Arabic Language N-Gram Frequencies

1st Anthony El Chemaly
Department of Computer Engineering
Holy Spirit University of Kaslik
Jounieh, Lebanon
anthony.c.elchemaly@net.usek.edu.lb

2nd Catherina El Khoury
Department of Computer Engineering
Holy Spirit University of Kaslik
Jounieh, Lebanon
catherina.n.elkhoury@net.usek.edu.lb

***Abstract*—Arabic is one of the most prevalent languages spoken globally, with more than 400 million people residing in the Middle East and North Africa claiming it as their first language. This language holds immense significance in fields such as literature, religion, and science. Its complexity, with a rich system of root-based word construction and a unique script, has long fascinated linguists and researchers.**

Understanding the frequency of letters, words, and combinations within any language is crucial for numerous analytical purposes. In the context of Arabic, studying letter frequencies is essential for linguistic analysis, as it reveals patterns in the language's structure that may otherwise remain hidden.

From a cryptographic perspective, letter frequency analysis plays a vital role. In classical cryptography, particularly in substitution ciphers, letter frequency distributions were used to crack encoded messages by comparing the frequency of letters in the ciphertext with typical frequency patterns of the language.

This paper presents a study focused on extracting the Arabic language n-gram letter frequencies to understand their patterns. We will work with a large Arabic Corpus applying preprocessing steps to ensure data consistency. While the primary focus of this study is on n-grams, we specifically analyzed sequences ranging from n = 1 to n = 7. However, the approach applied allows others using the same methodology to easily extend the analysis to additional n-gram levels.

The results provide a clear distribution of the most common letter combinations in Arabic, offering insights that can support research in areas such as security applications. The datasets, procedures, and outcomes are systematically documented for further study.

***Index Terms*—Arabic language, Language analysis, Letter frequency, N-Grams, Arabic word dataset**

I. Introduction

Arabic letter frequency analysis is an important area of study, especially in the cryptography field [1]. Understanding how often certain letters and combinations of letters (n-grams) appear in Arabic texts can provide valuable insights for developing better security systems.

In this project, we focus initially on studying Arabic letter n-grams (sequences of n letters). We started by picking a dataset of Arabic words and applied a systematic methodology to analyze the frequency of different letter combinations. The logic used is based on counting the appearances of each n-gram across the dataset and calculating their overall frequency relative to the total number of letter groups found.

While similar studies have been conducted for other languages [19], especially English, Arabic presents unique challenges. The Arabic language has a different script, a larger set of letter forms, and specific structural features such as the connection of letters within words. These characteristics mean that the frequency distribution in Arabic cannot be assumed to behave the same way as in English or other Latin-based languages, making this study particularly important for applications that involve Arabic texts [2].

The main objectives of this project are to document the datasets used, describe the procedures followed, present the frequency results, and discuss how the findings can be used in security contexts [3] .

This document outlines the steps taken throughout the project, explains the methodology in detail, and highlights the key results obtained from the analysis.

II. Literature Review

The analysis of letter frequencies has a long-standing history in linguistic studies and cryptographic research. In the English language, foundational works such as Mayzner and Tresselt's (1965) study [2] on letter and bigram frequency counts provided critical insights that have been extensively used [14] [15] [16]. Inspired by such studies, Peter Norvig revisited and extended these analyses, demonstrating how n-gram frequency extraction could further improve natural language processing models and predictive text applications.

On December 17th, 2012, Peter Norvig, Director of Research at Google, received a letter from Mark Mayzner, a retired 85-year-old researcher, whose 1965 publication had been cited in hundreds of scholarly articles. Mayzner inquired whether Norvig's team at Google would be interested in utilizing the massive computing power now available to significantly expand and produce new frequency tables similar to those Mayzner constructed over 50 years ago, but this time, using the Google Corpus Data. This moment marked a pivotal evolution in the study of letter frequencies, with Norvig revisiting Mayzner's work to produce more accurate, expansive letter frequency data for English. This revision [3] not only enhanced the understanding of English letter patterns but also contributed to advancements in text analysis and cryptography, influencing modern applications in natural language processing (NLP).

In the context of Arabic, research on letter frequency remains relatively underdeveloped, despite a few notable contributions [4] [9] [18]. One such effort is by Mustafa and Bouzoubaa (2012) [4], who introduced a bigram-based

method for constructing an exhaustive lexicon of Arabic triliteral roots, which required extracting bigrams from Arabic text. Additionally, Boudelaa, Perea, and Carreiras (2020) [9] provided comprehensive matrices detailing the frequency and visual similarity of Arabic letters and their allographs, offering valuable insights into the structural characteristics of Arabic script.

