre digramele de final de propoziție nu conțin niciun cuvânt din prima regiune Zipf. Comparând cu rezultate anterioare ale colectivului de cercetare (pentru digramele de inceput de frază / propoziție), se observă că rezultatele pentru digramele de cuvinte de final de frază sunt diferite față de digramele de cuvinte de început de frază, acestea din urmă conținând într-o măsură mai mare ambele cuvinte din prima regiune Zipf.

4. S-a analizat, de asemene, o conexiune între cuvintele din digramele de cuvinte de final de frază / propoziţie și mulțimea de 578 de digrame de cuvinte comune în cele 49 cărţi din corpusul de ansamblu. Aproximativ 1% dintre digramele de cuvinte comune (existente în formă identică de scriere în toate cărţile) sunt și digrame de cuvinte de final de frază / propoziţie.

5. Etapele 1 şi 2 au fost reluate şi la nivelul subcorpusului din cărţile celor 3 autori. Analiza pe acest subcorpus a pus în evidenţă o serie de valori cantitative relativ stabile (cu mici variaţii) în comparaţia făcută.

Privind pe ansamblul rezultatelor se poate aprecia că s-au obţinut rezultate inedite privind grupurile de două cuvinte succesive de final de frază / propoziţie, pe corpus literar de limbă română scrisă cu ortografie şi punctuaţie. Studiul a subliniat încă o dată impactul ortografiei şi punctuaţiei în modelul limbii şi importanţa cuvintelor de final şi a digramelor de cuvinte de final de propoziţie / frază în descrierea statistică a limbii.

2.5 Corpus literar reprezentativ

Am prezentat o metodă simplă de a colecta datele din corpusul NLCO şi de a crea o structură matriceală care a fost analizată mai departe dintr-o perspectivă rang-frecvenţă. Un aspect important pe care l-am avut în vedere când am colectat datele din corpus este distanţa minimă de independenţă, care a fost estimată în jurul valorii de 100 de cuvinte pentru limba română scrisă [12].

După ce datele au fost extrase din NLCO şi cele 3 seturi de interes au fost selectate, am putut observa similarităţi din punct de vedere al rangului şi frecvenţei relative. Cele mai frecvente cuvinte distincte din cele 3 seturi analizate sunt foarte corelate cu cele mai frecvente cuvinte distincte din corpusul NLCO. Aceasta este o primă dovadă care susţine ipoteza că seturile analizate sunt eşantioane de corpus reprezentativ pentru limba română scrisă.

Mai mult, 99% dintre cuvintele distincte comune celor 3 seturi de date (coloanele 1, 150 şi 300) fac parte din Aria 1 Zipf. În plus, 75% dintre cuvintele distincte comune celor 3 seturi de date fac parte şi din setul de cuvinte comune pentru subcorpusurile de autor din NLCO.

Noile rezultate susţin ideea de reprezentativitate a corpusului construit prin concatenări succesive ale mai multor subcorpusuri de autor, venind totodată şi cu beneficiile concatenării (lungime mai mare) pentru cercetătorul din domeniul procesării limbajului natural.

Capitolul 3

Psihologie şi teoria informaţiei

3.1 Introducere

Problematica abordată în acest capitol se bazează pe rezultatele anterioare dintr-un studiu experimental complex asupra impactului pe care îl au semnele de ortografie și punctuație în modelul limbii române scrise, când limba naturală este privită ca lanț de cuvinte [15, 21–26]. Rezultatele au fost publicate şi prezentate în cadrul Universităţii Titu Maiorescu, Bucureşti [27, 28].

Înainte de a avansa cu analiza statistică a corpusului, este important să introducem noțiunea de entropie informaţională, care va juca un rol fundamental în discuțiile următoare. În teoria informației, entropia Shannon sau entropia informațională măsoară incertitudinea asociată cu apariţia unui eveniment aleator. Pentru o sursă X binară asociată experimentului aruncării unei monede: de exemplu, o monedă cu probabilitățile de apariţie a feţelor p și 1-p, entropia sursei este dată de (3.1):

$$H(X) = -p \cdot log_2(p) - (1-p) \cdot log_2(1-p)[biti] \tag{3.1}$$

În cazul general, când sursa de informație X are n simboluri cu probabilitățile simbolurilor p_i, entropia se poate scrie ca in (3.2):

$$H(X) = -\sum_{i=1}^{n} p_i \cdot log_2(p_i)[biti] \tag{3.2}$$

În această cercetare am investigat cazul particular al digramelor de cuvinte din limba română și predicția cuvântului al doilea din digramă știind primul cuvânt:

1. O digramă de cuvinte este orice grup de două cuvinte consecutive ($Cuvant_i, Cuvant_j$)
2. Pe baza tuturor digramelor din corpusul analizat, se pot estima probabilitățile condiționate ale cuvintelor în funcție de cuvintele aflate cu o poziție în fața lor:
    - Pasul 1 – se calculează intrările în tabelul de ocurențe din Figura 3.1 (introdus deja în capitolul 2)

- Pasul 2 – se folosește informația din tabelul de ocurențe și se calculează probabilitățile condiționate conform (3.3)

$$P(Cuvant_j|Cuvant_i) = P(j|i) = \frac{n_{ij}}{n_i} \tag{3.3}$$

  - n_{ij} este numărul de ocurențe/apariţii corespunzător digramei (i,j), iar n_i este numărul de ocurențe corespunzător cuvântului i
  - Este important de menționat că voi folosi în continuare exprimarea Cuvântul i pentru cuvântul cu rangul i în ordinea descrescătoare a frecvențelor de apariție în corpus

Cuvinte / Cuvinte	$Cuvant_1$	$Cuvant_2$	...	$Cuvant_j$	...	$Cuvant_l$
$Cuvant_1$	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}
$Cuvant_2$	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$Cuvant_i$	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$Cuvant_l$	n_{l1}	n_{l2}	...	n_{lj}	...	n_{ll}

Figura 3.1 Tabelul de contingenţă pentru digrame de cuvinte

Odată inventariate digramele de cuvinte și probabilitățile condiționate estimate, este momentul să facem tranziția spre mărimea de interes în cadrul acestui studiu: entropia condiţionată. Este de la sine înțeles că fiecare cuvânt poate fi urmat (teoretic) de oricare dintre cuvintele din corpus. În practică, nu orice cuvânt are sens să urmeze după oricare cuvânt anterior, dar relația (3.4) rămâne valabilă:

$$\sum_{\substack{j=1 \\ i\,fixat}}^{l} P(j|i) = 1 \tag{3.4}$$

unde l reprezintă numărul total de cuvinte distincte din corpus.

