POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 1046 **from** 10-07-2023

Ph.D. THESIS SUMMARY

Alexandru DINU

MODELE ŞI METODE STATISTICE UTILIZATE ÎN STUDIUL ŞTIINŢELOR VIEŢII

STATISTICAL MODELS AND METHODS USED IN LIFE SCIENCES

THESIS COMMITTEE

Prof. Dr. Ing. Gheorghe BREZEANU Politehnica Univ. of Bucharest	President
Prof. Dr. Ing. Adriana VLAD Politehnica Univ. of Bucharest	PhD Supervisor
Prof. Dr. Ing. Corneliu BURILEANU Politehnica Univ. of Bucharest	Referee
Maître de conf*é*rences Mihai MITREA T*é*l*é*com Sud-Paris	Referee
Prof. Dr. Ing. Victor Adrian GRIGORAŞ Tehnical Univ. "Gheorghe Asachi" Iaşi	Referee

BUCHAREST 2023

Table of contents

Chapter 1

Introduction

The present doctoral thesis addresses a fertile and not enough explored field of study: life sciences viewed and analyzed from an interdisciplinary statistical perspective. By "life sciences," we mean any field of study, research, and knowledge related to life. This category includes biology, zoology, medicine, anthropology, psychology, sociology, and many others, either individually or in various combinations [1–3]. As it can be easily observed, the field of life sciences can be quite vast [4–6].

It will become evident in the following chapters that the topics addressed in this doctoral work share a common statistical and mathematical approach but vary greatly, ranging from the field of psychology combined with elements of information theory to the study of chaotic dynamic systems from a ubiquitous perspective. I believe that our progress and evolution as a species will occur only by considering and utilizing the benefits of ideas such as interdisciplinarity, multidisciplinarity, or transdisciplinarity, as more and more authors have recently begun to notice [7, 8]. It is astonishing how many similarities exist between the solutions found by researchers from seemingly unrelated fields of knowledge [9].

During my undergraduate studies at the University Politehnica of Bucharest, under the guidance of Professor Adriana Vlad, I became passionate about the field of cryptography and chaotic signals. Therefore, it is not surprising that the most important results from my undergraduate thesis were extended and published in the UPB Bulletin [10].

Upon beginning the doctoral studies, my research interest underwent slight modifications, shifting from the area of chaotic signals to psycholinguistics and statistics applied to written Romanian language. This shift aligns with the broader research theme chosen for the doctoral studies: models and statistical methods used in the study of life sciences, as well as with the complementary studies in psychology completed concurrently with doctoral studies.

Chapter 2 addresses the topic of written Romanian language analyzed from a dual perspective: psychological and statistical. In Section 2.1, based on the work presented at CONSILR 2019 [11], we revisited the concept of statistical independence for written Romanian language, using an existing literary corpus within the research group. The primary objective was to improve the perception and understanding of the concept of statistical independence for natural language and to use this concept to numerically evaluate the properties of written language. Section 2.2, based on the work presented at COMM 2020 [12], extends the initial results presented in 2.1 from n-grams to words. We invoked the notion of probability for Type II statistical error, which was extensively analyzed in [13]. Section 2.3 continues the investigation into statistical independence for written Romanian language seen as a chain of words. The study presented in section 2.4, advances the research conducted by the team led by Prof. Dr. Eng. Adriana Vlad regarding the word structure in Romanian. During this phase of research, results were obtained concerning digrams (successive groups of 2 words) occurring at the end of sentences/propositions, and the results regarding the beginning-of-sentence word digrams were enriched with additional explanations. New perspectives regarding Zipf's Law and analysis on an author-specific subcorpus were also investigated. Section 2.5 revisits the notion of a representative corpus for a linguistic corpus obtained by merging distinct literary corpora.

Chapter 3 continues the investigation initiated in Chapter 2, adding a new perspective: the human, psychological aspect of written language. This endeavor aims to explore the connection between psycholinguistic knowledge acquired by each of us throughout our lives and the validation and correlation of this knowledge with certain mathematical concepts and information theory: entropy and conditional probabilities.

Chapter 4 addresses a different research direction - chaos-based cryptography. The research included in this final chapter of the doctoral thesis attempts to provide a unifying vision of several perspectives coming from a very diverse set of disciplines: biology, genetics, economics, and cryptography, which appear to work in parallel to solve the same problem. All these transdisciplinary approaches from these domains aim to find a theory of everything, a unifying theory that brings order to chaos, sheds light on darkness, and accurately predicts the future based on past or present facts.

Chapter 5 includes the main conclusions of the doctoral thesis and summarizes the original results obtained during the PhD studies.

Chapter 2

Psycholinguistics - the study of written Romanian from a statistical and psychological perspective

2.1 Statistical independence for written Romanian - a case study for m-grams

2.1.1 Introduction

The new experimental results obtained concerning the minimum distance for statistical independence are illustrated for the case of letter m-grams (letters, digrams, trigrams) in a linguistic corpus consisting of 49 books written by 9 authors as follows: Isaac Asimov - 9 books; Constantin Chiriță - 5 books; Alexandre Dumas - 12 books; Colin Falconer - 1 book; Frank Herbert - 8 books; Niven Larry - 1 book; Orson Scott Card - 3 books; Michel Zevaco - 7 books; J. R. Tolkien - 3 books. The size of the corpus, including spelling and punctuation, is 36,898,820 characters (over 6 million words). The books were concatenated randomly to form the analyzed corpus [14].

In previous research by the authors, it was assumed that approximately 200 characters are sufficient to ensure statistical independence for written Romanian. This hypothesis had not been fully tested until we decided to revisit this important topic, but previous indirect results supported its validity.

2.1.2 Data Collection and Theoretical Considerations

Data Collection

The main question is: **what is the distance/number of characters between an m-gram and the next one, such that these two linguistic entities are not dependent/correlated with each other?** In other words, knowing $m - gram_1$, do we have any information about the m-gram that is d characters away, $m - gram_2$? If the answer is NO, then statistical independence has been achieved. Note: the m-gram located d characters away means that the difference between the index of the first letter in $m - gram_1$ and the index of the first letter in $m - gram_2$ is exactly d. The above can be mathematically expressed in (2.1) or (2.2):

$$P(m - gram_2 | m - gram_1) = P(m - gram_2) \tag{2.1}$$

$$P(m - gram_1, m - gram_2) = P(m - gram_1) \cdot P(m - gram_2) \tag{2.2}$$

The notion of statistical independence can also be easily explained based on the visualization in Table 2.1. If statistical independence is achieved for $d = 6$, then the probability that letter E follows 6 characters after letter A (or after any letter) would be identical to the probability of the letter E: $P(E|A) = P(E)$. This is essentially the translation of (2.1) for the case of unigrams/letters. Similarly, for digrams and assuming we are interested in the $d = 6$ case, $P(ER|AC) = P(ER)$ or $P(ERG|ACU) = P(ERG)$.

Table 2.1 Examples of digrams and trigrams in natural language.

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

Information Theory

An approach we used to assess the minimum distance for statistical independence is derived from the field of information theory. M-grams are collected from the written text as visualized in Table 2.1, Figure 2.1, and Figure 2.2.

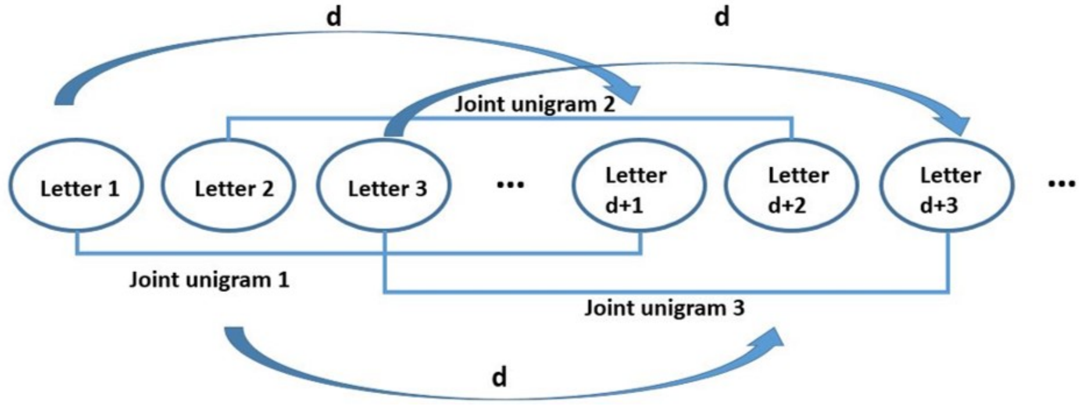


Fig. 2.1 Collecting letters from the corpus, without spacing

Figure 2.3 presents the visualization of the contingency table when analyzing letters, with letters ordered from most frequent in the corpus to least frequent (top-down and left-right).