They emphasized the distinct features of Arabic, including its root-based morphology and the complexities introduced by diacritical marks and the varying shapes of letters based on their positions within words.

In fact, Arabic natural language processing has consistently faced unique challenges due to the language's complex morphology and orthographic diversity. As Habash (2010) points out, effective Arabic NLP systems must account for both the rich derivational structures and the impact of optional diacritics, which influence word meaning and pronunciation. These features introduce sparsity issues when modeling n-grams, particularly for larger values of n. Moreover, the shape-shifting nature of Arabic letters depending on their word position adds a layer of visual variability that further complicates tokenization and frequency analysis. Thus, any meaningful extraction of Arabic n-gram frequencies demands meticulous preprocessing, including normalization, diacritic removal, and root-based grouping of semantically related tokens [12].

Similarly, Nafea et al. [5] and other studies [10] highlight how preprocessing — such as normalization, diacritic removal, tokenization, and handling variations in letter forms — is essential before performing any meaningful linguistic analysis on Arabic datasets. However, the modeling of n-grams in Arabic remains challenging due to its morphological richness and orthographic variability.

This points to the importance of preprocessing steps, such as normalization and diacritic removal, in achieving consistent frequency counts for Arabic text [8].

While these efforts have contributed valuable insights into Arabic computational linguistics, there is still a significant gap in comprehensive studies that document Arabic n-gram frequencies across a wide range of n-values (e.g., up to 7-grams).

Despite these advancements, comprehensive n-gram studies for Arabic, especially those using large-scale corpora and systematic preprocessing, remain limited. This study aims to address this gap by providing a detailed analysis of Arabic letter n-gram frequencies, from n = 1 to n = 7, using a large, curated corpus.

From a cryptographic perspective, letter frequency analysis has long served as a cornerstone of classical cryptanalysis. In monoalphabetic substitution ciphers, each plaintext letter is consistently replaced with another letter, making the statistical properties of the language — particularly letter frequencies — a powerful tool for decryption. Cryptanalysts historically leveraged the fact that certain letters occur more frequently than others to infer potential substitutions in ciphertext. This technique, known as frequency analysis, was famously used by Arab polymath Al-Kindi in the 9th century, who is credited with developing the earliest known method of cryptanalysis based on frequency statistics [13]. His foundational work demonstrated that by analyzing the occurrence of characters in encrypted texts, one could make educated guesses about the original plaintext — a principle that continues to underpin various statistical and computational approaches in modern-day cryptography. His foundational work demonstrated that by analyzing the occurrence of characters in encrypted texts, one could make educated guesses about the original plaintext — a principle that continues to underpin various statistical and computational approaches in modern-day cryptography [20], [21].

Additionally, in the 16th century, the Italian cryptanalyst Giovanni Soro expanded on Al-Kindi's methods by formalizing more systematic approaches to breaking substitution ciphers, which influenced later works in cryptography [22]. The effectiveness of frequency analysis was further emphasized in the 20th century during the deciphering of German Enigma machine codes, as statistical methods such as letter and digraph frequency analysis played a crucial role in breaking the cipher [23], [24]. Today, frequency analysis continues to inform techniques used in breaking simple encryption methods and serves as a foundation for cryptanalysis in the context of modern cipher attacks.

By documenting the methodology and results, this paper seeks to provide a valuable resource for future linguistic, cryptographic, and computational research in Arabic.

III. Methodology

In this project, we focus on efficiently extracting n-grams from the 101 Billion Arabic Words Dataset [6] and analyzing the stability of these n-grams over multiple processing checkpoints. The goal is to identify the most frequent n-grams (up to 7-grams) in Arabic and determine when the frequency distribution of these n-grams stabilizes. Stability is measured using a log-scale graph and mean absolute difference to track the changes in n-gram frequencies over time.

A. Dataset Collection:

The 101 Billion Arabic Words Dataset [6] is a large-scale, publicly available corpus curated by the Clusterlab team. This dataset consists of a massive 101 billion words extracted from web content, representing a mixture of Modern Standard Arabic (MSA) and various Arabic dialects. The diversity in the dataset is crucial for developing robust language models that are capable of understanding both formal and colloquial forms of Arabic.

Moreover, this dataset is licensed under the Apache 2.0 license [17] and is freely accessible for research. It has been widely used for training and fine-tuning Large Language Models (LLMs), particularly those focused on Arabic natural language processing (NLP).

The dataset includes text from various domains such as news, blogs, forums, and social media, providing a comprehensive representation of Arabic as used in modern contexts. This diversity allows for the extraction of rich and varied linguistic

patterns, which are essential when studying the structure and evolution of the Arabic language.

