[1] **Saviour Prakash Gnana Prakasam Louis Raja**

[2] **V.V.Ramalingam**

# Natural Language Processing using Latin Square Design to Examine Tamil Grammar for the Lyrics of the Lyricist with Efficiency.

Journal of Electrical Systems

***Abstract:*** - The concept of grammaralization provides an explanation for studying grammatical categories. This study examines the emergence and structuring of grammatical categories in a topologically generalized language. The vallinam and idaiyinam languages' nuances have been studied. This paragraph focuses on the use of the aforementioned grammatical framework in spoken and written dialogue. In addition, we have looked into every element that influences language development. Information theory, linguistics, and computer science employ the levenshtein distance, a string metric, to measure the dissimilarity amid two sequences. The levenshtein distance between two words is defined as the number of single-character modifications vital to alteration one word into another. By using quantitative methods for drawing trustworthy conclusions, controlled language researchers have progressed toward these goals. Researchers employed the Latin square design to compare the frequency of appearance of the three terms vallinam and idaiyinam in the compositions of various Tamil lyricists. The findings have led us to some reasonable conclusions.

***Keywords:*** Analysis of Variance (ANOVA), Idaiyinam, Latin Square Design, Levenshtein Distance, Lyricist, Tamil, Vallinam

## I. INTRODUCTION *(HEADING 1)*

Tamil Meyyeluthukal (consonants) are fundamental to the language's phonetic and grammatical structure. Here's how NLP (Natural Language Processing) is applied to study and analyze Meyyeluthukal: Using natural language processing techniques, the phonetic features of Meyyeluthukal may be examined. This necessitates familiarity with the phonetics of the language and the relationships between different consonant sounds. Phonetic analysis are often used by linguists to investigate the acoustic characteristics of Tamil consonants.

Based on their phonetic properties, we classify Tamil consonants into three main categories: Vallinam (Hard Consonants): க் (k), ச் (ch), ட் (ṭ), த் (t), ப் (p), ற் (ṟ). Mellinam (Soft Consonants): ங் (ṅ), ஞ் (ñ), ண் (ṇ), ந் (n), ம் (m), ன் (ṉ). Idayinam (Medium Consonants): ய் (y), ர் (r), ல் (l), வ் (v), ழ் (ḷ), ள் (ḻ).

Articulatory phonetics studies the production of these consonants. A significant amount of airflow obstruction produces Vallinam, resulting in a harder sound. Less obstruction leads to the production of Mellinam, which produces softer sounds. Moderate obstruction produces Idayinam, which produces sounds that are neither too rigid nor too soft.

Meyyeluthukal's pronunciation can change depending on their position in a word and the surrounding sounds. Aspiration: In certain contexts, one may aspirate certain consonants, followed by a burst of air. While these consonants are produced, the vocal cords may vibrate (voiced) or not (voiceless).

Here are some examples of Meyyeluthukal in words:

- க் (k): கல் (kal) - stone

- ஞ் (ñ): பஞ்சு (panju) - cotton

- ட் (ṭ): பட்டு (paṭṭu) - silk

- ந் (n): பந்து (pandhu) - ball

[1] *Corresponding author: Saviour Prakash Gnana Prakasam Louis Raja 1 Department of Computing Technologies, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Tamilnadu,603 203, India. saviourgl@live.com.

[2] V.V.Ramalingam 2 Department of Computing Technologies, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Tamilnadu,603 203, India. ramalinv@srmist.edu.in.

- ய் (y): மெய் (mey) - truth

- ழ் (l̤): தாழ்வு (thāl̤vu) - humility

Analyzing the phonetic properties of Meyyeluthukal involves challenges such as dialectal variations in different Tamil dialects. Coarticulation is The influence of surrounding sounds can alter the pronunciation of consonants. Due to technological limitations, accurate phonetic analysis necessitates advanced tools and techniques that may not always be available.

## II. Related Work

Your literature survey provides a comprehensive overview of the significant contributions made by statisticians and researchers in the field of language analysis using statistical approaches over the past 65 years. Here are some key points and additional insights based on your summary:

### *A. Early Contributions:*

- [1], [2], [3], [4], [5], and [6] laid the groundwork for statistical analysis in linguistics.
- [7] Continued this tradition, emphasizing the importance of quantitative methods in linguistic studies.