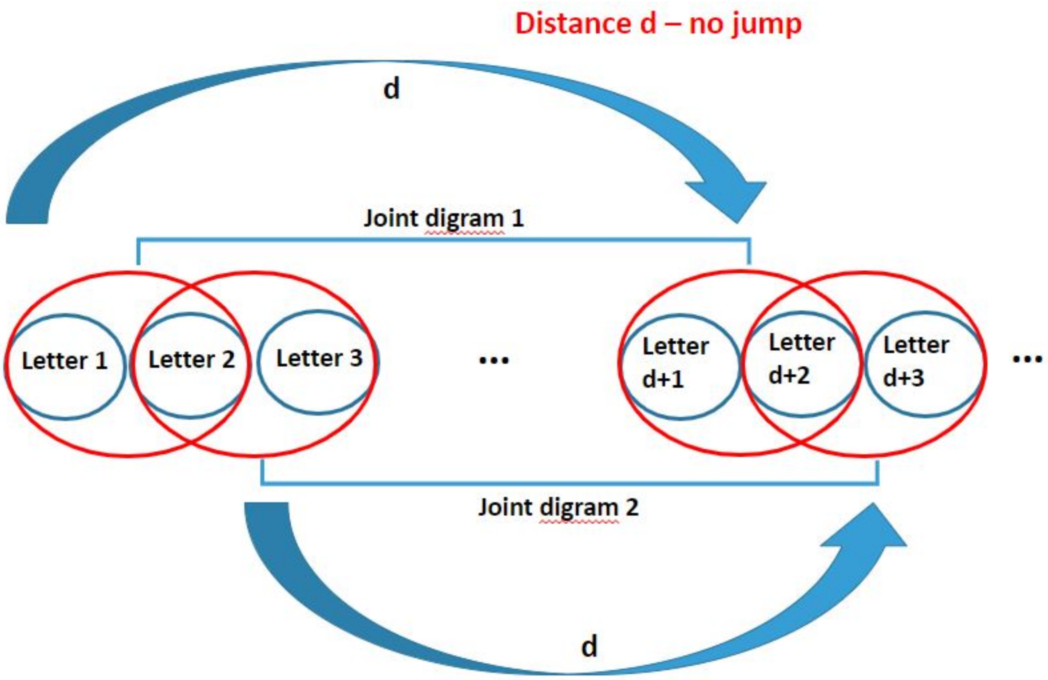
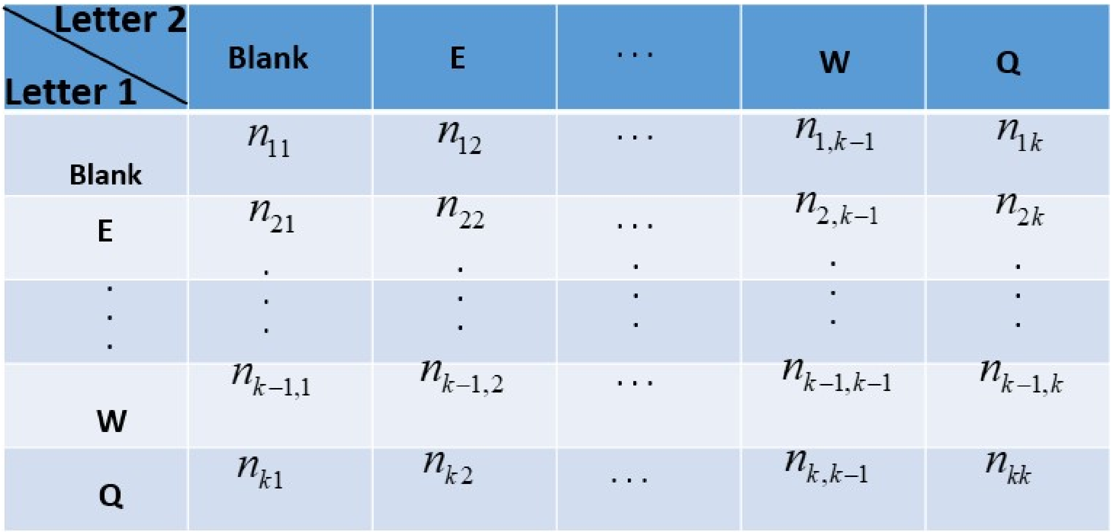


Fig. 2.2 Collecting digrams from the corpus, without spacing



Letter 2 / Letter 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}

Fig. 2.3 Occurrences for pairs of letters separated by distance d

Figure 2.3 corresponds to the information/noise channel for distance d, where the m-grams on the rows correspond to channel inputs, and the m-grams on the columns correspond to outputs [15]. By dividing these two quantities, an estimation for the conditional probability of a specific pair of m-grams separated by distance d can be obtained (see (2.3)).

$$P(m-gram_2|m-gram_1) \approx \frac{Occurrences \text{ for pair } (m-gram_1, m-gram_2)}{Occurrences \text{ for } m-gram_1} \tag{2.3}$$

When equations (2.1) and (2.2) are satisfied, statistical independence is achieved, and the value of transinformation becomes 0. A quick way to assess whether statistical independence has been achieved is to visually inspect the table of conditional probabilities. It has the same structure as presented in Figure 2.3. The deviation of each conditional probability from the unconditioned value is expressed in (2.4). When independence is reached, we expect the error matrix to be very close to 0.

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

Table 2.2 presents the conditional probabilities versus the unconditioned probabilities for $d = 13$. In fact, Table 2.2 numerically presents what Figure 2.5 shows through the associated color code. The two types of probabilities should be equal when statistical independence is achieved. It can be easily observed that the approximation between the two types of probabilities is more visible for symbols in the top-left corner (frequent symbols), while the largest errors are in the bottom-right corner (the rarest letters).

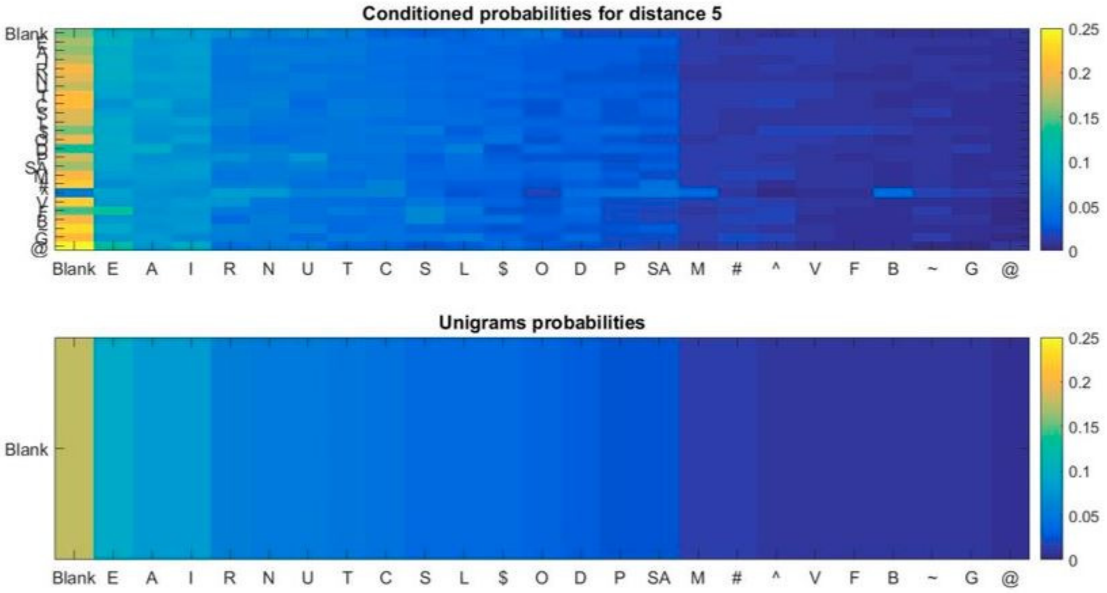


Fig. 2.4 Conditional probabilities for letters $P(Letter_2|Letter_1)$ for $d = 5$ (X-axis - $Letter_1$; Y-axis - $Letter_2$)

Comment on Table 2.2: the conditional probability that letter A follows letter E after 13 letters is 8.02% (we began the analysis with $d = 13$ based on the results from the section using the Chi-squared test-based approach. SA is the artificial symbol obtained by concatenating the least frequent 8 characters from the corpus alphabet.

The data from Table 2.2 was used to calculate the errors between conditional and unconditional probabilities based on (2.4). The result can be visualized in Figure 2.6 for $d = 5$ and Figure 2.7 for $d = 13$. When statistical independence is achieved, the error matrix should become 0, and this should correspond to a completely blue matrix according to the color code used.

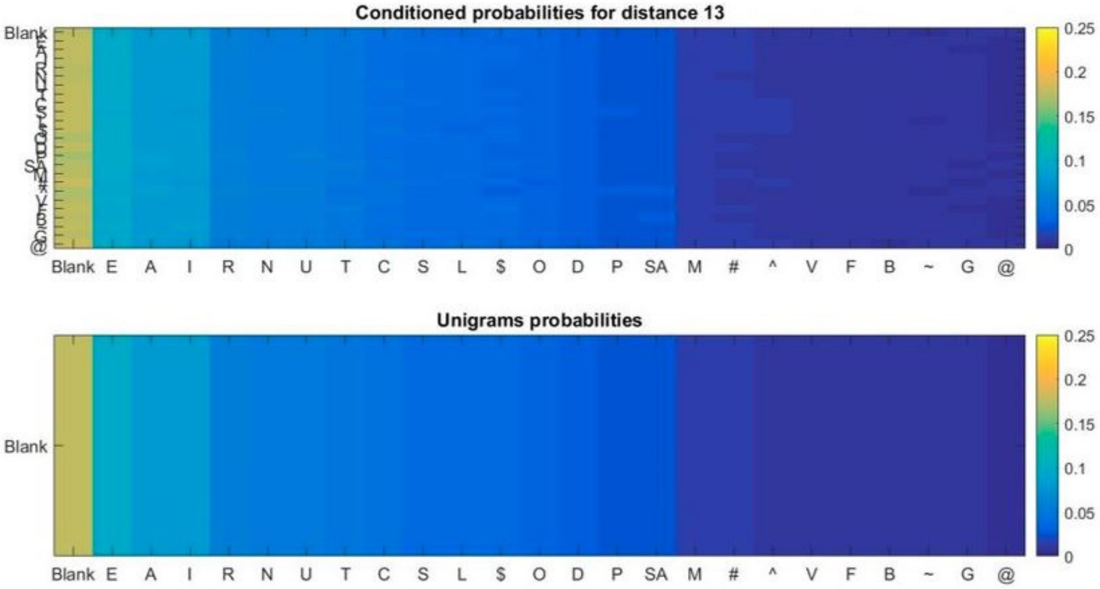


Fig. 2.5 Conditional probabilities for letters $P(Letter_2|Letter_1)$ for $d = 13$ (X-axis - $Letter_1$; Y-axis - $Letter_2$)