To ensure the ability to monitor the stabilization of n-gram frequencies over time, this extensive dataset was selected. This extensive corpus allowed for continued counting until statistical stability was achieved, thereby ensuring that the analysis captured a representative and diverse distribution of letter combinations within the Arabic language.

B. Preprocessing:

Preprocessing was essential to ensure the consistency and cleanliness of the data before n-gram extraction [7].

For Arabic corpora, preprocessing involves several stages to handle the complexities of Arabic script and language structure. The steps included:

1) Removing Diacritics: Arabic text often includes diacritics (such as fatha, damma, kasra) that provide vowel sounds but are typically omitted in modern Arabic writing, especially in informal contexts. Diacritics are removed to standardize the text and focus on the root structure of words.

2) Removing Punctuation, Special Characters, Foreign Language Characters, and Numbers: All punctuation marks, special characters (such as parentheses, quotation marks, etc.), foreign language characters, and numbers are removed from the text. This ensures that the focus remains purely on the linguistic content, avoiding formatting, syntactic, or non-relevant features that could interfere with meaningful n-gram extraction.

3) Normalization: Arabic characters such as alif, ta marbuta ة, and others have different forms. For example, alif can appear in multiple shapes: ا or أ. These forms are normalized to a single representation. This normalization is important for improving consistency and reducing the variety of word forms that can occur due to orthographic differences.

After preprocessing, the dataset was a clean list of Arabic words ready for n-gram extraction.

For Arabic corpora, preprocessing involves several stages to handle the complexities of Arabic script and language structure. The steps included:

C. Tokenization

Tokenization is a fundamental preprocessing step in natural language processing (NLP) that involves segmenting a text into smaller, meaningful units, typically referred to as tokens [11]. As part of tokenization:

1) Word-level Tokenization: The text is split into individual words, where each word is treated as an atomic unit. The importance of word-level tokenization is particularly evident when studying n-grams, which are contiguous sequences of n words drawn from a given text. Constructing meaningful n-grams critically depends on an accurate identification of word boundaries. Without proper word-level tokenization, the resulting n-gram models would suffer from noise and inconsistencies, leading to unreliable linguistic patterns.

D. N-Gram Extraction

Once the text is preprocessed, the next step is to extract n-grams from the data. N-grams are continuous sequences of n items (characters or words) from the text. In this study, n-grams are extracted using a sliding window approach, where a window of size n is applied to the word, and the corresponding n-gram is captured for each window.

1) Sliding Window Approach: The sliding window approach ensures that overlapping n-grams are extracted from the text.

The window moves one step at a time, ensuring that all potential n-grams are captured.

E. Frequency Calculation

After extracting all the n-grams from the dataset, their frequencies were calculated through the following steps. Each token is processed by:

- Counting letter occurrences within each word
- Generating and counting all possible n-grams of lengths 2-7 within each word

The n-grams were then sorted in descending order based on their frequency to highlight the most commonly occurring letter sequences. This approach was applied consistently across bigrams, trigrams, 4-grams, and 5-grams, 6-grams and 7-grams.

F. Incremental Processing and Convergence Analysis

This approach allows us to determine when sufficient data has been processed to achieve stable distribution estimates: The corpus is processed in batches of 1,000 examples, defining an epoch. At each epoch, a checkpoint is created with current distribution statistics.

The following steps are applied:

- The mean absolute difference between consecutive distribution checkpoints is calculated. This difference metric quantifies the stability of the distributions over time.
- The top contributors to distribution changes are identified and tracked.

The mean absolute difference is calculated as follows:

- Raw counts are converted to percentage distributions
- We compute the absolute percentage difference for each element (letter or n-gram)
- We proceed by averaging these differences across all elements

IV. Results

In this section, we present the results of the n-gram extraction and frequency analysis performed on the 101 Billion Arabic Words Dataset.

A. Distribution Convergence Analysis

A principal objective of this study was to determine the point at which statistical distributions of Arabic characters and n-grams stabilize. Distribution convergence was measured by calculating the mean absolute difference in percentage frequency between consecutive processing epochs. Each epoch is 1000 samples from the dataset. This metric quantifies how much the distribution changes as additional data is processed, effectively measuring the "learning rate" of the statistical model.