Putem asocia apariţiei fiecărui cuvânt un experiment similar aruncării cu un zar, pe fiecare față a zarului fiind unul dintre cuvintele j posibile care pot urma după i. Cu alte cuvinte, pentru fiecare cuvânt i avem o mini-sursă de informație (dată de limba română scrisă), alegerile sau evenimentele posibile fiind reprezentate de cuvintele j posibile care pot urma după un cuvânt i fixat. Relaţia (3.4) este binecunoscuta condiţie de la matricea

de zgomot a unui canal de informaţie - suma pe linii este 1 (liniile corespund intrărilor canalului de informaţie - cuvintelor i, iar coloanele corespund ieşirilor - cuvintelor j).

Pentru fiecare dintre aceste surse de informație (cuvânt i) se poate calcula o entropie condiționată (3.5):

$$H_i = H(j|i) = -\sum_{j=1}^{l} P_(j|i) \cdot log_2(P(j|i)[bii] \tag{3.5}$$

Entropia informațională înglobează informații importante despre surpriza pe care un observator o are când o anumită față a zarului apare. În cazul acestui studiu, surpriza se transpune în uimire că un anumit cuvânt (j) urmează după un alt cuvânt (i).

Figura 3.2 prezintă grafic valorile entropiei condiţionate pentru primele cele mai frecvente 500 de cuvinte din limba română scrisă. Se poate observa o ușoară scădere a valorilor entropiei pe măsură ce cuvintele devin mai puțin frecvente în limbă. Acest fapt este explicabil, pentru că primele cuvinte, cele mai frecvente, pot fi urmate foarte/la fel de probabil de multe alte cuvinte, pe când cuvintele mai puțin frecvente prezintă anumite preferințe – sunt urmate mai frecvent doar de anumite cuvinte, deci au o entropie mai mică.

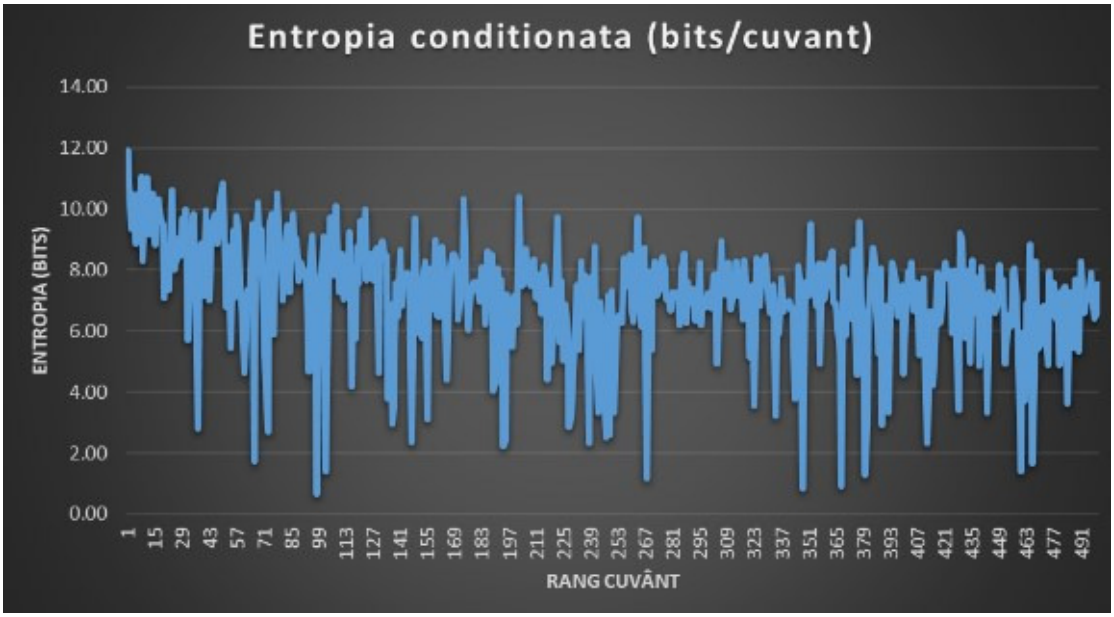

Figura 3.2 Entropia condiţionată pentru primele cele mai frecvente 500 de cuvinte distincte din corpus

Am pornit la drum cu următoarele ipoteze:

- Ipoteza 1: rata de succes în "ghicirea" cuvântului următor după un cuvânt de interes este invers proporțională/corelată negativ cu valoarea entropiei asociată cuvântului fixat.
- Ipoteza 2: răspunsurile obținute la chestionar sunt bine corelate cu rezultatele obținute pe baza corpusului literar.

Prima abordare este bazată pe analiza statistică pe un corpus literar de limba română scrisă obținut prin concatenarea a 49 de cărți (beletristică). Pe baza acestuia, am extras statistici descriptive cum ar fi frecvența cuvintelor sau grupurilor de cuvinte cele mai frecvente în limba română. În același timp, se pot calcula foarte ușor și estimații pentru probabilitățile condiționate ale cuvintelor din limba română, iar pe baza acestor probabilități se pot asocia entropii condiționate fiecărui cuvânt (grupul de interes a fost reprezentat de primele cele mai des folosite 500 de cuvinte) - (3.5).