Tamil Language Analysis:

- Subbarayan (1995) applied the Warrens-Herdan distribution to analyze Peyarccolkal from the Muvar Tevaram, highlighting the use of statistical methods in Tamil language studies.

### *B. Recent Developments:*

- [8], [9] and [10] investigated the grammaticalization and evolution of Tamil grammar, contributing to our understanding of the Dravidian language family.
- [11], [12], and [13]G. L. Saviour Prakash & Ramalingam (2024) have made recent contributions to the application of statistical methodologies in Tamil linguistics.
- [14] Emphasized the importance of quantitative methods in language studies, particularly focusing on special letters or consonant-vowel ligatures in Tamil.

### *C. Statistical Methods in Linguistics:*

- [15] Levenshtein Distance algorithm is being utilized to provide query suggestions in a web-based drugs e-Dictionary, enhancing its functionality and efficiency.
- [16] Discussed the use of t-tests and ANOVA for examining linguistic structures.
- [17] Explored the use of ANCOVA and MANCOVA to analyze mean differences in grammatical structures.

### *D. Challenges and Future Directions:*

The analysis of special letters, consonant-vowel ligatures, and underutilized consonant clusters remains a critical area of study. The development of new alphabets and the evolution of current Tamil grammar forms are ongoing research areas. Our survey highlights the evolution of statistical approaches in linguistic studies and underscores the importance of quantitative methods in understanding language structures.

## III. BUILDING A DATA FRAMEWORK

Through their poetic works, several Tamil lyricists have made significant contributions to the structure and richness of Tamil grammar. Here are some notable figures:

### *A. Pulavar Pulamaipithan*

Pulavar Pulamaipithan is renowned for his deep understanding of Tamil grammar and ability to weave complex grammatical structures into his lyrics. His works often reflect a profound respect for classical Tamil literature.

*B. Gangai Amaran*

Gangai Amaran, a versatile lyricist and composer, has contributed to the modern Tamil music scene. His lyrics often incorporate contemporary language while maintaining grammatical integrity, making them accessible yet rich in meaning.

*C. Kannadasan*

Kannadasan is one of the most celebrated Tamil lyricists. His mastery over the language and its grammar is evident in his songs, which often explore philosophical and emotional themes. His ability to use simple yet profound language has left a lasting impact on Tamil literature.

*D. Panchu Arunachalam*

Panchanathan Arunachalam, known as Panchu Arunachalam, has penned numerous songs that are both grammatically sound and emotionally resonant. His contributions have enriched Tamil cinema and music.

*E. Vairamuthu Ramasamy*

Vairamuthu is a modern-day poet and lyricist whose works are known for their poetic beauty and grammatical precision. He has received numerous awards for his contributions to Tamil literature and cinema.

*F. Muthulingam*

Muthulingam's lyrics are known for their simplicity and adherence to grammatical norms. His works often reflect everyday life and emotions, making them relatable to a wide audience.

These lyricists have not only enriched Tamil music with their creative expressions, but they have also contributed to the preservation and evolution of Tamil grammar. Do we have a favorite lyricist or a particular song that we find grammatically absorbing Table.1?

*TABLE 1 DETAILS OF LYRICIST*

| S.no | Authors / Lyricists | Given Name |
|---|---|---|
| 1. | Pulavar Pulamaipithan. | A1 |
| 2. | Gangai Amaran | A2 |
| 3. | Kannadasan | A3 |
| 4. | Panchanathan Arunachalam | A4 |
| 5. | Vairamuthu Ramasamy | A5 |
| 6. | Muthulingam | A6 |

## IV. DATASET FOR LYRICIST

Our research on Tamil lyrics from the early years of the Talkies era is fascinating! Analyzing the Meyyeluttukal sequences from the lyrics of six prominent Tamil lyricists over the last six months must have provided some insightful data. Here's a brief overview of the key features and statistical information you might have find in our analysis Table.2:

*TABLE 2 BASIC STATISTICAL INFORMATION ON THE LYRIC'S ATTRIBUTES.*

| Information pertaining to the metrics of the lyrics | |
|---|---|
| Years range | 1952 – 2022 |
| Songs / Lyrics | 156 |
| Vocabulary / words | 8036 |

| | |
|---|---|
| Authors / Lyricists | 06 |
| Average Lyricists / lyrics | 20 |
| Average Lyrics / vocabulary | 136 |
| Minimum Lyrics / vocabulary | 39 |
| Maximum Lyrics / vocabulary | 170 |