Our analysis revealed a consistent pattern of exponential decay in distribution change, indicating that each new batch of data contributed progressively less new information about the underlying frequency distributions. By epoch 599 (599,000 examples processed), the following stability metrics were observed:

- Letter distribution change: 0.00003306%
- 2-gram distribution change: 0.00000547%
- 3-gram distribution change: 0.00000071%
- 4-gram distribution change: 0.00000012%
- 5-gram distribution change: 0.00000003%
- 6-gram distribution change: 0.00000002%
- 7-gram distribution change: 0.00000002%

These extremely low percentage changes between consecutive epochs indicate that the distributions had effectively converged, with longer n-grams (5-7) reaching stability earlier than shorter sequences. This pattern suggests that after processing approximately 600,000 examples, the additional information gain from further data processing becomes negligible.

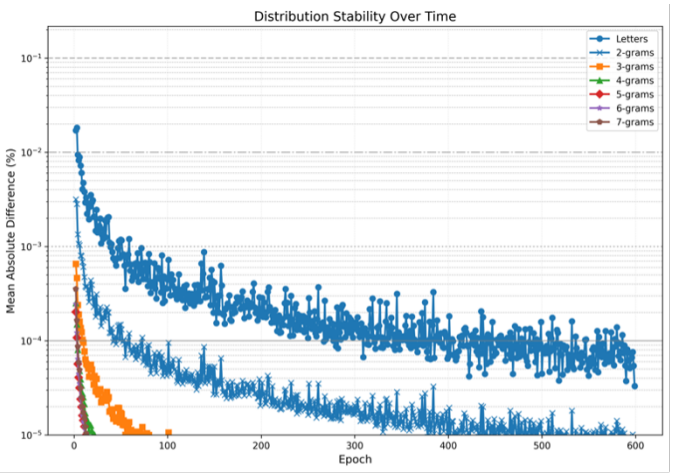

Fig. 1. Distribution Stability Over Time

As presented in Fig. 1, stability is determined by a smaller value and a flattening of the curve on the graph. When the mean absolute difference approaches very small values and the lines on the log-scale graph flatten out horizontally, it indicates the distribution has stabilized. This means that processing more data doesn't significantly change the letter or n-gram frequency distributions anymore.

B. Dataset Processing Statistics

The analysis processed a total of 600,000 examples from the 101 billion Arabic words dataset, extracting 1,799,391,295 Arabic letters after preprocessing. This substantial corpus provided a robust foundation for statistical analysis of Arabic character distributions.

C. Letter Frequency Distribution

Analysis of single letter frequencies revealed distinctive patterns characteristic of Arabic text. The complete letter frequency distribution has been documented in tabular format, showing both raw counts and relative percentages for standard Arabic letters and their variants. The most frequent letters in the corpus showed expected patterns consistent with prior linguistic understanding of Arabic.

D. Interpretation of High-Frequency Letters

The frequency of individual letters in Arabic provides valuable insight into the language's structure and the role of specific letters in word formation. The most common letters, based on their frequency, often serve crucial grammatical and morphological functions. The interpretation of the top five most frequent letters in Arabic is as follows:

1) ا (Alif): The most frequent letter in the dataset, ا, functions both as a root letter and as a grammatical marker. It appears in the definite article "ال" (al-), as part of verb forms, and as a placeholder for various long vowels.

2) ل (Lam): Often paired with ا, ل appears frequently due to its role in definite articles (e.g., البيت, "the house"). It is also a common consonant in Arabic root systems, contributing to its high standalone frequency.

3) ي (Ya): The third most common letter, ي, often functions as a suffix denoting possession (بيتي, "my"), and is frequently found in root words and verb conjugations, especially in the imperfect tense (e.g., يكتب, "he writes").

4) م (Meem): A morphologically productive letter, م is used to form nouns and participles (e.g., مدرس, "teacher," from the verb درس). Its utility in multiple grammatical forms elevates its frequency.

TABLE I
FREQUENCY OF ARABIC LETTERS

Rank	Letter	Frequency	%	Rank	Letter	Frequency	%	Rank	Letter	Frequency	%
1	ا	314,447,080	17.5072%	11	ع	59,756,349	3.3270%	21	ش	19,204,048	1.0692%
2	ل	208,235,162	11.5937%	12	د	53,618,053	2.9852%	22	خ	16,639,780	0.9264%
3	ي	148,421,728	8.2635%	13	س	49,545,896	2.7585%	23	ز	12,624,121	0.7029%
4	م	117,461,708	6.5398%	14	ف	44,573,587	2.4817%	24	ض	10,590,117	0.5896%
5	ه	97,788,303	5.4445%	15	ك	40,073,325	2.2311%	25	ث	8,831,377	0.4917%
6	ر	92,304,653	5.1392%	16	ح	36,321,961	2.0223%	26	ذ	8,088,845	0.4504%
7	و	91,897,301	5.1165%	17	ق	35,852,410	1.9961%	27	غ	7,412,574	0.4127%
8	ن	85,710,173	4.7720%	18	ج	29,068,243	1.6184%	28	ئ	7,093,993	0.3950%
9	ت	85,077,904	4.7368%	19	ص	20,896,114	1.1634%	29	ء	6,022,905	0.3353%
10	ب	63,135,041	3.5151%	20	ط	19,885,231	1.1071%	30	ظ	3,982,191	0.2217%
31	ؤ	1,544,500	0.0860%								