Table 2.2 Conditional probabilities for letters $P(Letter_2|Letter_1)$ for $d = 13$ (rows - $Letter_1$; columns - $Letter_2$)

Letters/Probabilities	**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 (artificial symbol)	17.53	9.48	8.30	...	2.63	...	0.79
...	...	...	...	...	...	...	...
â	18.03	9.73	7.92	...	2.38	...	0.86
Column header letter probability	17.63	9.54	8.09	...	2.57	...	0.75

It can be observed that as the distance $d = 5$ increases towards $d = 13$ and beyond, the errors become much smaller (Figure 2.7), with a few more colored areas (large errors between conditional and unconditional probabilities - Figure 2.6) still present at a distance of $d = 5$.

Chi-square Test in Contingency Tables

This test provides an overview of the independence situation and relies on the representation in Figure 2.3 (and implicitly on the noise/error matrix associated with it), and, together with the approach based on information theory, leads to a faster response regarding the issue of statistical independence.

A very important requirement for the Chi-square test is that experimental data must come from independent and identically distributed random variables, the test's sensitivity

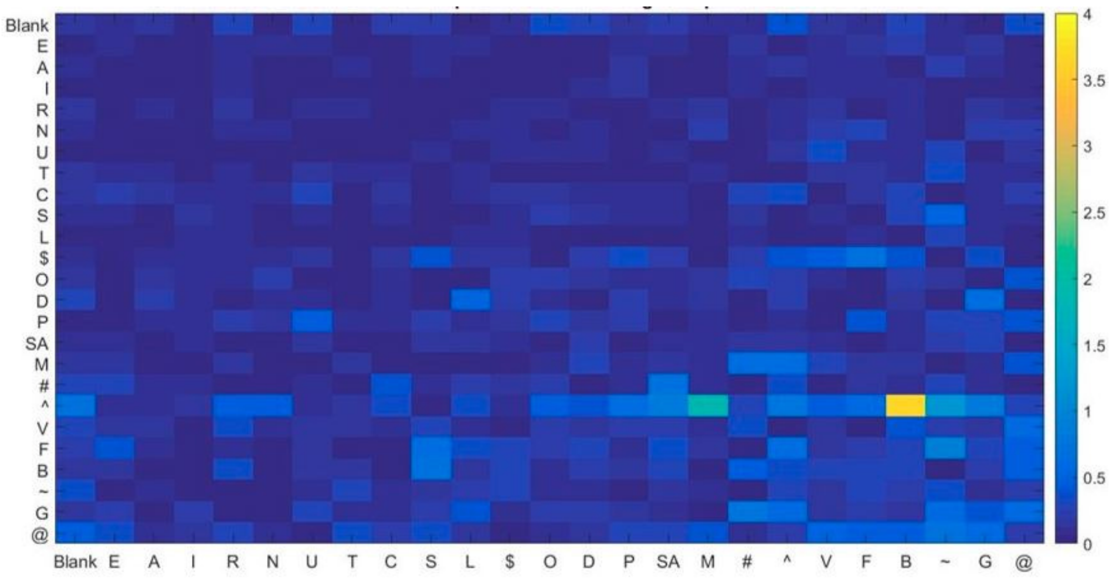


Fig. 2.6 Error matrix (%) for letters for $d = 5$

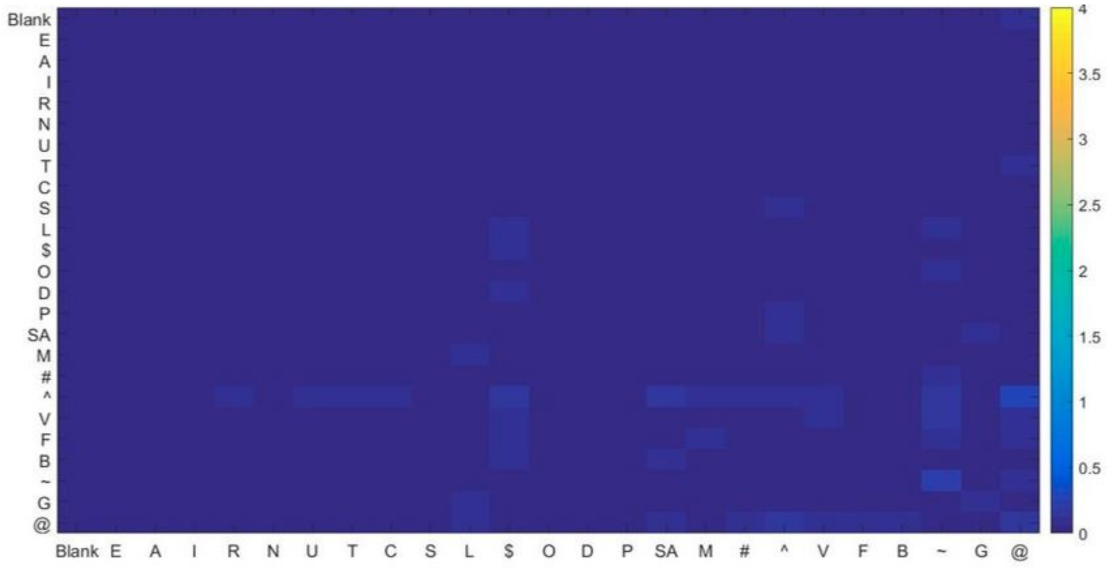


Fig. 2.7 Error matrix (%) for letters for $d = 13$

to i.i.d. data being a well-known aspect in the literature [16]. Figure 2.8 presents the procedure used to collect the i.i.d. data from the corpus needed for the Chi-square test.

The condition regarding the minimum number of occurrences of an m-gram applies in this case. When the DeMoivre-Laplace condition is not met, the corresponding rows and columns are concatenated into an artificial symbol, thus reducing the size of the contingency table. For example, in the case of letters, where the corpus alphabet has a size of 32, the reduced contingency table has a smaller number of symbols, specifically 25. The algorithm used to reduce the size of the contingency table can be visualized in Figure 2.9.

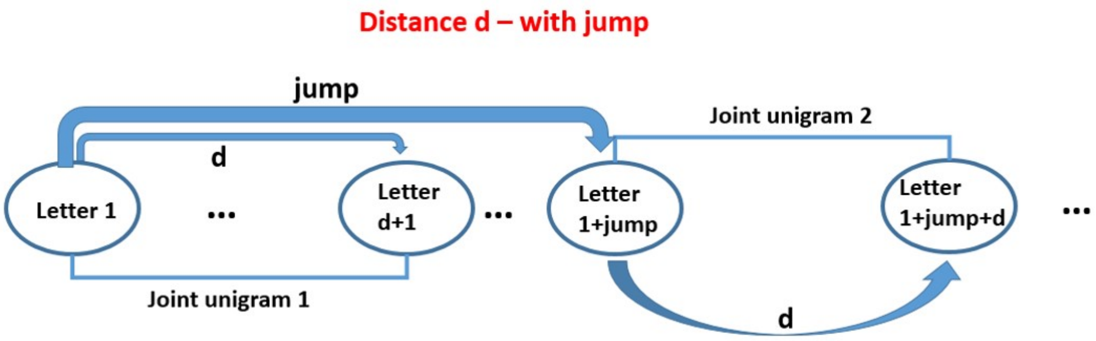
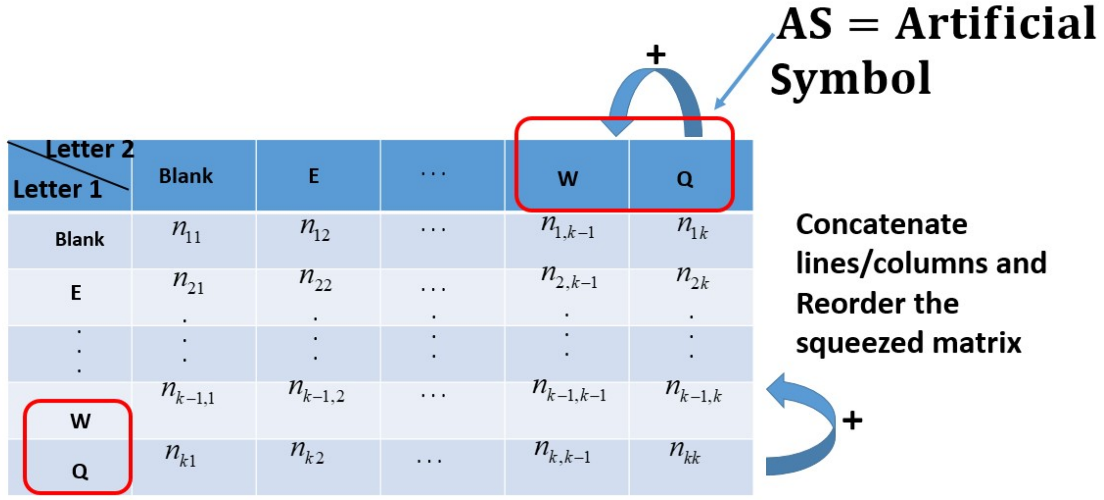


Fig. 2.8 Collecting i.i.d. (with spacing) data for the Chi-square test in contingency tables



Letter 2 / Letter 1	Blank	E	$\cdots$	W	Q
Blank	n_{11}	n_{12}	$\cdots$	$n_{1,k-1}$	n_{1k}
E	n_{21}	n_{22}	$\cdots$	$n_{2,k-1}$	n_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
W	$n_{k-1,1}$	$n_{k-1,2}$	$\cdots$	$n_{k-1,k-1}$	$n_{k-1,k}$
Q	n_{k1}	n_{k2}	$\cdots$	$n_{k,k-1}$	n_{kk}

Fig. 2.9 Reducing the contingency table for the Chi-square test

The algorithm is straightforward: for each distance d, a test value is calculated according to (2.5) and compared to an accepted/reference value (the alpha-quantile of a Chi-square distribution with $(k-1)\cdot(k-1)$ degrees of freedom, denoted as z_α).