În al doilea rând, pe baza analizei realizate pe corpusul literar menționat mai sus, am selectat 30 de cuvinte reprezentative din punct de vedere informațional și al entropiei condiționate: 10 cuvinte cu entropie mică, 10 cuvinte cu entropie medie și 10 cuvinte care au asociată o valoare mare a entropiei (ipoteza de la care am plecat și care e bazată pe observațiile lui C.E. Shannon [29] este că entropia este într-o relaţie de invers proporționalitate cu cantitatea de surpriză asociată unui anume cuvânt referitor la cuvântul care îi urmează). Cele 30 de cuvinte astfel selecționate din corpus au fost folosite pentru a realiza un chestionar cu întrebări de forma: **“Care credeți că este cel mai frecvent cuvânt în limba română dupa cuvântul . . . ?”**.

În continuare, au fost realizate corelații între mărimile de interes: rata de succes a participanților la chestionar (prin raportare la varianta considerată corectă şi care e bazată pe rezultatele statistice din corpus) și valoarea entropiei condiționate pentru fiecare cuvânt inclus în chestionar. Variabile secundare, cum ar fi vârsta, genul sau nivelul de studii al participanţilor la studiu, au fost utilizate de asemenea pentru elaborarea unor concluzii despre subiectul de interes.

Chestionarul a fost realizat folosind mediul online Google Forms și a constat în 30 de întrebări de forma: Care credeti că este cuvântul cel mai des folosit în limba română după cuvântul CEEA? Chestionarul a fost trimis și a primit răspunsuri de la participanții la studiu în perioada 12-15 Aprilie 2020. Respondenții au avut de ales între 6 variante, toate posibile și cu sens, constând în cuvintele cele mai frecvente după cuvântul de interes (CEEA, în exemplul de mai sus). Completarea chestionarului a fost anonimă, participanții oferind doar informații generale, cum ar fi: vârsta, genul (M/F) și nivelul de studii. La chestionar au răspuns 101 de persoane. Distribuția respondenţilor în funcție de informaţiile colectate este prezentată în cele ce urmează.

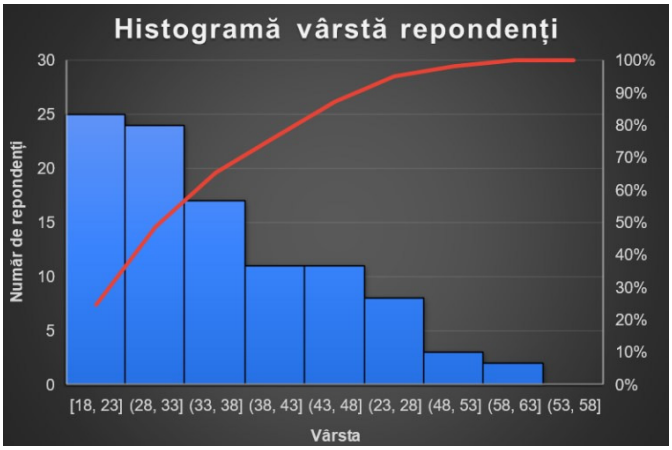

Figura 3.3 Histogramă vârstă respondenți

Figura 3.3 prezintă distribuția participanților la studiu pe baza vârstei lor. Se observă că peste jumătate dintre participanți au sub vârsta de 35 de ani, restul fiind relativ uniform distribuiți și acoperind gama 35 – 58 de ani.

Din punct de vedere al nivelului de studii, majoritatea respondenților au absolvit studii de licență (peste 90%). Rezultatele detaliate privind distribuția participanților în funcție de nivelul de studii poate fi vizualizat în Figura 3.4.

În ceea ce privește repartizarea respondenților în funcție de gen, rezultatele sunt după cum urmează:

- 74 sunt de gen feminin
- 27 sunt de gen masculin

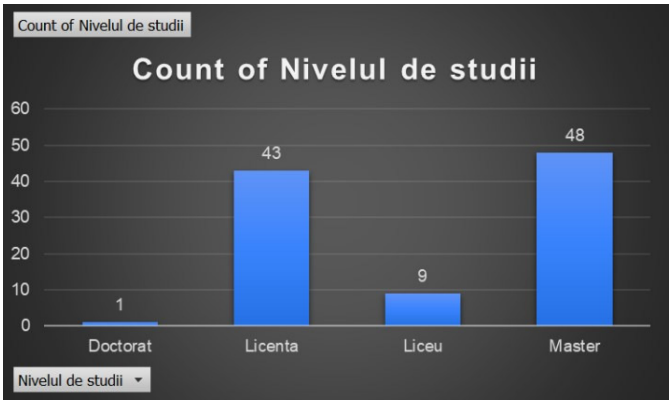

Figura 3.4 Histogramă nivel studii respondenți

3.1.1 Interpretarea cantitativă și calitativă a datelor

Pentru a calcula cu ușurință corelațiile dorite, răspunsurile la chestionar au fost sumarizate astfel:

1. Pentru fiecare întrebare am calculat câte persoane au ales fiecare dintre variantele posibile de răspuns
2. Pe baza numărului de preferințe pentru fiecare variantă dintre cele 6, am calculat procentul de succes pentru fiecare variantă prin comparaţie cu rezultatele corecte, de referinţă, provenind din analiza statistică a corpusului literar disponibil
3. Astfel, pentru fiecare întrebare am obținut un vector de 6 valori procentuale pentru fiecare variantă de răspuns, cel mai important fiind procentul pentru cel mai frecvent cuvânt de interes din acea întrebare
4. Acest procent de succes experimental, bazat pe răspunsurile participanților la studiu, a fost corelat/comparat cu procentul teoretic de succes obținut din analiza corpusului literar
5. Este de așteptat ca procentul de succes pentru fiecare întrebare să depindă de valoarea entropiei condiționate asociată acelui cuvânt (Ipoteza 1) și să fie puternic corelat cu procentul de succes rezultat din analiza statistică a corpusului literar (Ipoteza 2)