5) ه (Ha): Serving both as a consonant and a possessive suffix (ـه, "his"), ه is integral to personal pronoun constructions and contributes to syntactic cohesion in sentences.

E. Bi-gram Frequency Distribution

In this section, we analyze the frequency of bigrams—two consecutive letters—found within the Arabic language corpus. A bigram analysis is essential for understanding letter pairing patterns and exploring how often specific combinations of letters appear together.

F. Bi-gram Frequency Interpretation

Looking at this table of the 30 most frequent Arabic bigrams, we can provide several significant observations:

1) Highly skewed distribution: The most frequent bigram ال accounts for 8.62% of all bigrams, which is more than 4 times the frequency of the second most common bigram. This creates a "long tail" distribution where a small number of bigrams occur with extremely high frequency.

2) Top 30 Bigrams: The top 30 bigrams collectively represent approximately 28.5% of all Arabic bigram occurrences in the corpus, highlighting how a relatively small set of character combinations forms a substantial portion of written Arabic.

3) Definite article dominance: The most frequent bigram ال corresponds to "al" which is the Arabic definite article (similar to "the" in English). This reflects its crucial grammatical role in the language.

4) Negation patterns: The second most common bigram لا is used for negation ("la" meaning "no" or "not"), showing how frequently negation occurs in Arabic writing.

5) Key grammatical particles: Several high-ranking bigrams represent common grammatical elements:

- في (fi) meaning "in" (ranked 10th)
- من (min) meaning "from" (ranked 11th)
- ما (ma) which can function as a question word or negation particle (ranked 15th)

6) Common prefixes and suffixes: : Many bigrams represent common affixes:

- ات (at) is a common feminine plural ending (ranked 4th)
- ين (in) is a common dual/plural ending (ranked 19th)
- ها (ha) is a common feminine possessive suffix (ranked 27th)

G. Trigram Frequency Distribution

Table 3 presents the frequency and percentage distribution of the most common Arabic trigrams.

H. Trigram Frequency Interpretation

The most striking feature of this trigram frequency table is the overwhelming dominance of trigrams beginning with the definite article "ال" (al). Out of the top 30 trigrams, 17 begin with "ال" (including ranks 1-12, 14-15, 17-21, 23, and 30). This pattern reveals fundamental characteristics of Arabic language structure.

1) Definite Article Combinations: The definite article "ال" (al) combines with nearly every letter of the Arabic alphabet in these high-frequency trigrams. This demonstrates how the definite article attaches to nouns beginning with different letters to form common trigrams.

TABLE II
FREQUENCY OF ARABIC BIGRAMS

Rank	Bigram	Frequency	%	Rank	Bigram	Frequency	%	Rank	Bigram	Frequency	%
1	ال	123,827,674	8.6153%	11	من	13,683,963	0.9521%	21	عل	10,140,305	0.7055%
2	لا	29,943,807	2.0833%	12	ام	13,176,843	0.9168%	22	نا	9,968,861	0.6936%
3	لم	23,228,081	1.6161%	13	را	13,038,886	0.9072%	23	دي	9,630,837	0.6701%
4	ات	20,406,270	1.4198%	14	ري	12,720,844	0.8851%	24	بي	9,002,684	0.6264%
5	يه	20,390,057	1.4186%	15	ما	12,409,513	0.8634%	25	ره	8,967,699	0.6239%
6	ان	19,353,643	1.3465%	16	با	11,363,320	0.7906%	26	له	8,711,546	0.6061%
7	لي	19,175,942	1.3342%	17	يا	11,094,908	0.7719%	27	ها	8,557,678	0.5954%
8	وا	17,321,445	1.2051%	18	لت	10,795,426	0.7511%	28	ني	8,494,140	0.5910%
9	ار	16,434,231	1.1434%	19	ين	10,403,436	0.7238%	29	ير	8,156,542	0.5675%
10	في	15,301,328	1.0646%	20	اس	10,187,224	0.7088%	30	لل	8,137,516	0.5662%