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

If $z<z_\alpha$, we say that the Chi-Square test passes and statistical independence has been achieved; otherwise, the test does not pass, and the next distance d is checked [16–18].

2.1.3 Conclusions

The results mentioned in Section 2.1 confirm the idea that a distance of 80-100 characters is more than enough to ensure statistical independence in the case of m-grams extracted from natural texts written in the Romanian language. Additionally, the independence distance is analyzed through two methods: the Chi-square test in contingency tables and information theory.

2.2 Statistical Independence for Written Romanian Language - Case Study - Words

2.2.1 Experimental Results

Figures 2.10 and 2.11 present the results of the Chi-square test when words are collected from the corpus with a jump of 100 characters. Figure 2.10 shows the evolution of the test value z compared to z_α for different distances d. Figure 2.11 presents the size of the reduced contingency table. It is interesting to note that not only does z vary with distance, but z_α is also variable. This is logical and happens because the size of the contingency table also changes with distance.

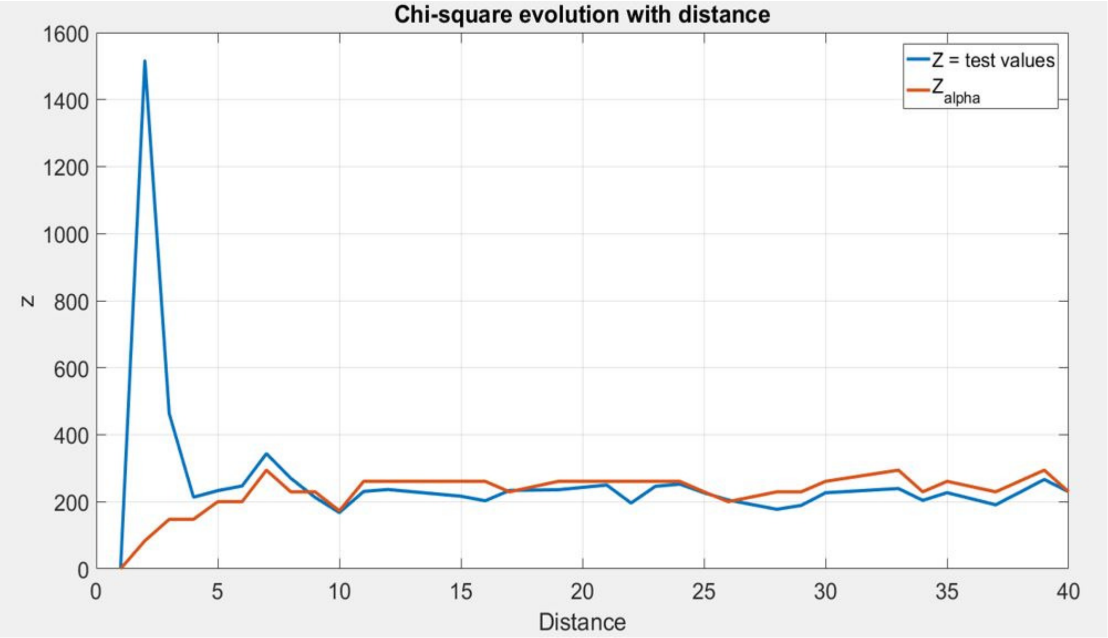


Fig. 2.10 Results of the Chi-square test with skip for words (X-axis = distance d; Y-axis - test value z - see (2.5))

A first conclusion obtained after applying the Chi-square test is that statistical independence is achieved for $d = 9$ in the case of words.

Information Theory

Figures 2.12, 2.13, and 2.14 show how as the distance d between words increases, the conditional probabilities (upper half of the figures) start to resemble the individual/unconditional probabilities (lower half of the figures). Approximately constant values for conditional probabilities indicate that statistical independence is achieved.

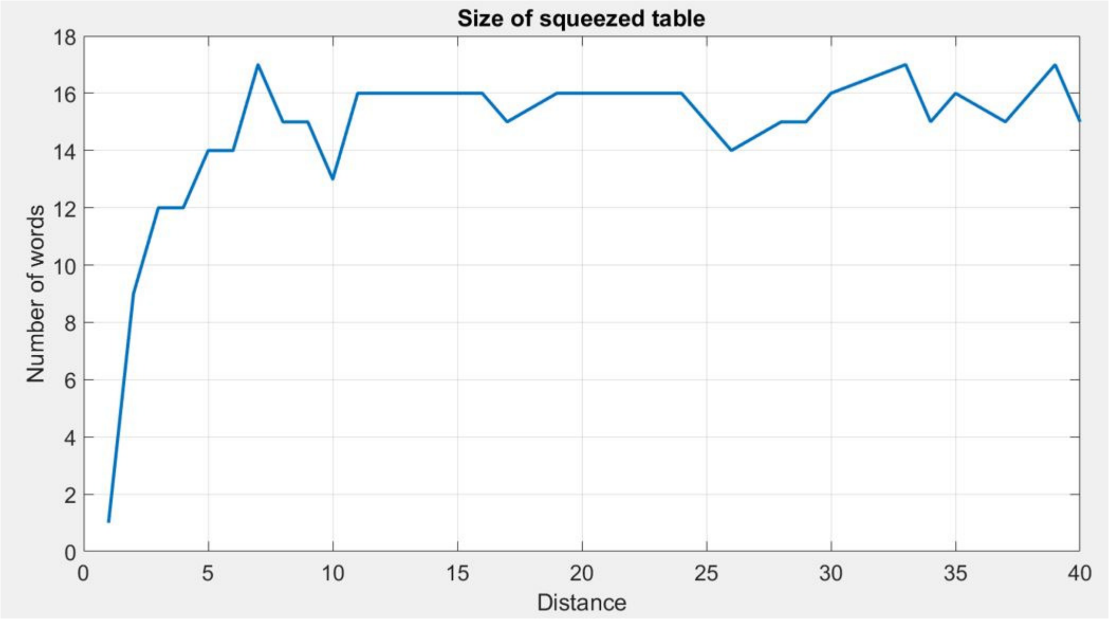


Fig. 2.11 Results of the Chi-square test with skip for words (X-axis = distance d; Y-axis - number of words in the reduced contingency table)

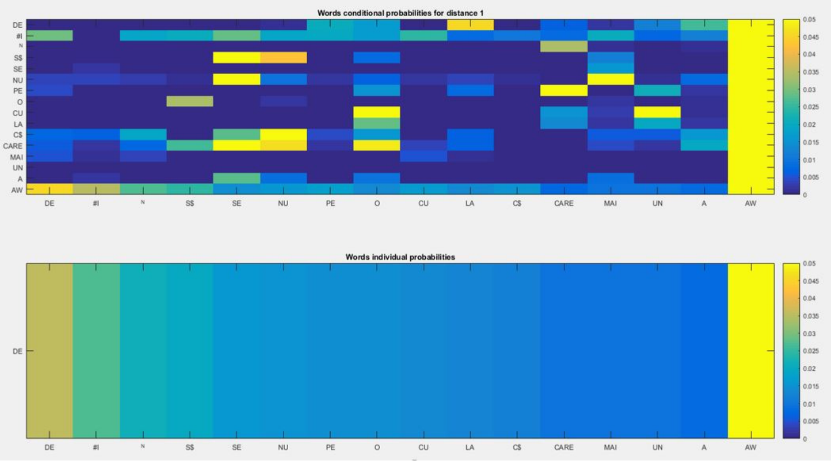


Fig. 2.12 Conditional probabilities for words $P(Word_2|Word_1)$ for $d = 1$ (X-axis = $Word_1$; Y-axis = $Word_2$)

A highly intense yellow column can be easily observed on the right side of each figure (see Figure 2.14): this corresponds to the Artificial Word (AW) that encompasses words covering approximately 76

2.2.2 Conclusions

We have managed to demonstrate that approximately 100 words are sufficient to guarantee statistical independence for both a corpus with an alphabet of 32 symbols and a corpus with an alphabet of 47 symbols.