*A. Vallinam And Idaiyinam, Two Elementary Data Structures, Described*

Three primary groups in Tamil grammar categorize consonants: Vallinam, Mellinam, and Idaiyinam. Let's focus on Vallinam and Idaiyinam:

**Vallinam (Hard Consonants)**

In Tamil, the hard consonants are referred to as Vallinam. Pronouncing these consonants with more force is crucial in maintaining the phonetic structure of the language. There are six Vallinam consonants:

க் (k), ச் (ch), ட் (ṭ), த் (t), ப் (p), ற் (ṟ).

These consonants are characterized by their strong, explosive sounds, which are essential for the correct pronunciation and meaning of words.

**Idaiyinam (medium consonants)**

The medium consonants, pronounced with moderate force, are known as idaiyinam. These consonants serve as a bridge between the hard and soft consonants, providing a balanced sound. There are six Idaiyinam consonants:

ய் (y), ர் (r), ல் (l), வ் (v), ழ் (zh), ள் (ḷ)

These consonants are important for the fluidity and rhythm of the Tamil language.

**Comparison and Usage**

Words use Vallinam consonants to convey strength and emphasis. For example, the word "கடல்" (kadal) meaning "sea" uses the Vallinam consonant "க்".

Idaiyinam consonants provide a softer, more flowing sound. For instance, "வாழ்" (vazh) meaning "live" uses the Idaiyinam consonant "ழ்".

Understanding these primary data structures is essential for mastering Tamil pronunciation and grammar.

*B. Vallinam's and Idaiyinam Data Structure*

To determine the total number of letters in a word, the algorithm.1 technique is used. When read from the text file Tamilsong.txt, the lyrics are referred to as a string array. Furthermore, grammatical structure is assessed by comparing and contrasting writers' styles and use.

1. **Read_file:** Reads the content of the file "TamilSong.txt".
2. **Character_Count:** Counts the occurrences of each character in the provided list (characters) within the song.
3. **Vallinam and Idaiyinam:** Lists of Vallinam and Idaiyinam consonants.
4. **Song:** The content of the song file.
5. **Vallinam_count and Idaiyinam_count**: Dictionaries storing the counts of each Vallinam and Idaiyinam consonant.
6. **Print:** Outputs the counts of Vallinam and Idaiyinam consonants.
7. This should help you accurately count the occurrences of Vallinam and Idaiyinam consonants in your Tamil song.

*ALGORITHM 1 METHOD FOR VALLINAM AND IDAIYINAM COUNTING*

## V. EMPRICAL ANALYSIS

*A. Distance between Vallinam and Idaiyinam according to Levenshtein*

- Matrix Initialization: A 2D list dp is created to store the distances. The size of the matrix is (len(word1) + 1) x (len(word2) + 1).
- Base Cases: The first row and column are filled through indices, instead of the cost of changing an empty string to the other string by insertions or deletions.
- Dynamic Programming: The matrix is filled by comparing characters of word1 and word2. If the characters match, the cost is the same as the diagonal value. If they don't match, the cost is the minimum of the three possible operations (insertion, deletion, substitution) plus one.
- Result: The value at dp[len(word1)][len(word2)] gives the Levenshtein distance.

Word 1 = 'க்ச்ட்த்ப்ற்'

Word 1 = 'ய்ர்ல்வ்ழ்ள்'

Levenshtein (word 1, word 2)

6.0

This implementation efficiently calculates the Levenshtein distance using dynamic programming.

*B. Latin Square Design and Data Representation*

Latin Square Design is a statistical method used to control for two blocking variables, reducing the number of experimental units required. Here's a detailed explanation:

**a) Key Features of Latin Square Design**

Blocking Variables: Unlike randomized block designs that control for one blocking variable, Latin Square designs control for two.

b) **Reduction in Experimental Units:** This design significantly reduces the number of experimental units needed. For example, a simple random design with six treatments would require 216 units, but a Latin Square design only needs 36 units, a 75% reduction.

**c) Assumptions**

- **Minimal Interaction Terms:** Assumes that interaction terms are minimal and can be ignored.
- **Main Effects Only:** Focuses on the main effects of treatments, row factors, and column factors on the response.