TABLE III
FREQUENCY OF ARABIC 3-GRAMS

Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%
1	الم	18,934,382	1.74%	11	الله	4,047,043	0.37%	21	الص	3,073,813	0.28%
2	الأ	17,207,832	1.58%	12	الج	3,754,376	0.34%	22	لام	2,983,858	0.27%
3	الت	8,554,130	0.78%	13	بال	3,729,707	0.34%	23	اله	2,925,490	0.27%
4	وال	6,906,213	0.63%	14	الف	3,584,828	0.33%	24	ليه	2,919,999	0.27%
5	علي	6,676,190	0.61%	15	الر	3,474,282	0.32%	25	لاس	2,918,933	0.27%
6	العـ	6,593,709	0.60%	16	اره	3,316,888	0.30%	26	اني	2,752,867	0.25%
7	الي	5,718,337	0.52%	17	الن	3,198,522	0.29%	27	رات	2,748,961	0.25%
8	الـ	5,336,290	0.49%	18	الد	3,189,561	0.29%	28	است	2,734,673	0.25%
9	الس	5,001,728	0.46%	19	الق	3,180,875	0.29%	29	لان	2,659,491	0.24%
10	الب	4,194,665	0.38%	20	الش	3,106,691	0.28%	30	الخ	2,532,691	0.23%

2) Conjunctions, Prepositions, and Pronouns: Many of the most frequent trigrams involve combinations of conjunctions, prepositions, and pronouns—often interacting with each other or with the definite article.

- والـ (wa+al, rank 4, 0.63%): The conjunction ”and” + the definite article (”and the”), forming a critical connective structure in Arabic discourse. This demonstrates how Arabic frequently links defined nouns and phrases in sequence.
- علي (ala+y/ali, rank 5, 0.61%): Either ”upon me” (combining the preposition على with the first-person pronoun suffix) or the proper name ”Ali”. This high-frequency trigram shows how prepositions fuse with pronouns in Arabic.
- بال (bi+al, rank 13, 0.34%): The preposition ”with/by” + the definite article (”with the/by the”), creating instrumental or adverbial phrases that establish relationships between elements in a sentence.

These functional trigrams highlight Arabic’s synthetic nature, where grammatical relationships are frequently expressed through the fusion of multiple elements rather than through separate words. This system creates efficient and compact expressions where conjunctions, prepositions, and pronouns blend seamlessly to form the connective tissue of Arabic discourse.

I. N-gram Frequency Distribution

In this section, we present the frequency distributions of Arabic n-grams for sequences ranging from 4-grams to 7-grams. Each table lists the most common 4-grams, 5-grams, 6-grams, and 7-grams in the dataset, along with their frequency percentages.

1) Semantic Shift Toward Meaningful Words: One of the most striking observations is the semantic shift that occurs as n-gram length increases:

- 4-grams: Still largely dominated by partial morphological

elements and word fragments (e.g., الاس, المو)

- 5-grams: Begin to show complete meaningful units (e.g., كساره "crusher", العرب "the Arab")
- 6-grams: Almost entirely composed of complete words (e.g., العربي "the Arab", كتروني part of "electronic")
- 7-grams: Predominantly full words or highly predictable word combinations (e.g., الكترون "electron/electronic", استخدام "usage")

This shift demonstrates how Arabic morphology consolidates into semantically meaningful units at the 5-7 character length, which aligns with the typical word length in written Arabic.

2) Structural Patterns: The definite article ال (al) continues to dominate even in longer n-grams, appearing at the beginning of many high-frequency sequences across all lengths. This confirms the fundamental grammatical importance of definiteness marking in Arabic discourse.

V. Conclusion

This study successfully analyzed the letter and n-gram frequencies of the Arabic language using a carefully curated dataset and systematic methodology. By focusing on n-grams ranging from n = 1 to n = 7, we captured key patterns in the structure of Arabic words, providing valuable insights for linguistic research and security-related applications such as cryptography.

The findings demonstrate that Arabic's unique script, its root-based word construction, and the connection between letters significantly influence letter frequency patterns, making language-specific analysis essential for accurate cryptographic work.

The documented datasets, procedures, and results not only serve as a foundation for future research but also offer a framework that can be expanded to deeper levels of n-gram analysis or adapted for other Arabic corpora.

Overall, this work contributes to a deeper understanding of Arabic text structure and its potential applications in various technical and academic fields.

References

[1] W. Stallings, Cryptography and Network Security, Pearson Education, 2017.

[2] M. Mayzner and M. Tresselt, Tables of Single-letter and Digram Frequency Counts for Various Word-length and Letter-position Combinations, Psychonomic Press, 1965.