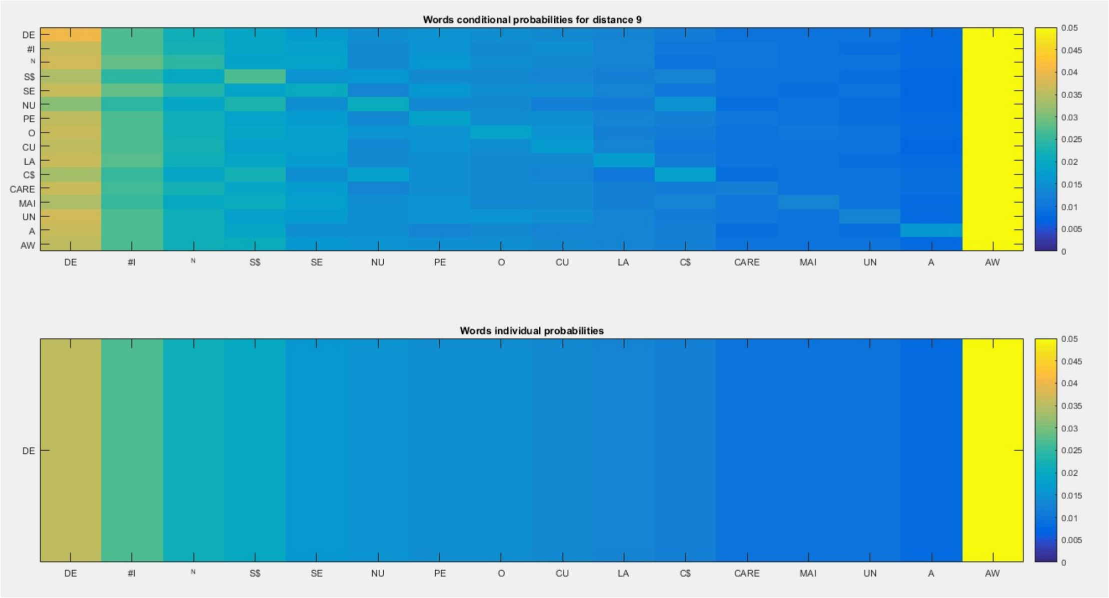


Fig. 2.13 Conditional probabilities for words $P(Word_2|Word_1)$ for $d = 9$ (X-axis = $Word_1$; Y-axis = $Word_2$)

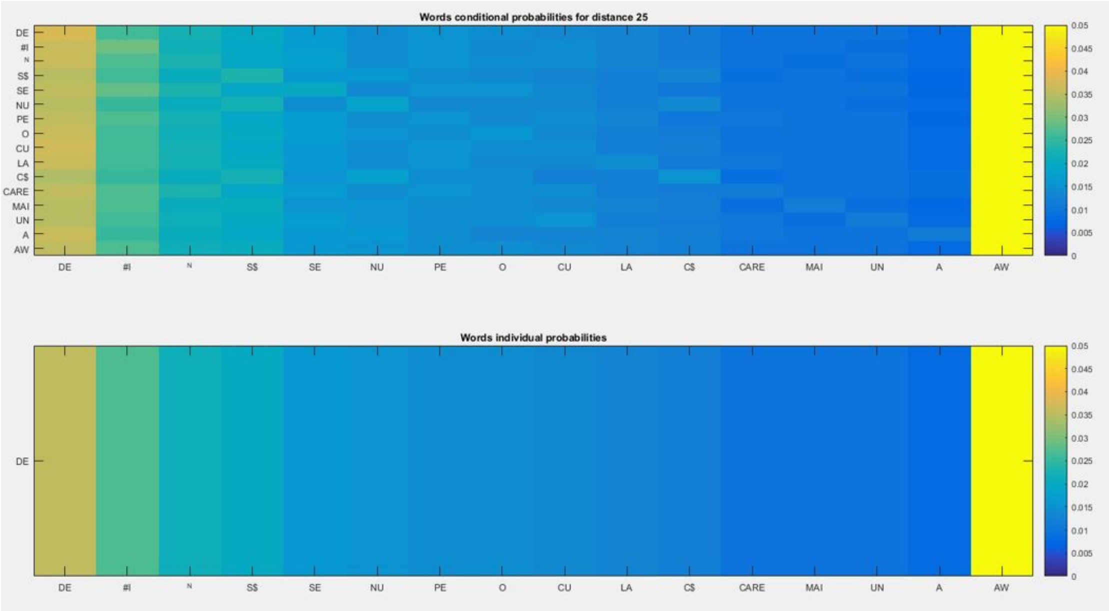


Fig. 2.14 Conditional probabilities for words $P(Word_2|Word_1)$ for $d = 25$ (X-axis = $Word_1$; Y-axis = $Word_2$)

The following sections of Chapter 2 build upon and extend the results from Sections 2.1 and 2.2, so we will only present the conclusions, with the details included extensively in the doctoral thesis.

2.3 Statistical Independence and Type II Statistical Error Probability (β)

Section 2.3 revisits the concept of statistical independence for written Romanian language. Here, the focus is on the Type II statistical error probability, which plays an important role in the discussion about independence, and the previous results of the authors' team [19, 20] are correlated with the results presented in [11, 12] from this perspective.

For the analysis involving words and their probabilities, the corpus used is not extensive enough to consider all distinct words in the statistical independence investigation. Certain aspects, such as the corpus length (as in this work, over 36 million characters long) and considering the Type II statistical error probability in evaluating word probabilities, lead to setting a minimum probability value for the words involved in the probability analysis to obtain statistically significant results.

Independence investigations based on the Chi-square test in contingency tables have found these limits, expressed by reducing the size of the associated contingency table and introducing Artificial Words obtained by combining several less frequent distinct words in the corpus (AWs created to meet the β requirements). Furthermore, multiple scenarios for creating AWs were analyzed. In all four scenarios, the independence results obtained were similar (1 AW, 3 equally probable AWs, 10 equally probable AWs, 10 AWs with different probabilities). The minimum independence distance is not dependent on the way AWs are created (an expected result based on the ergodicity of natural language). The only requirement to be met is the minimum AW probability value, closely related to the accepted level of the Type II statistical error probability.

2.4 Analysis Regarding Successive Two-Word Groups at the Beginning and End of Sentences/Phrases in a Literary Corpus of Written Romanian with Orthography and Punctuation

The overall corpus used for this study consists of 6,377,720 words. The corpus comprises 49 books (Romanian novels, as well as translations) written by 9 authors and was previously constructed by the research team. The study also includes a comparative analysis with a subcorpus composed of the works of 3 authors, namely #CHIRIŢĂ, #HERBERT, and #DUMAS (a total of 20 books out of the 49 in the overall corpus).

In this analysis, four punctuation marks have been considered as sentence/phrase endings or beginnings:

1. Period

2. Question mark
3. Exclamation mark
4. Ellipsis

The analysis was conducted in several stages:

1. Extraction of word digrams at the end of sentences/phrases regardless of the preceding punctuation mark. It is worth noting that this can only be done for sentences/phrases with a length of at least 2 words. These starting digrams were counted and sorted, and the paper presents tabulated results for the most significant digrams at the end of sentences/phrases. In the overall corpus, there are over 522,000 sentences with a length equal to or greater than 2 words and a total of 289,894 distinct digrams at the end of sentences/phrases.

2. The analysis from step 1 was repeated separately for each of the 4 punctuation marks (period/question mark/exclamation mark/ellipsis).

3. Previous work by the authors presented several results regarding the distribution of words in the corpus into 3 different regions along the Zipf's Law graph, based on the number of occurrences for each distinct word. Region 1 corresponds to words with more than 200 occurrences. These are the first 2762 words, covering approximately 72% of the corpus. We are interested in how many of the digrams at the end of sentences include words from Zipf's Region 1. The results are as follows: over 19% of the digrams at the end of sentences contain both words from Zipf's Region 1, approximately 64.4% of the digrams include a single word from Region 1, and 16.5% of the digrams at the end of sentences do not contain any word from Region 1. Comparing these results with previous results by the research team (for digrams at the beginning of sentences/phrases), it can be observed that the results for digrams at the end of sentences are different from those for digrams at the beginning of sentences, with the latter containing both words from Zipf's Region 1 to a greater extent.

4. Additionally, an analysis was performed on the connection between the words in the digrams at the end of sentences/phrases and the set of 578 common word digrams found in all 49 books in the overall corpus. Approximately 1% of the common word digrams (presented in identical form in all books) are also digrams at the end of sentences/phrases.

5. Steps 1 and 2 were repeated at the subcorpus level, focusing on the books of the 3 authors. The analysis on this subcorpus highlighted a series of relatively stable quantitative values (with minor variations) in the comparison made.

In summary, the overall results show that novel findings have been obtained regarding successive two-word groups at the end of sentences/phrases in a literary corpus of written Romanian with orthography and punctuation. The study once again emphasizes the impact of spelling and punctuation on the language model and the importance of end words and digrams at the end of sentences/phrases in the statistical description of the language.

2.5 Representative Literary Corpus

We presented a simple method to collect data from the NLCO corpus and create a matrix structure that was further analyzed from a rank-frequency perspective. An important aspect we considered when collecting data from the corpus is the minimum independence distance, which was estimated in section 2.2 to be around 100 words for written Romanian [12].

Once the data was extracted from NLCO, and the three sets of interest were selected, we observed similarities in terms of rank and relative frequency. The most distinct words in the three analyzed sets are highly correlated with the most frequent unique words in the NLCO corpus. This is the first evidence supporting the hypothesis that the analyzed sets are samples from a representative corpus of written Romanian.