Pentru a înțelege pașii de mai sus, voi prezenta în continuare un exemplu numeric:

1. La întrebarea: Care credeți ca este cuvântul cel mai des folosit in limba română dupa cuvântul CEEA?, rezultatele la chestionar se prezintă ca în Figura 3.5:

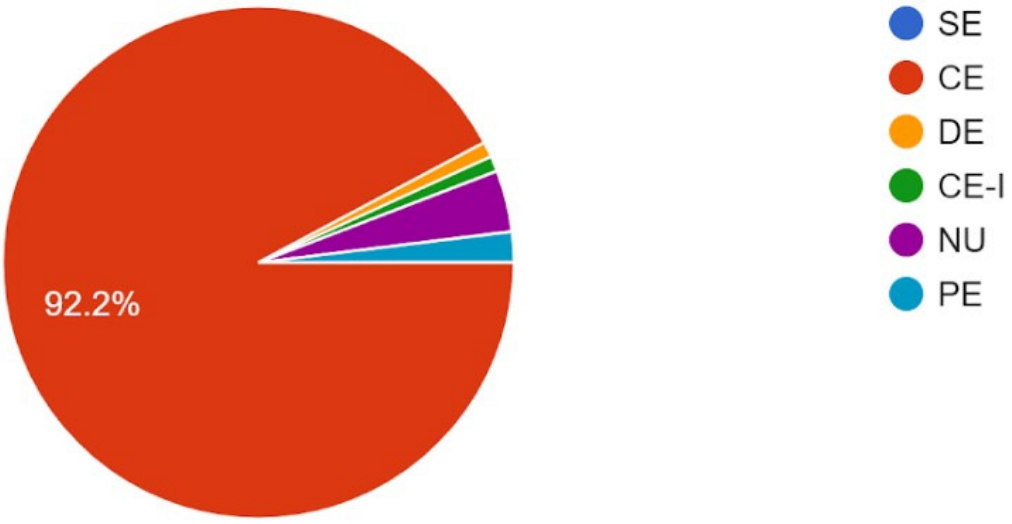

Figura 3.5 Rezultate chestionar pentru întrebarea 1 (entropie mică)

2. Pe de altă parte, din analiza statistică a corpusului (care a fost punctul de start pentru dezvoltarea chestionarului), rezultatatele arată foarte similar cu cele din Figura 3.5 - vezi Tabelul 3.1.

Tabel 3.1 Rezultatele analizei corpusului pentru cuvântul CEEA şi cele 6 cuvinte Follower ale sale (întrebarea 1)

Rang cuvânt	**Entropia condiționată (bits/cuvânt)**	**Cuvânt 1**	**Follower opțiunea 1**	**Follower opțiunea 2**	**Follower opțiunea 3**	**Follower opțiunea 4**	**Follower opțiunea 5**	**Follower opțiunea 6**
98	0.68	CEEA	CE (97.8%)	CE-I (2.04%)	SE (0.03%)	DE (0.03%)	PE (0.01%)	NU (0.01%)

3. Rezultate similare se obțin și pentru celalte nouă întrebări care abordează cuvinte de entropie mică (una dintre probabilitățile condiţionate este foarte mare).

Rezultatele prezentate în secţiunile anterioare şi în lucrarea de doctorat în detaliu, permit următoarele observaţii:

1. Ipoteza 1 a fost validată: Cu cât entropia condiționată a unui cuvânt este mai mică, cu atât este mai ușor/probabil ca un vorbitor nativ să știe ce cuvânt urmează după cuvântul de interes (coeficientul de corelație = -0.76)

2. Ipoteza 2 a fost validată: Participanții la chestionar au ales într-o mare proporţie aceleași cuvinte următoare după cele 30 de cuvinte de interes, aşa cum era așteptat din analiza statistică a corpusului literar diponibil (coeficientul de corelație = 0.84)

3. În plus, vârsta și nivelul de studii nu joacă un rol important semnificativ în scorurile totale obținute de către participanții la chestionar. Ce produce totuși o diferență între răspunsuri este genul, persoanele de gen Feminin răspunzând în medie mai bine cu o întrebare față de persoanele de gen Masculin.

4. O observație general valabilă și menționată și în partea de introducere a lucrării este că aceste cunoștințe psiholingvistice native, inconștiente şi colective sunt bine conservate pentru fiecare dintre noi și fac ca în timp, după ani și ani de practică, să alegem cuvintele care urmează și formează frazele într-un mod foarte interesant: cu sens și în același timp urmând un tipar statistic foarte bine definit.

3.2 Perspective de dezvoltare ulterioară

Posibile direcții de cercetare viitoare ar putea fi reprezentate de clarificarea rezultatelor inițiale obținute în acest studiu și anume: scorul obținut de persoanele de gen feminin este mai mare decât cel obținut de repondenții de gen masculin. Este aceasta o întâmplare, se datorează mărimii eșantionului selectat sau este o concluzie general valabilă? Și mai ales, dacă răspunsul la întrebarea anterioară este pozitiv, care este motivul subiacent?

În plus, un alt domeniu foarte interesant ar putea fi generarea de text aleatoriu, pe baza structurii de probabilități și entropii condiționate rezultate din analiza corpusului

literar disponibil, și validarea/rafinarea nivelului de sens obținut prin acest procedeu cu ajutorul unor participanți/respondenți la un studiu viitor.