**d) Randomization Procedure**

Pick a Design: Select a design randomly from accessible orthogonal designs.

- **Allot Row Points:** Randomly allocate levels of the row aspect to the rows.
- **Allot Column Points:** Randomly allocate levels of the column aspect to the columns.
- **Allot Treatments:** Randomly allocate treatments to the treatment numbers or letters.

**Example with Six Treatments**

- **Treatments:** A1, A2, A3, A4, A5, A6
- **Experimental Units**: Reduced from 216 to 36 units.
- **Orthogonal Latin Squares:** Combining two orthogonal Latin Squares can further reduce the number of units required.

**e) Experimental Layout**

Here's an example layout for a six-treatment Latin Square experiment:

This table.3 ensures that each treatment appears exactly once in each row and column. GS stands for the grammatical structure of the things mentioned before. In Table 3, we can see how the author employs the grammatical structures.

*TABLE 3 SIX TREATMENTS OF LATIN SQUARE |DESIGN*

| GS | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| R1 | A1 | A2 | A3 | A4 | A5 | A6 |
| R2 | A2 | A3 | A4 | A5 | A6 | A1 |
| R3 | A3 | A4 | A5 | A6 | A1 | A2 |
| R4 | A4 | A5 | A6 | A1 | A2 | A3 |
| R5 | A5 | A6 | A1 | A2 | A3 | A4 |
| R6 | A6 | A1 | A2 | A3 | A4 | A5 |

**f) Benefits**

- **Efficiency:** Drastically reduces the number of experimental units.
- **Control:** Effectively controls for two blocking variables.

**g) Conclusion**

Latin Square designs are powerful tools in experimental design, especially when dealing with multiple factors. They provide a structured way to minimize experimental units while controlling for two sources of variability.

*C. Latin Square Design for Vallinam and Idaiyinam*

The Latin Square Design is a powerful tool for studying two-directional variations, which is particularly useful when comparing objects with multiple factors. In this context, we can use it to analyze the variations between Vallinam (hard consonants) and Idaiyinam (medium consonants) in Tamil.

- **Vallinam and Idaiyinam**

Vallinam (க்ச்ட்த்ப்ற்) - Hard consonants, often referred to as chest tone.

Idaiyinam (ய்ர்ல்வ்ழ்ள்) - Medium consonants, often referred to as neck tone.

- **Data Representation**

To illustrate the use of Latin Square Design, we can create a matrix where rows represent Vallinam and columns represent Idaiyinam. Here's an example layout:

From the segmented picture, the suggested model calculates the size of the drawing and extracts its characteristics. The approach creates data augmentation drawings of varying sizes based on the sketch characteristic. For this purpose, we have trained and tested with an enhanced feature set from Fig.1.

| Authors | Vallinam | | | | | | Idaiyinam | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | க் | ச் | ட் | த் | ப் | ற் | ய் | ர் | ல் | வ் | ழ் | ள் |
| A1 | 256 | 80 | 80 | 80 | 80 | 80 | 80 | 147 | 268 | 10 | 10 | 140 |
| A2 | 276 | 37 | 37 | 37 | 37 | 37 | 37 | 77 | 194 | 10 | 10 | 88 |
| A3 | 187 | 25 | 25 | 25 | 25 | 25 | 25 | 127 | 290 | 35 | 35 | 195 |
| A4 | 267 | 72 | 72 | 72 | 72 | 72 | 72 | 134 | 282 | 24 | 24 | 160 |
| A5 | 239 | 80 | 80 | 80 | 80 | 80 | 80 | 105 | 280 | 23 | 23 | 96 |
| A6 | 221 | 74 | 74 | 74 | 74 | 74 | 74 | 127 | 250 | 38 | 38 | 135 |

*FIGURE 1 BELOW SHOWS THE DATA STRUCTURE WITH REGARD TO MELLINAM AND VALLINAM.*

We may find the typical data structure for Vallinam and Idaiyinam in Algorithm 1.