Furthermore, 99% of the distinct words common to the three datasets (columns 1, 150, and 300) belong to Zipf's Area 1. Additionally, 75% of the distinct words common to the three datasets are also part of the common word set for author subcorpora in NLCO.

The new results support the idea of representativeness of the corpus constructed by concatenating multiple author subcorpora, bringing the benefits of concatenation (greater length) to natural language processing researchers.

Chapter 3

Psychology and Information Theory

3.1 Introduction

The topic addressed in this chapter is building on previous results from a complex experimental study on the impact of punctuation and orthography symbols on the model of written Romanian language when viewed as a word chain [15, 21–26]. The results from this section have been published and presented at Titu Maiorescu University, Bucharest [27, 28].

Before proceeding with the statistical analysis of the corpus, it is important to introduce the concept of informational entropy, which plays a fundamental role in the following paragraphs. In information theory, Shannon entropy or informational entropy measures the uncertainty associated with an event or a random variable. For a binary source X, for example, a coin with probabilities of showing heads as p and tails as 1-p, the entropy of the source is given by (3.1):

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

In the general case, when the information source X has n symbols with probabilities of the symbols being p_i, the entropy can be written as in (3.2):

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

In this research, we investigated the particular case of word digrams in the Romanian language and the prediction of the second word in a digram given/knowing the first word:

1. A word digram is any group of two consecutive words.

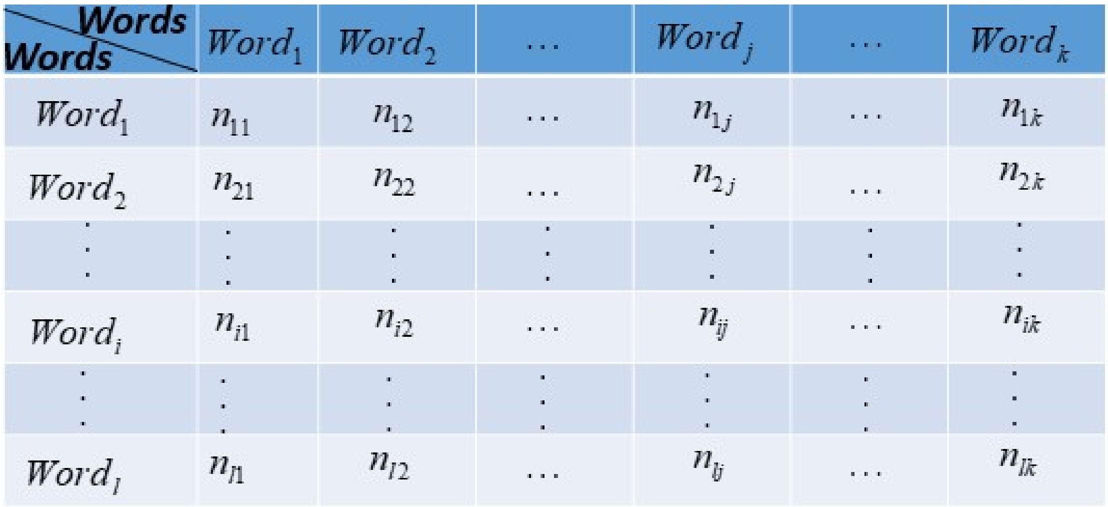
2. Based on all the digrams in the analyzed corpus, conditional probabilities of words can be estimated based on the words that precede them:

    - Step 1 – compute entries in the contingency table from Figure 3.1 (already introduced in Chapter 2).

- Step 2 – use the information from the contingency table to calculate the conditional probabilities as in (3.3):

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

  - n_{ij} is the number of occurrences corresponding to the digram (i,j), and n_i is the number of occurrences corresponding to word i.
  - It is important to note that I will continue to use the expression "Word i," which signifies the word with rank i in the descending order of appearance frequencies in the corpus.



Words Words	$Word_1$	$Word_2$	$\cdots$	$Word_j$	$\cdots$	$Word_k$
$Word_1$	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}
$Word_2$	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Word_i$	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Word_l$	n_{l1}	n_{l2}	$\cdots$	n_{lj}	$\cdots$	n_{lk}

Fig. 3.1 Contingency table for word digrams

Once the word digrams and estimated conditional probabilities are inventoried, it is time to transition to the main interest in this study: entropy. It is self-evident that each word can theoretically be followed by any word in the language. In practice, not all words make sense to follow any word before them, but the relationship (3.4) remains valid:

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

where l represents the total number of distinct words in the corpus.

We can associate the appearance of each word with a similar experiment of rolling a dice, where each face of the dice represents one of the possible words that can follow Word i. In other words, for each Word i, we have a mini-source of information (given by the written Romanian language), with possible choices or events represented by the possible Words j that can follow a fixed Word i. Equation (3.4) is the well-known condition for the noise matrix of an information channel - the sum over rows is 1 (rows correspond to the input of the information channel - Words i, and columns correspond to the output - Words j).

For each of these sources of information (word i), a conditional entropy (3.5) can be calculated:

$$H + i = H(j|i) = -\sum_{j=1}^{l} P_{(}j|i) \cdot \log_2(P(j|i))[bits/word] \qquad (3.5)$$

Informational entropy encapsulates important information about the surprise an observer has when a certain face of the dice appears. In the case of this study, surprise is translated into the wonder that a particular word (j) follows another word (i).

Figure 3.2 graphically presents the values of conditional entropy for the first 500 most frequent distinct words in written Romanian. A slight decrease in entropy values can be observed as words become less frequent in the language. This can be explained because the most frequent words, the first ones, can be followed by many other words with nearly equal probabilities, whereas less frequent words exhibit certain preferences and are more frequently followed by specific words, resulting in lower entropy.

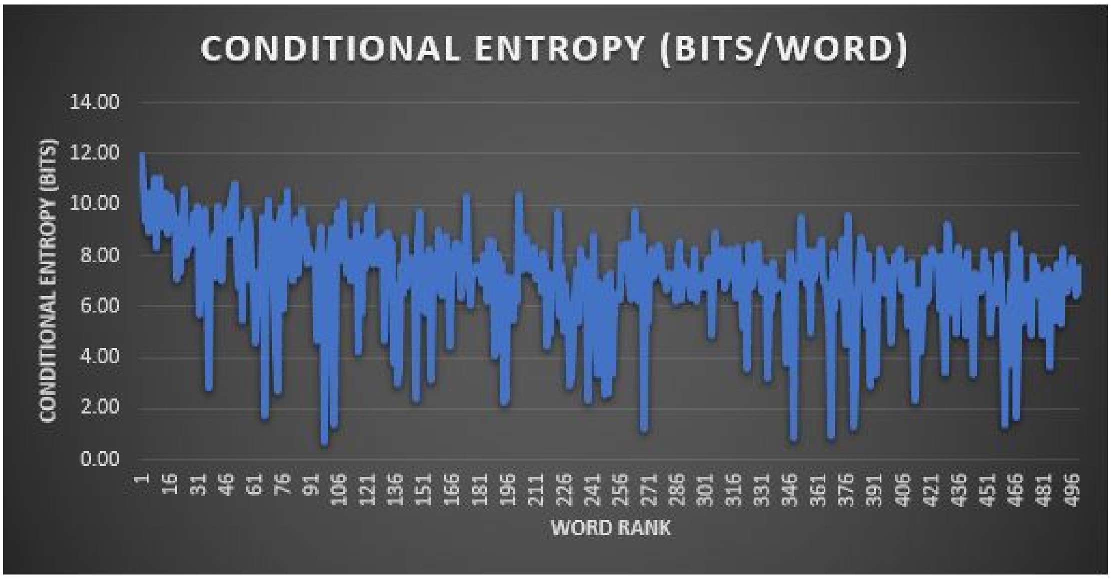


Fig. 3.2 Conditional entropy for the first 500 most frequent distinct words in the corpus

We started with the following hypotheses:

- Hypothesis 1: The success rate in "guessing" the next word after a word of interest is inversely proportional/negatively correlated with the entropy value associated with the fixed word.

- Hypothesis 2: The responses obtained in the questionnaire are strongly correlated with the results obtained based on the literary corpus.

The first approach is based on the statistical analysis of a literary corpus in the Romanian language obtained by concatenating 49 books (fiction). Based on this corpus, we extracted descriptive statistics such as word frequency or the frequency of word groups most commonly used in the Romanian language. At the same time, conditional probabilities of Romanian words can be easily calculated and estimated based on these probabilities, conditioned entropies can be associated with each word (the group of interest was represented by the first 500 most frequently used words) - (3.5).

Secondly, based on the analysis performed on the aforementioned literary corpus, we selected 30 representative words in terms of information and conditional entropy: 10 words with low entropy, 10 words with medium entropy, and 10 words associated with high entropy values (the hypothesis on which we started, based on the observations of C.E. Shannon [29], is that entropy is in an inverse relationship with the amount of surprise associated with a particular word and the following word). These 30 words selected from the corpus were used to create a questionnaire with questions like: "What do you think is the most frequently used word in the Romanian language after word ...?".

Furthermore, correlations were made between the variables of interest: the success rate of participants in the questionnaire (compared to the correct version based on statistical results from the corpus) and the conditional entropy value for each word included in the questionnaire, or the success rate of the following words obtained based on the information available from the literary corpus. Secondary variables, such as age, gender, or the level of education of the study participants, were also used to draw conclusions and conjectures on the subject of interest.