Mai mult, abordarea prezentată în acest capitol ar putea fi folosită la crearea unui test care să fie folosit pentru detectarea timpurie sau monitorizarea persoanelor ce suferă de boli neurodegenrative. Există în prezent mai multe teste recunoscute internaţional care realizează acest deziderat (MMSE - Mini Mental State Examination, testul ceasului, etc.), dar un test lingvistic, bazat pe specificul limbii române ar putea furniza rezultate mai bune pentru populaţia din România.

Capitolul 4

Criptografie şi teoria haosului

Acest ultim capitol al tezei de doctorat propune o schimbare de paradigmă de la secţiunile anterioare. Focusul va fi pe domeniul criptografiei cu haos şi este bazat pe mai multe lucrări publicate de autor încă din timpul studiilor de licenţă [10] sau de doctorat [30–32].

4.1 Funcția cort compusă și conexiunea dintre codurile Gray și recuperarea condiției inițiale

Funcția cort este un sistem haotic discret unidimensional definit de următoarea ecuație:

$$x_{k+1} = \begin{cases} \dfrac{x_k}{p} & 0 \leq x_k \leq p \\ \dfrac{1-x_k}{1-p} & p < x_k \leq 1 \end{cases} \tag{4.1}$$

unde $p \in (0,1)$, p este parametrul funcției cort. Datorită ergodicității și sensibilității la condiția inițială x_0 și la parametrul de control p (care trebuie să fie diferit de 0,5, altfel, prin aplicarea formulei (4.1) cu p = 0,5, obținem o secvență care nu mai este haotică), completate de distribuția uniformă de probabilitate a valorilor x_k, funcția cort poate fi utilizată cu succes în aplicații criptografice bazate pe haos.

În majoritatea studiilor care au ca subiect funcția cort, semnalul haotic este folosit ca generator de secvențe de criptare [33–35]. Secvența de criptare ("cheia") este obținută prin binarizarea unei traiectorii (iterațiile succesive ale funcției cort) definite prin formula (4.1), folosind un anumit prag c:

- Dacă $x_k \leq c$, asignăm valoarea binară $b_k = 0$.
- Dacă $x_k > c$, asignăm valoarea binară $b_k = 1$.

Una dintre modalitățile de a crea criptograma este prin suma modulo 2 (simbol cu simbol) a mesajului clar și a "cheii". Așadar, întrebarea care apare este: "Poate cineva, având o parte (un șir de biți) din secvența binară ce corespunde cheii, să recupereze

condiția inițială x_0 care a generat respectiva traiectorie a funcției cort?". Dacă acest lucru este posibil, se poate reconstrui întreaga "cheie", indiferent cât de lungă este, și în acest caz, condiția inițială nu poate fi folosită ca element în cheia secretă.

Am arătat în secţiunea 4.1 că în anumite cazuri condiția inițială nu ar trebui inclusă în cheia secretă în criptografie, deoarece poate fi găsită foarte ușor. O soluție pentru a face mai dificilă găsirea acesteia este discretizarea / binarizarea cu $c = 0.5$ și prelevarea de date i.i.d. prin eşantionarea funcţii haotice folosite. O securitate suplimentară în acest sens este obținută prin aplicarea unei abordări de tip "cheie în mișcare" / running-key, așa cum a fost avansată pentru funcţia cort în [36]. Această abordare nu a fost folosită în această investigaţie.

Ultima parte a secţiunii 4.1 se concentrează pe problema funcţiei cort compuse. Am găsit o formulă generală relativ simplu de înţeles pentru a descrie $f_n(x)$. În plus, am arătat de ce se poate găsi x_0 dintr-un număr mic de doar 16 încercări, un rezultat enunțat în literatură, dar nu atât de convingător dovedit. Acest rezultat este valabil când pragul de binarizare e egal cu parametrul funcţiei cort ($c = p$), atunci când se iau în considerare iteraţii succesive, fără a eşantiona funcţia cort. Mai mult, în acest caz, am observat o subtilă conexiune cu codurile Gray. Descrierea funcţiei compuse aduce un suport suplimentar ideii că eşantionarea funcţiei cort ar putea fi un obstacol serios în recuperarea condiției inițiale.

Pot apărea unele dificultăți atunci când cineva dorește să utilizeze formula noastră pentru funcţia cort compusă. Acestea sunt în principal legate de ordinea în care sunt efectuate operațiile care definesc funcţia compusă. Cu toate acestea, acest nou semnal - funcţia cort compusă - are aceleași proprietăți ca (4.1) (aceeași lege uniformă de probabilitate; distanța minimă de eșantionare a independenței statistice scade cu ordinul m al funcţiei cort compuse, de exemplu, pentru $p = 0.4$ și $f_{15}(x)$, valorile succesive sunt practic i.i.d.) și poate fi o sursă de cercetare nouă în acest domeniu.

4.2 Sistemul haotic Lorenz, independența statistică și frecvența de eșantionare

Scopul acestui studiu [30] a fost să investigheze relația dintre sistemul haotic Lorenz, independența statistică și frecvența de eșantionare/pasul de timp folosit pentru a rezolva ecuațiile Lorenz. Ipoteza noastră a fost că nu ar trebui să existe o dependență de frecvența de eșantionare și independența statistică ar trebui să fie obținută în mod similar, independent de cât de fin sau grosier sunt rezolvate ecuațiile Lorenz.

Prin intermediul testului Hi-pătrat spațial, am putut să demonstrăm că datele statistic independente pot fi extrase din spațiul soluțiilor ecuațiilor Lorenz. Mai mult, am arătat că timpul de independență de aproximativ $30s$ este independent de pasul de timp/frecvența de eșantionare utilizate în rezolvarea sistemului de ecuații diferențiale. Acesta este un re-

zultat foarte important, deoarece în practică ar putea fi mai ușor, mai eficient și mai puțin costisitor din punct de vedere computational să obținem soluții ale sistemului Lorenz cu un pas de timp mai mare/frecvență de eșantionare mai mică. Datele independente necesare pentru diverse aplicații practice pot fi apoi selectate astfel încât să îndeplinească cerințele de timp de independență.