[3] P. Norvig, "English Letter Frequency Counts: Mayzner Revisited," [Online]. Available: http://norvig.com/mayzner.html. [Accessed 26 April 2025].

[4] E. Mustafa and K. Bouzoubaa, "A Bi-Gram Approach for an Exhaustive Arabic Triliteral Roots Lexicon," Languages, vol. 8, no. 1, p. 83, 2023.

[5] A. A. Nafea, M. S. Muayad, R. R. Majeed, A. Ali, O. M. Bashaddadh, M. A. Khalaf, S. N. Abu Baker and A. Steiti, "A Brief Review on Preprocessing Text in Arabic Language Dataset: Techniques and Challenges," Babylonian Journal of Artificial Intelligence, p. 46–53, 2024.

[6] M. Aloui, H. Chouikhi, G. Chaabane, H. Kchaou and C. Dhaouadi, 101 Billion Arabic Words Dataset, arXiv, 2024.

[7] A. Elnagar, S. Yagi and A. Bou Nassif, "Systematic Literature Review of Dialectal Arabic: Identification and Detection," IEEE Access, vol. 9, pp. pp. 31010-31042, 2021.

[8] K. Shaalan, "Nizar Y. Habash, Introduction to Arabic natural language processing (Synthesis lectures on human language technologies," Machine Translation, pp. 285-289, December 2010.

[9] Boudelaa, S., Perea, M. & Carreiras, M. Matrices of the frequency and similarity of Arabic letters and allographs. Behav Res 52, 1893–1905 (2020).

[10] Elbarougy, R., Behery, G., El Khatib, A. (2020). A Proposed Natural Language Processing Preprocessing Procedures for Enhancing Arabic Text Summarization. In: Abd Elaziz, M., Al-qaness, M., Ewees, A., Dahou, A. (eds) Recent Advances in NLP: The Case of Arabic Language. Studies in Computational Intelligence, vol 874. Springer, Cham.

[11] Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 12, 2025. https://web.stanford.edu/ jurafsky/slp3.

[12] Habash, N. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers.

[13] I. A. Al-Kadit, 'ORIGINS OF CRYPTOLOGY: THE ARAB CONTRIBUTIONS', Cryptologia, vol. 16, no. 2, pp. 97–126, 1992.

[14] Jones, M.N., Mewhort, D.J.K. Case-sensitive letter and bigram frequency counts from large-scale English corpora. Behavior Research Methods, Instruments, & Computers 36, 388–396 (2004).

[15] D. R. Ridley and B. M. Lively, 'English Letter Frequencies and Their Applications: Part I', Perceptual and Motor Skills, vol. 96, no. 2, pp. 545–548, 2003.

[16] D. R. Ridley and B. M. Lively, 'English Letter Frequencies and Their Applications: Part II—Digraph Frequencies', Psychological Reports, vol. 95, no. 3, pp. 787–794, 2004.

[17] Apache Software Foundation, Apache License, Version 2.0, Jan. 2004. [Online]. Available: http://www.apache.org/licenses/LICENSE-2.0

[18] Intellaren, "A study of Arabic letter frequency analysis," Intellaren.com, 2020. http://www.intellaren.com/articles/en/a-study-of-arabic-letter-frequency-analysis

[19] G. Grigas and A. Juskeviciene, 'Letter Frequency Analysis of Languages Using Latin Alphabet', International Linguistics Research, vol. 1, p. 18, 03 2018.

[20] : Singh, S. (1999). The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography. Doubleday.

[21] : Kahn, D. (1996). The Codebreakers: The Story of Secret Writing. Macmillan.

[22] : Soro, G. (1547). De Cryptographia. (Early Renaissance cryptanalysis treatise).

[23] : Welchman, G. (1982). The Hut Six Story: Breaking the Enigma Codes. McGraw-Hill.

[24] : Turing, A. (1950). Computing Machinery and Intelligence. Mind, 59, 433-460.