The questionnaire was created using the online platform Google Forms and consisted of 30 questions of the form: "What do you believe is the most commonly used word in the Romanian language after the word CEEA?". Respondents had to choose from 6 possible and meaningful options, all consisting of the most frequently used words after the word of interest (CEEA, in the example above). Filling out the questionnaire was anonymous, with participants providing only general information such as age, gender (M/F), and education level. A total of 101 individuals responded to the questionnaire. The distribution of respondents based on the collected information is presented below.

Figure 3.3 presents the distribution of study participants based on their age. It can be observed that over half of the participants are under the age of 35, with the rest being relatively evenly distributed and covering the range of 35 to 58 years.

In terms of education level, the majority of respondents have completed undergraduate studies (over 90

Regarding the distribution of respondents by gender, the results are as follows:

- 74 female
- 27 male

The questionnaire was sent and received responses from study participants during the period of April 12-15, 2020.

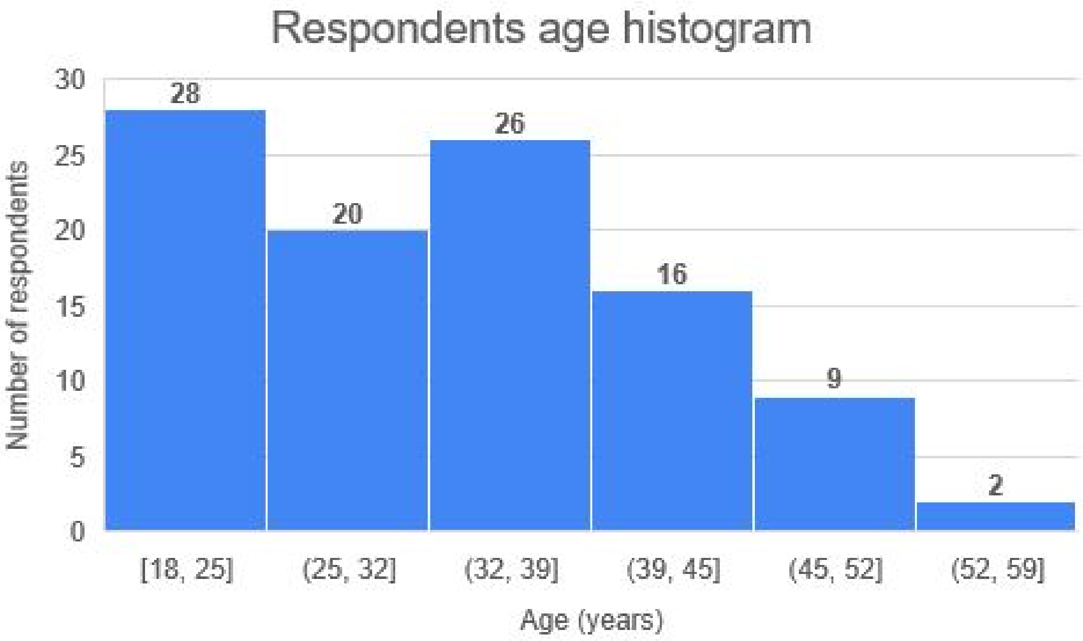


Fig. 3.3 Age Distribution of Respondents

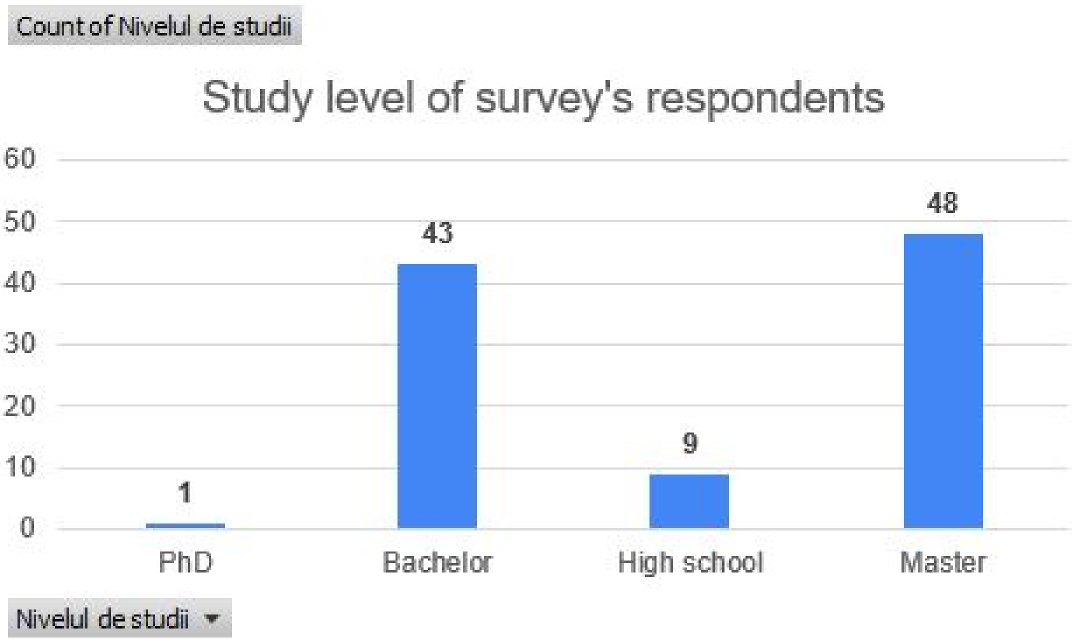


Fig. 3.4 Education Level Distribution of Respondents

3.1.1 Quantitative and Qualitative Interpretation of Data

To easily calculate the desired correlations, responses to the questionnaire were summarized as follows:

1. For each question, we calculated how many individuals chose each of the possible answer options.

2. Based on the number of preferences for each option out of the 6, we calculated the success rate for each option by comparing it to the correct reference results obtained from the statistical analysis of the available literary corpus.
3. Thus, for each question, we obtained a vector of 6 percentage values for each answer option, with the percentage for the most frequent word of interest in that question being the most important.
4. This experimental success rate, based on the responses of study participants, was correlated/compared with the theoretical success rate obtained from the statistical analysis of the literary corpus.
5. It is expected that the success rate for each question depends on the conditional entropy value associated with that word (Hypothesis 1) and is strongly correlated with the success rate resulting from the statistical analysis of the literary corpus (Hypothesis 2).

To better understand the above steps, I will now present a numerical example:

1. For the question: "What do you think is the most frequently used word in the Romanian language after the word CEEA?", the questionnaire results are as shown in Figure 3.5:

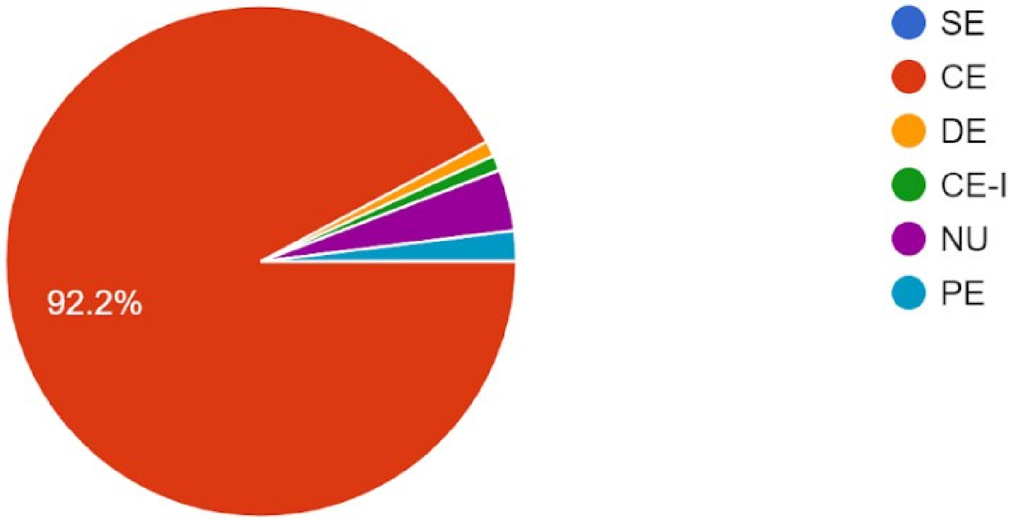


Fig. 3.5 Questionnaire Results for Question 1 (Low Entropy)

2. On the other hand, from the statistical analysis of the corpus (which was the starting point for developing the questionnaire), the results are very similar to those in Figure 3.5 - see Table 3.1.

Table 3.1 Results of Corpus Analysis for the Word CEEA and Its 6 Followers (Question 1)

Word Rank	**Conditional Entropy (bits/word)**	**Word 1**	**Follower Option 1**	**Follower Option 2**	**Follower Option 3**	**Follower Option 4**	**Follower Option 5**	**Follower Option 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. Similar results are obtained for the other nine questions that address words with low entropy (one of the conditional probabilities is very high).

The results presented in the previous sections and detailed in the doctoral thesis allow the following observations:

1. Hypothesis 1 has been validated: The lower the conditional entropy of a word, the easier/probable it is for a native speaker to know what word comes after the word of interest (correlation coefficient = -0.76).

2. Hypothesis 2 has been validated: Participants in the questionnaire selected with a high probability the same following words after the 30 words of interest, as expected from the statistical analysis of the available literary corpus (correlation coefficient = 0.84).

3. Additionally, age and education level do not play a significantly important role in the total scores obtained by the participants in the questionnaire. However, what does make a difference between responses is gender, with female participants performing better on average with one question compared to male participants.