4.3 Singularitate, observabilitate și independența statistică în contextul sistemelor haotice

Rezultatele detaliate în lucrarea de doctorat arată că există o corelație între coeficientul de observabilitate și suprapunerea între atractor și manifoldul de singularitate. Se poate observa ușor că, cu cât coeficientul de observabilitate este mai mare pentru o anumită variabilă de stare, cu atât suprapunerea dintre regiunea de singularitate și atractorul acelui sistem haotic este mai mare. În practică, sistemele sunt alese astfel încât suprapunerea să fie cât mai mică posibil, deoarece acest lucru permite mai multă flexibilitate în timpul aplicării datelor în diferite aplicații criptografice. În general, din punctul de vedere al observabilității-singularității, din cele trei sisteme analizate, Ikeda pare cel mai promițător, urmat de Tinkerbell și Clifford. Cu toate acestea, singularitatea și observabilitatea nu sunt singurele concepte care contează. Din punct de vedere al independenței statistice, harta Clifford este singura capabilă să fie utilizată ca generator pseudoaleator (PRNG).

În concluzie, acest demers de cercetare prezintă o procedură nouă de analiză pentru hărți dinamice, având în vedere concepte esențiale precum independența statistică, singularitatea și observabilitatea. Marele avantaj al fluxului de gândire propus în acest articol este că nu se bazează exclusiv pe o singură noțiune, oricât de puternică ar fi aceea. Singularitatea, observabilitatea și independența statistică au fost abordate separat în literatură, și este un fapt cunoscut că acestea pot oferi, separat, perspective importante asupra problemelor specifice de interes în criptografie. Cu toate acestea, folosirea rezultatelor din toate cele trei perspective diferite considerate împreună duce la o concluzie convergentă și mai puternică, care poate fi utilizată și investigată mai departe de către cercetătorii din domeniul sistemelor dinamice cu aplicații criptografice.

Abordarea propusă poate fi utilizată pentru orice sistem, independent de numărul de variabile de stare. Rezultatele experimentale arată că există un echilibru fragil între conceptele care pot fi utilizate pentru a selecta un sistem pentru utilizare în criptografie, nu există un concept de tipul "one size fits all", și există mai degrabă un compromis.

Capitolul 5

Concluzii şi lista publicaţiilor originale

5.1 Concluzii

Lingvistica, psihologia şi ingineria par subiecte fără o legătură aparentă între ele, însă am demonstrat cum concepte de teoria informaţiei şi statistică reuşesc să dea răspunsuri interesante la probleme de tip umanist.

Capitolul 2 se axează pe studiul limbii române scrise şi analiza din mai multe perspective a corpusului literar disponibil în colectivul de cercetare [19, 14].

Studiul prezentat în capitolul 3 și-a propus să investigheze și să explice într-un mod simplu și vizual legătura între concepte la îndemâna fiecăruia dintre noi, pe care le folosim clipă de clipă (cuvintele și frazele cu sens pe care le utilizăm pentru a ne înțelege unii cu ceilalți), chiar fără a conștientiza mereu acest lucru, și procedee și mărimi matematice foarte complexe, la o primă vedere (entropia, probabilitățile condiționate sau valori medii, variațe și corelații).

Analiza bazată pe interdisciplinaritate şi ubicuitate este continuată şi în capitolul 4, când concepte de criptografie sunt abordate statistic şi chiar sunt conectate cu preocupări fundamentale din domeniul fizicii sau transdisciplinarităţii. Am arătat că în anumite cazuri condiția inițială nu ar trebui inclusă în cheia secretă utilizată în criptografie, deoarece poate fi găsită foarte ușor. O soluție pentru a face mai dificilă găsirea acesteia este discretizarea / binarizarea cu $c = 0.5$ și prelevarea de date i.i.d. prin eşantionarea hărţii haotice folosite. Scopul studiului din secţiunea 4.2 a fost să investigheze relația dintre sistemul haotic Lorenz, independența statistică și frecvența de eșantionare/pasul de timp folosit pentru a rezolva ecuațiile Lorenz [30]. Având în vedere rezultatele prezentate în Secțiunea 4.3, pot fi trase mai multe concluzii importante cu privire la procedura propusă pentru testarea unificată a diferitelor sisteme haotice și selectarea generatoarelor pseudoaleatoare (PRNG). În general, din punctul de vedere al observabilității-singularității, din cele trei sisteme analizate, Ikeda pare cel mai promițător, urmat de Tinkerbell și Clifford. Cu toate acestea, singularitatea și observabilitatea nu sunt singurele concepte care contează. Din punct de vedere al independenței statistice, sistemul Clifford este

singurul capabil să fie utilizat ca generator pseudoaleator (PRNG). Acest demers de cercetare prezintă o procedură nouă de analiză pentru hărți dinamice, având în vedere concepte esențiale precum independența statistică, singularitatea și observabilitatea. Marele avantaj al fluxului de gândire propus în acest articol este că nu se bazează exclusiv pe o singură noțiune, oricât de puternică ar fi aceea. Singularitatea, observabilitatea și independența statistică au fost abordate separat în literatură, și este un fapt cunoscut că acestea pot oferi, separat, perspective importante asupra problemelor specifice de interes în criptografie. Cu toate acestea, folosirea rezultatelor din toate cele trei perspective diferite considerate împreună duce la o concluzie convergentă și mai puternică, care poate fi utilizată și investigată mai departe de către cercetătorii din domeniul sistemelor dinamice cu aplicații criptografice. Abordarea propusă poate fi utilizată pentru orice sistem, independent de numărul de variabile de stare. Rezultatele experimentale arată că există un echilibru fragil între conceptele care pot fi utilizate pentru a selecta un sistem pentru utilizare în criptografie, nu există un concept de tipul "one size fits all", și vorbim mai degrabă un compromis în funcţie de aplicaţia de interes.