TABLE IV
FREQUENCY OF ARABIC 4-GRAMS

Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%
1	الاس	2,438,016	0.31%	11	التي	1,332,574	0.17%	21	والم	1,014,772	0.13%
2	كسار	1,901,867	0.24%	12	المت	1,283,359	0.16%	22	شركه	996,761	0.13%
3	المو	1,788,925	0.23%	13	العر	1,177,673	0.15%	23	جديد	985,903	0.13%
4	الان	1,735,311	0.22%	14	المع	1,177,656	0.15%	24	المر	985,085	0.13%
5	المن	1,688,695	0.22%	15	ستخد	1,105,632	0.14%	25	لاست	983,282	0.13%
6	العا	1,670,646	0.21%	16	الات	1,104,604	0.14%	26	السي	974,247	0.12%
7	ساره	1,634,309	0.21%	17	المح	1,099,025	0.14%	27	الما	963,174	0.12%
8	ارات	1,572,895	0.20%	18	الال	1,096,588	0.14%	28	اليه	962,448	0.12%
9	الام	1,517,334	0.19%	19	التع	1,094,512	0.14%	29	الاو	957,365	0.12%
10	المس	1,405,149	0.18%	20	والا	1,087,911	0.14%	30	الدو	914,884	0.12%

TABLE V
FREQUENCY OF ARABIC 5-GRAMS

Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%
1	كساره	1,589,310	0.31%	11	الكتر	647,111	0.12%	21	معدات	554,856	0.11%
2	العرب	768,552	0.15%	12	لكترو	645,482	0.12%	22	عربيه	550,507	0.11%
3	مطحنه	747,977	0.14%	13	للبيع	616,986	0.12%	23	سعودي	548,657	0.11%
4	الاست	725,648	0.14%	14	الاول	612,542	0.12%	24	الدول	540,483	0.10%
5	لعربي	680,714	0.13%	15	الموا	592,605	0.11%	25	العام	539,414	0.10%
6	الحجر	676,674	0.13%	16	اليوم	588,553	0.11%	26	العمل	525,485	0.10%
7	المست	668,085	0.13%	17	لعالم	584,107	0.11%	27	تخدام	525,033	0.10%
8	العال	653,787	0.13%	18	ستخدم	579,329	0.11%	28	ستخدا	525,030	0.10%
9	كترون	649,496	0.13%	19	استخد	569,723	0.11%	29	منتدي	524,639	0.10%
10	تروني	648,370	0.12%	20	الانت	566,821	0.11%	30	لالكت	519,055	0.10%

TABLE VI
FREQUENCY OF ARABIC 6-GRAMS

Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%
1	العربي	664,538	0.21%	11	السعود	493,957	0.16%	21	الجديد	352,930	0.11%
2	كتروني	646,327	0.20%	12	لعربيه	472,474	0.15%	22	معلوما	349,543	0.11%
3	الكترو	644,730	0.20%	13	الرياض	432,115	0.14%	23	علومات	349,269	0.11%
4	لكترون	640,118	0.20%	14	سعوديه	427,090	0.14%	24	المتحد	313,140	0.10%
5	العالم	574,048	0.18%	15	التعلي	375,705	0.12%	25	لتعليم	312,573	0.10%
6	ستخدام	524,903	0.17%	16	الثاني	375,615	0.12%	26	اكتوبر	306,345	0.10%
7	استخدا	524,491	0.17%	17	اجتماع	375,462	0.12%	27	كسارات	304,523	0.10%
8	لالكتر	518,450	0.16%	18	برنامج	373,986	0.12%	28	الموقع	299,323	0.09%
9	الالكت	516,249	0.16%	19	منتجات	368,663	0.12%	29	لمتحده	295,700	0.09%
10	لسعودي	497,029	0.16%	20	الرئيس	367,660	0.12%	30	لتعدين	294,403	0.09%

TABLE VII
FREQUENCY OF ARABIC 7-GRAMS

Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%	Rank	N-gram	Frequency	%
1	الكترون	639,376	0.35%	11	التعليم	284,192	0.16%	21	الانترن	217,029	0.12%
2	لكتروني	637,070	0.35%	12	لاجتماع	268,447	0.15%	22	لانترنت	216,932	0.12%
3	استخدام	524,380	0.29%	13	الاسلام	267,558	0.15%	23	المملكه	206,883	0.11%
4	لالكترو	517,890	0.28%	14	التعدين	265,568	0.15%	24	لمعلوما	202,006	0.11%
5	الالكتر	515,842	0.28%	15	اجتماعي	264,932	0.15%	25	المعلوم	200,558	0.11%
6	السعودي	491,655	0.27%	16	الاجتما	264,588	0.14%	26	الانسان	200,523	0.11%
7	العربيه	457,595	0.25%	17	العالمي	235,230	0.13%	27	افريقيا	199,182	0.11%
8	لسعوديه	409,557	0.22%	18	لامارات	230,563	0.13%	28	الجديده	195,202	0.11%
9	معلومات	349,098	0.19%	19	الرئيسي	227,973	0.12%	29	المحمول	194,901	0.11%
10	المتحده	295,671	0.16%	20	الامارا	227,345	0.12%	30	لاسلامي	192,026	0.11%