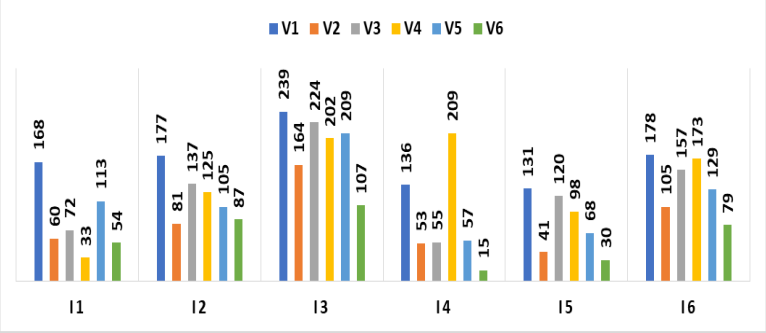


*FIGURE 2 PARENT DATA VALLINAM AND IDAIYINAM*

## VI. RESULT AND DISCUSSIONS

The authors, Vallinam, and Idaiyinam, have created an ANOVA table to analyze the results of their study. The table contains columns of six, with the row's assessment Vallinam in the head Mellinam column and the treatment plan in the second column. The experiment is designed according to the guidelines for randomization, and the results are used to replace the random numbers in A6. The LM method is applied to the data, with the first row of factors fixed, the next column fixed, then the first treatment (Treatment 1) fixed. The answer variable is A6.

To run the ANOVA, the "Up to 1-Way" option is selected in the Model window. This forces the tool to generate an incorrect interaction term, resulting in visible results. The authors, Vallinam, and Idaiyinam's ANOVA table is shown in Table.4. The authors' ANOVA table is a valuable tool for understanding the relationship between treatment and the model window. By following these guidelines, the results of the study can be analyzed and interpreted in a more accurate and meaningful way.

*TABLE 4 ANALYSIS OF VARIANCE TABLE (BASED ON 4.B)*

| Sources of variation | Degrees of freedom | Sum of squares | Mean Sum of squares | $F_0$ | $F_e$ |
|---|---|---|---|---|---|
| Vallinam | 5 | 52507.8 | 10501.56 | 21.18 | 2.71 |
| Idaiyinam | 5 | 49499.8 | 9899.96 | 19.97 | 2.71 |
| Authors | 5 | 8025.81 | 1605.16 | 3.23 | 2.71 |
| Error | 20 | 9914.22 | 495.711 | * | * |
| Total | 35 | 119948 | * | * | * |

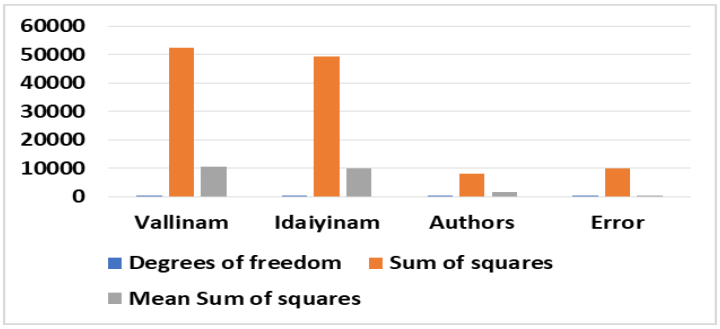


FIGURE 3 ANOVA TABLE OF VALLINAM, IDAIYINAM AND AUTHORS

Fig. 3 results from the analysis of variance show that there are notable variations in the grammatical structures used by Tamil lyricists. A significantly higher F-value (Fo) of 21.18 than the predicted F-value (Fe) of 2.71 was discovered for Vallinam phrases. It seems that the usage of Vallinam words varies significantly among rows. In a similar vein, the observed F-value of 19.97 for Idaiyinam words is higher than the predicted F-value of 2.71; this suggests that their usage varies significantly between columns. Additionally, the authors' research reveals that the anticipated F-value is 2.71 and the actual F-value is 3.2, suggesting that there is a considerable variance in the way various writers utilize the words Vallinam and Idaiyinam. These findings bring attention to the fact that Tamil lyricists' usage of grammar patterns has changed significantly over the last 65 years.