5.2 Publicaţii originale

Rezultatele incluse în această lucrare de doctorat au fost publicate in revistele şi la conferinţele menţionate mai jos (o publicaţie Q1 şi una de tip Q3 incluse în totalul de 5 publicaţii WOS şi 6 BDI).

5.2.1 Articole de revistă

1. Dinu, A. and Frunzete, M. (2023b). Observability and statistical independence in the context of chaotic systems. Mathematics, 11(2) - **WOS indexed, CCC: 000916377100001, Q1 journal**

2. Dinu, A. and Vlad, A. (2014). The compound tent map and the connection between Gray codes and the initial condition recovery. UPB Sci. Bull. Ser. A Appl. Math. Phys, 76(1), **WOS:000332914700002, Q3 journal**.

3. Dinu, A. and Frunzete, M. (2023a). Determinism and chaos – a story about Big Bang, singularity and the future of mankind. Ann Math Phys 6(1): 041-043. DOI: 10.17352/amp.000075.

5.2.2 Conferinţe

1. Dinu, A., Vlad, A., Hanu, B., and Mitrea, A. (2020a). Beginning and end of sentence word digrams for printed romanian language. Proceedings of the 15th International Conference Linguistic Resources and Tools for Natural Language Processing, pages 53–63, ISSN 1843-911X, **WOS:000659362800005**

2. Dinu, A., Vlad, A., Mitrea, A., and Hanu, B. (2020b). The statistical independence for words in printed Romanian language. 13th International Conference on Communications (COMM2020), pages 319-324, Bucharest, 2020, **WOS:000612723900056**

3. Alexandru Dinu and Adriana Vlad. Romanian printed language, statistical independence and the type II statistical error. In International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pages 120–125, Bucarest, 2021, **WOS:000786794700022**

4. Dinu, A. and Vlad, A. (2022). Revisiting the idea of a representative linguistic corpus. 14th International Conference on Communications (COMM2022), Bucharest, Romania, 2022, pages 1–5, DOI: 10.1109/COMM54429.2022.9817208

5. Dinu, A. (2020b). Psycholinguistics, statistics and the unconscious mind. Conferinţa Internaţională Educaţie şi Creativitate pentru o Societate Bazată pe Cunoaştere - Psihologie, Bucureşti, Universitatea Titu Maiorescu, 2020, pages 190–195, ISSN 2248-003X, ISBN 978-3-9503145-6-4

6. Dinu, A., Vlad, A., Hanu, B., and Mitrea, A. (2019). Revisiting the statistical independence for the printed Romanian language. Proceedings of the 14th International Conference Linguistic Resources and Tools for Natural Language Processing, pages 99–113, ISSN 1843-911X

7. Hanu, B., Vlad, A., Dinu, A., and Mitrea, A. (2019). Looking along Zipf's Law for the distribution of words beginning and ending sentences in literary printed Romanian corpora. Proceedings of the 14th International Conference Linguistic Resources and Tools for Natural Language Processing, pages 51–63, ISSN 1843-911X

8. Dinu, A. and Frunzete, M. (2021). The Lorenz chaotic system, statistical independence and sampling frequency. 2021 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, pages 1–4, DOI: 10.1109/ISSCS52333.2021.9497431

5.2.3 Rapoarte de cercetare şi alte publicaţii

1. 4 rapoarte de cercetare științifică din cadrul SD-ETTI:

   - Raportul științific nr. 1/2019, "The Chi-square test in contingency tables and its applications for natural language processing"
   - Raportul științific nr. 2/2019, "Revisiting the statistical independence for the printed Romanian language"

- Raportul ştiinţific nr. 3/2020, "The statistical independence for words in printed Romanian language"
- Raportul ştiinţific nr. 4/2020, "The Lorenz chaotic system, statistical independence and sampling frequency"

2. Două rapoarte de cercetare din cadrul subprogramului 3: "Analiză privind grupurile de două cuvinte succesive de început şi de final de frază/propoziţie, pe corpus literar de limbă română scrisă cu ortografie şi punctuaţie", al programului de cercetare "Resurse şi tehnologii pentru limba română în context multilingv standardizat", la Institul de cercetări pentru inteligenţă artificială "Mihai Drăgănescu" al Academiei Române, în care am fost implicat împreună cu colectivul de cercetare coordonat de doamna profesor Adriana Vlad (2019-2020).

3. Dinu, A. (2020a). Psycholinguistics, statistics and the unconscious mind. Lucrare licenţă în Psihologie susţinută la Universitatea Titu Maiorescu, Bucureşti.

Bibliografie

[1] I.B. Djordjevic. Markov chain-like quantum biological modeling of mutations, aging, and evolution. *Life*, 5(3):1518–1538, 2015.

[2] F. Alcántara-López, C. Fuentes, C. Chávez, J. López-Estrada, and F. Brambila-Paz. Fractional growth model with delay for recurrent outbreaks applied to covid-19 data. *Mathematics*, 10(5), 2022.

[3] A. Finnemann, D. Borsboom, S. Epskamp, and H.L.J. van der Maas. The theoretical and statistical ising model: A practical guide in r. *Psych*, 3(4):593–617, 2021.

[4] L. Kolbe, F. Oort, and S. Jak. Bivariate distributions underlying responses to ordinal variables. *Psych*, 3(4):562–578, 2021.

[5] T.D. Martinho. Researching culture through big data: Computational engineering and the human and social sciences. *Social Sciences*, 7(12), 2018.

[6] K. Loeber. Big data, algorithmic regulation, and the history of the cybersyn project in chile, 1971–1973. *Social Sciences*, 7(4), 2018.

[7] K. Sell, F. Hommes, F. Fischer, and L. Arnold. Multi-, inter-, and transdisciplinarity within the public health workforce: A scoping review to assess definitions and applications of concepts. *International Journal of Environmental Research and Public Health*, 19(17), 2022.