4. A generally applicable observation, also mentioned in the introduction of the work, is that these native, unconscious, and collective psycholinguistic knowledge is well-preserved for each of us and, over time, after years and years of practice, we choose the following words and form sentences in a very interesting dual manner: with meaning and at the same time following a very well-defined statistical pattern.

3.2 Future development prospects

Possible directions for future research could include clarifying the initial results obtained in this study, namely, the score obtained by female individuals is higher than that obtained by male respondents. Is this a coincidence, due to the size of the selected sample, or is it a generally valid conclusion? And more importantly, if the answer to the previous question is positive, what is the underlying reason?

Additionally, another very interesting area could be the generation of random text based on the probability structure and conditional entropy resulting from the analysis of the available literary corpus, and the validation/refinement of the level of meaning obtained through this process with the help of future study participants/respondents.

Moreover, the approach presented in this chapter could be used to create a test that could be used for early detection or monitoring of individuals suffering from neurodegenerative diseases. There are currently several internationally recognized tests that achieve this goal (MMSE - Mini Mental State Examination, clock test, etc.), but a linguistic test based on the specifics of the Romanian language could provide better results for the population in Romania.

Chapter 4

Cryptography and Chaos Theory

This final chapter of the doctoral thesis proposes a paradigm shift from the previous sections. The focus will be on the field of cryptography with chaos and is based on several works published by the author since undergraduate studies [10] or doctoral research [30–32].

4.1 The compound tent map and the connection between Gray Codes and initial condition recovery

The tent map is a one-dimensional discrete chaotic system defined by the following equation:

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

where $p \in (0,1)$, and p is the parameter of the tent map. Due to ergodicity and sensitivity to the initial condition x_0 and the control parameter p (which must be different from 0.5, otherwise, applying formula (4.1) with $p = 0.5$ yields a non-chaotic sequence), combined with the uniform probability distribution of values x_k, the tent map can be successfully used in chaos-based cryptographic applications.

In most studies involving the tent map, the chaotic signal is used as a generator for encryption sequences [33–35]. The encryption sequence ("the key") is obtained by binarizing a trajectory (successive iterations of the tent map) defined by formula (4.1), using a certain threshold c:

- If $x_k \leq c$, assign the binary value $b_k = 0$.
- If $x_k > c$, assign the binary value $b_k = 1$.

One of the ways to create the ciphertext is by modulo-2 addition (symbol by symbol) of the plaintext and the "key." Therefore, the question that arises is: "Can someone, having a part (a bit sequence) of the binary sequence corresponding to the key, recover the initial condition x_0 that generated the respective trajectory of the tent map?" If this is possible, the entire "key," no matter how long it is, can be reconstructed, and in this case, the initial condition cannot be used as an element in the secret key.

We have shown that in certain cases, the initial condition should not be included in the cryptographic key since it can be easily found. One solution to make it harder to find is discretization/binarization with $c = 0.5$ and sampling data i.i.d. by sampling the used chaotic map. Additional security in this regard could be obtained by applying a "running-key" approach for the tent map, as proposed for the logistic map in [36]. This approach was not used in this investigation.

We found a relatively simple general formula to describe $f_n(x)$. Furthermore, we demonstrated why x_0 can be found from only a small number of 16 trials, a result stated in the literature but not convincingly proven. This result holds for $c = p$, when considering successive iterations without sampling the tent map. Moreover, in this case, we noticed a subtle connection with Gray codes. Describing the compound tent map provides additional support for the idea that sampling the tent map could be a significant obstacle in recovering the initial condition.

Some difficulties may arise when someone wants to use our formula for the compound tent map. These are mainly related to the order in which the operations defining the compound function are performed. However, this new signal - the compound tent map - has the same properties as (4.1) (the same uniform probability law; the minimum sampling distance for statistical independence decreases with the order m of the composed map, for example, for $p = 0.4$ and $f_{15}(x)$, successive values are practically i.i.d.), and it can be a new research source in this field.

4.2 The Lorenz chaotic system, statistical independence and sampling frequency

The purpose of this study [30] was to investigate the relationship between the Lorenz chaotic system, statistical independence, and the sampling frequency/time step used to solve the Lorenz equations. Our hypothesis was that there should be no dependence on the sampling frequency, and statistical independence should be achieved similarly, regardless of how finely or coarsely the Lorenz equations are solved.

Through the spatial Chi-squared test, we were able to demonstrate that statistically independent data can be extracted from the solution space of the Lorenz equations. Furthermore, we showed that the independence time of about $30s$ is independent of the time step/sampling frequency used in solving the system of differential equations. This is a very important result because, in practice, it might be easier, more efficient, and less computationally expensive to obtain solutions of the Lorenz system with a larger time step/smaller sampling frequency. Independent data required for various practical applications can then be selected to meet the independence time requirements.

4.3 Singularity, observability and statistical independence in the context of chaotic systems

The detailed results in the doctoral work show a correlation between the observability coefficient and the overlap between the attractor and the singularity manifold. It can be easily observed that the higher the observability coefficient is for a specific state variable, the greater the overlap between the singularity region and the attractor of that chaotic system. In practice, systems are chosen so that the overlap is as small as possible, as this allows more flexibility when applying data in various cryptographic applications. Overall, in terms of observability-singularity, out of the three analyzed systems, Ikeda seems the most promising, followed by Tinkerbell and Clifford. However, singularity and observability are not the only concepts that matter. In terms of statistical independence, the Clifford map is the only one capable of being used as a pseudo-random number generator (PRNG).

In conclusion, this research effort presents a new analysis procedure for dynamic maps, considering essential concepts such as statistical independence, singularity, and observability. The significant advantage of the proposed thought process in this article is that it does not rely exclusively on a single notion, no matter how powerful it is. Singularity, observability, and statistical independence have been addressed separately in the literature, and it is a known fact that they can provide important insights into specific problems of interest in cryptography. However, using the results from all three different perspectives considered together leads to a convergent and stronger conclusion that can be used and further investigated by researchers in the field of dynamic systems with cryptographic applications.

Chapter 5

Conclusions and list of original publications

5.1 Conclusions

Linguistics, psychology, and engineering may seem like unrelated subjects, but we have demonstrated how concepts from information theory and statistics can provide interesting answers to humanistic questions.

Chapter 2 focuses on the study of written Romanian language and analyzes from multiple perspectives the literary corpus available in the research collective [19].

The study presented in Chapter 3 aimed to investigate and explain in a simple and visual manner the connection between concepts that each of us uses every moment (words and meaningful phrases used for mutual understanding), even without always realizing it, and very complex mathematical procedures and quantities at first glance (entropy, conditional probabilities, or mean values, variances, and correlations).

The interdisciplinary and ubiquitous analysis continues in Chapter 4, when concepts of cryptography are statistically approached and even connected to fundamental concerns in physics or transdisciplinarity. We showed that in certain cases, the initial condition should not be included in the secret key used in cryptography because it can be easily found. A solution to make it harder to find is discretization/binarization with $c = 0.5$ and the sampling of the chaotic map used. The aim of the study in Section 4.2 was to investigate the relationship between the chaotic Lorenz system, statistical independence, and the sampling frequency/time step used to solve the Lorenz equations [30]. Given the results presented in Section 4.3, several important conclusions can be drawn regarding the proposed procedure for the unified testing of different chaotic systems and the selection of pseudo-random number generators (PRNGs). In general, from the perspective of observability-singularity, out of the three analyzed systems, Ikeda seems the most promising, followed by Tinkerbell and Clifford. However, singularity and observability are not the only concepts that matter. From the perspective of statistical independence, the Clifford map is the only one capable of being used as a pseudo-random generator (PRNG). This research endeavor presents a new procedure for the analysis of dynamic maps, taking into account essential concepts such as statistical independence, singularity, and observability. The significant advantage of the proposed line of thinking in this article is that it is not based exclusively on a single notion, however strong it may be. Singularity, observability, and statistical independence have been separately addressed in the literature, and it is a known fact that they can separately provide important perspectives on specific issues of interest in cryptography. However, using the results from all three different considered perspectives together leads to a convergent and stronger conclusion that can be used and further investigated by researchers in the field of dynamic systems with cryptographic applications. The proposed approach can be used for any system, regardless of the number of state variables. Experimental results show that there is a fragile balance between the concepts that can be used to select a system for use in cryptography, and there is no concept of "one size fits all," but rather a compromise depending on the application of interest.

5.2 List of Original Publications

The results included in the doctoral thesis have been published in the journals and conferences listed below (1 Q1 and 1 Q3 articles included in the total of 5 WOS-indexed publications and 6 BDI).

5.2.1 Journal Articles

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 Conferences

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 Research Reports and Other Publications

1. 4 scientific research reports within SD-ETTI:

   - Scientific Report No. 1/2019, "The Chi-square test in contingency tables and its applications for natural language processing"
   - Scientific Report No. 2/2019, "Revisiting the statistical independence for the printed Romanian language"

- Scientific Report No. 3/2020, "The statistical independence for words in the printed Romanian language"
- Scientific Report No. 4/2020, "The Lorenz chaotic system, statistical independence, and sampling frequency"

2. Two research reports within the "Resources and Technologies for the Romanian Language in a Standardized Multilingua. Context" research program at the Institute of Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy, in which I was involved with the research team coordinated by Professor Adriana Vlad.

3. Dinu, A. (2020a). Psycholinguistics, statistics, and the unconscious mind. Bachelor’s Thesis in Psychology presented at Titu Maiorescu University, Bucharest.

References

[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.