## VII. CONCLUSIONS

Examining the grammatical patterns utilized by Tamil lyricists over the last half-century has resulted to intriguing findings in the research. The grammar of the lyricist is rooted in the traditions of Idaiyinam, Vallinam, and Mellinam. The words are shaped by these form components. For this strategy, we consulted the canons of six lyricists whose impact has grown over time. There is a unique structure to the data system. It is worth mentioning that there has been a dearth of research on quantitative analysis of grammar structure in recent years. For this in-depth analysis of grammar, we relied on the Latin Square Design Model and the Levenshtein Distance. Our analysis of the Levenshtein distance reveals that there is a 6.0 distance between each of the three evaluations involving Vallinam, Mellinam, and Idaiyinam. Keep in mind that the writers of Vallinam and Idaiyinam use terms with different grammatical structures, and that this difference is substantial enough to warrant mentioning. There are significant variations in the usage of grammar by Tamil lyricists, as the research makes abundantly obvious.

## ACKNOWLEDGMENT

We are grateful to the Lord God Almighty for His unending grace, to SRM IST Institutions and the Professor who provided us with the necessary knowledge and advice to successfully complete this paper, and to our loving family who consistently provided financial support and encouragement, enabling us to truly excel in this research paper. Without the support of these individuals, the feasibility and value of this researcher's work would not have been possible. We would like to say, "Thank you, and May God Almighty bless us all," to show our appreciation.

## REFERENCES

[1] W. W. Greg and G. U. Yule, “The Statistical Study of Literary Vocabulary,” *Mod. Lang. Rev.*, vol. 39, p. 291, 1944.
[2] F. Mosteller and D. Wallace, “Inference and disputed authorship: The Federalist.(1964),” 1964.
[3] G. Herdan, “Quantitative linguistics,” 1964.
[4] L. Doležel, “A framework for the statistical analysis of style,” *Stat. style*, pp. 10–35, 1969.
[5] C. B. Williams, *Style and vocabulary: numerical studies*. Griffin, 1970.
[6] M. Mepham, “Introduction to the Mathematics of Language Study, by Barron Brainerd. (Mathematical Linguistics and Automatic Language Processing, 8). New York: American Elsevier, 1971. Pp. ix + 313.,” *Can. J. Linguist. Can. Linguist.*, vol. 18, no. 2, pp. 181–183, 1973, doi: DOI: 10.1017/S0008413100007428.
[7] S. T. Gries, *Statistics for Linguistics with R: A Practical Introduction*. De Gruyter Mouton, 2021. doi: doi:10.1515/9783110718256.
[8] S. Steever, “Verb + verb sequences in Dravidian,” 2021, pp. 327–353. doi: 10.1093/oso/9780198759508.003.0013.
[9] K. Sarveswaran, G. Dias, and M. Butt, “ThamizhiMorph: A morphological parser for the Tamil language,” *Mach. Transl.*, vol. 35, no. 1, pp. 37–70, 2021, doi: 10.1007/s10590-021-09261-5.
[10] V. Renganathan, “Expressives in Sangam, Medieval and Modern Tamil,” in *Expressives in the South Asian Linguistic*

*Area*, Brill, 2020, pp. 125–153.

[11] S. Raja and R. Venkatesan, "The grammatical structure used by a Tamil lyricist: a linear regression model with natural language processing," *Soft Comput.*, vol. 27, pp. 1–11, Oct. 2023, doi: 10.1007/s00500-023-09263-w.

[12] Z. Zhang, K. Lasocki, Y. Yu, and A. Takasu, "Syllable-level lyrics generation from melody exploiting character-level language model," *EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Find. EACL 2024*, pp. 1336–1346, 2024.

[13] V. V. G. L. Saviour Prakash, Ramalingam, "To investigate the occurrence of tamil lyricists' words throughout poems using truncated Poisson distribution." AIP Conf. Proc. 3075, 020019 (2024), 2024. doi: https://doi.org/10.1063/5.0217119.

[14] "Announcement," *Rev. English Stud.*, vol. 75, no. 320, p. 279, Sep. 2024, doi: 10.1093/res/hgae052.

[15] H. T. Sadiah, M. Saad Nurul Ishlah, and N. Najwa Rokhmah, "Query Suggestion on Drugs e-Dictionary Using the Levenshtein Distance Algorithm," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 10, no. 3, p. 193, 2019, doi: 10.24843/lkjiti.2019.v10.i03.p07.

[16] T. Toprak, "Analysis of differences between groups," 2019, pp. 179–197. doi: 10.4324/9781315187815-9.

[17] Z. Li and M. Chen, "Application of ANCOVA and MANCOVA in language assessment research," 2019, pp. 198–218. doi: 10.4324/9781315187815-10.