[8] C.N. Knapp, R.S. Reid, M.E. Fernández-Giménez, J.A. Klein, and K.A. Galvin. Placing transdisciplinarity in context: A review of approaches to connect scholars, society and action. *Sustainability*, 11(18), 2019.

[9] A. Classen. Transdisciplinarity—a bold way into the academic future, from a european medievalist perspective and or the rediscovery of philology? *Humanities*, 10(3), 2021.

[10] A. Dinu and A. Vlad. The compound tent map and the connection between gray codes and the initial condition recovery. *UPB Sci. Bull. Ser. A Appl. Math. Phys*, 76(1), 2014.

[11] A. Dinu, A. Vlad, B. Hanu, and A. Mitrea. Revisiting the statistical independence for the printed Romanian language. *Proceedings of the 14th International Conference Linguistic Resources and Tools for Natural Language Processing*, pages 99–113, 2019.

[12] A. Dinu, A. Vlad, A. Mitrea, and B. Hanu. The statistical independence for words in printed Romanian language. *13th International Conference on Communications (COMM)*, pages 319–324, 2020.

[13] A. Dinu and A. Vlad. Romanian printed language, statistical independence and the type ii statistical error. In *International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 120–125, 2021.

[14] A. Vlad, A. Mitrea, A. Luca, and O. Hodea. Considerations regarding the statistical compatibility of two romanian literary corpora with orthography and punctuations marks included. *Towards Multilingual Europe 2020: A Romanian Perspective, Dan Tufis, Vasile Rus, Corina Forascu Eds., The Publishing House of the Romanian Academy*, pages 99–122, 2013.

[15] C.E. Shannon. Prediction and entropy of printed english. *Bell Syst. Tech. J.*, 30:50–64, 1951.

[16] V. Craiu. Verificarea ipotezelor statistice. *Editura Didactica si Pedagogica, Bucuresti, Romania*, 1972.

[17] R. Walpole, R. Myers, S. Myers, and K. Ye. Probability & statistics for engineers & scientists, Mylab statistics update (9th edition). *Pearson*, 2016.

[18] J. Devore. Probability and statistics 7th (seventh) edition. *Duxburry Press*, 2008.

[19] A. Vlad, Mitrea A., and M. Mitrea. Limba română scrisă ca sursă de informație. *Editura Paideia*, 2003.

[20] A. Vlad, A. Mitrea, and M. Mitrea. Information sources approximating to printed romanian: The role of type ii statistical error. In *Proceedings of the Romanian Academy, Series A*, pages 329–337, 2004.

[21] A. Vlad and A. Mitrea. Contribuţii privind structura statistică de cuvinte in limba română scrisă. *Limba Română în Societatea Informaţională - Societatea Cunoaşterii. Editura Expert, Bucuresti, Romania, 2002*, pages 209–236, 2002.

[22] S. Ciuca, A. Vlad, and A. Mitrea. A mathemathical comparison between several single author corpora. *U.P.B. Sci. Bull., Series A, Vol. 74, Iss. 1, 2012*, 2012.

[23] A. Mitrea, A. Vlad, and A. Luca. On the occurrences of two successive words în a literary romanian corpus. *Proc. of The 8th International Conference on Communications "COMM 2010", June 10-12, 2010, Bucharest*, pages 115–118, 2010.

[24] A. Mitrea, A. Vlad, and A. Luca. Statistical study on a literary romanian corpus for the beginning and ending of the words. *Proc. 9th IEEE International Conference on Communications (COMM 2012), Bucharest*, pages 81–84, 2012.

[25] A. Mitrea, A. Vlad, O. Hodea, and R. Dragomir. A study on the common words found in different literary romanian corpora. *Proc. 10th IEEE International Conference on Communications (COMM 2014), Bucharest*, pages 123–127, 2014.

[26] B. Say and V. Akman. Current approaches to punctuation în computational linguistics. *Computer and the Humanities*, 30:457–469, 1997.

[27] A. Dinu. Psycholinguistics, statistics and the unconscious mind. *Lucrare de licenţă în Psihologie susţinută la Universitatea Titu Maiorescu, Bucureşti*, 2020.

[28] A. Dinu. Psycholinguistics, statistics and the unconscious mind. *Conferinţa Internaţională Educaţie şi Creativitate pentru o Societate Bazată pe Cunoaştere – PSIHOLOGIE, Bucureşti, Universitatea Titu Maiorescu, 2020*, pages 190–195, 2020.

[29] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[30] A. Dinu and M. Frunzete. The lorenz chaotic system, statistical independence and sampling frequency. *2021 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania*, pages 1–4, 2021.

[31] A. Dinu and M. Frunzete. Determinism and chaos – a story about big bang, singularity and the future of mankind. *Ann Math Phys 6(1): 041-043. DOI: 10.17352/amp.000075*, 2023.

[32] A. Dinu and M. Frunzete. Observability and statistical independence in the context of chaotic systems. *Mathematics*, 11(2), 2023.

[33] D. Arroyo. Framework for the analysis and design of encrypt. PhD Thesis, 2009.

[34] A. Ilyas, A. Luca, and A. Vlad. A study on binary sequences generated by tent map having cryptographic view. *In Proc.9thInternational conference on Communications (COMM), Bucharest, June 21-23*, pages 23–26, 2012.

[35] A. Luca, A. Ilyas, and A. Vlad. Generating random binary sequences using tent map. *In Proc.10th.International Symposium on signals, Circuits and Systems (ISSCS), Iasi, Romania, June 30-July 1*, pages 81–84, 2011.

[36] A. Vlad, A. Luca, O. Hodea, and R. Tataru. Generating chaotic secure sequences using tent map and a running-key approach. *PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A*, 14:295–302, 2013.