Esperanto: A Cross-Linguistic Examination with Turkish and Spanish

Mayank Kumar
Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen
Keplerstraße 2, 72074 Tübingen, Germany
mayank.kumar@student.uni-tuebingen.de

Abstract

This report examines the hypothesis that Esperanto, a constructed international language, has evolved to exhibit more similarities with Turkish—a non-Indo-European language—than with Spanish, a European language closely related to Esperanto's primary lexical contributors. Utilizing sentence pairs translated among Esperanto, Turkish, and Spanish, this study employs a series of linguistic metrics, including average word length, character-word ratio, and direct comparison of character and word counts between languages. My analysis, based on statistical computations and visual data representations, indicates that contrary to the initial hypothesis, Esperanto shares closer structural characteristics with Spanish than with Turkish.

1 Introduction

Esperanto(Janton, 1993), created in 1887 by L. L. Zamenhof, was designed as a neutral, easy-to-learn language to foster international communication and understanding. Drawing primarily from Indo-European languages, particularly Romance and Germanic languages, Esperanto features a highly regular grammar, and a vocabulary(Forster, 1982) that is largely recognizable to speakers of European languages.

This report aims to test the hypothesis that Esperanto has become increasingly similar to Turkish, distancing itself from the European language characteristics it originally shared, especially with languages like Spanish. By examining structural and lexical features across translated sentence pairs in Esperanto, Turkish, and Spanish, this study seeks to understand whether Esperanto's usage aligns more closely with the non-European structural patterns of Turkish or remains consistent with its European roots, as exemplified by Spanish.(Comrie, 1989)

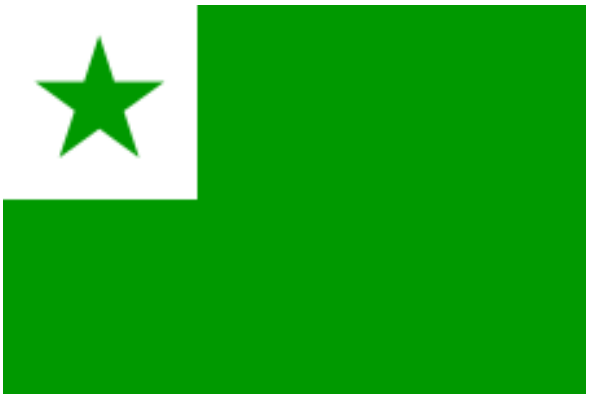

Figure 1: Esperanto Flag

2 Data

2.1 Data Acquisition

The datasets used in this study were obtained from Tatoeba(Tiedemann, 2020), a large database of example sentences aimed at language learners. Tatoeba's datasets are particularly valuable for linguistic research due to the extensive range of languages covered and the availability of parallel sentences directly translated by contributors. For the purpose of this analysis, we specifically utilized datasets consisting of parallel sentences in Esperanto-Turkish and Esperanto-Spanish.

2.2 Data Prepration and Filtering

Following the verification of data integrity, the datasets were prepared for analysis. This involved merging the Esperanto-Turkish and Esperanto-Spanish sentence pairs based on the Esperanto sentences. By merging these datasets, we were able to directly compare how a single Esperanto sentence was translated into both Turkish and Spanish, thus providing a straightforward method for assessing linguistic similarity across the two languages. The merging was executed using standard Pandas data manipulation techniques, which ensured that the resultant dataset was well-structured for the analyt-

ical methods applied later in the study.

3 Method

3.1 Data Analysis Techniques

To evaluate the linguistic similarities between Esperanto, Turkish, and Spanish, several analytical techniques were employed to examine the structural and lexical characteristics of the sentences:

3.1.1 Average Word Length Calculation:

The average word length for each sentence in the three languages was calculated. This measure assesses the morphological complexity(Song, 2014) of words typically used in each language, which is indicative of linguistic richness and syntactic structure.

3.1.2 Character-Word Ratio:

This metric was computed by dividing the total number of characters by the number of words in each sentence for the three languages. The character-word ratio provides insights into the average word length in a different manner, emphasizing the compactness or expansiveness of the lexical items used.

3.1.3 Direct Comparison of Character and Word Counts:

The total number of characters and words in the Esperanto sentences were compared against their Turkish and Spanish counterparts. This approach quantifies the linguistic distance between Esperanto and each of the other two languages based on sentence length and word count, providing a straightforward metric of similarity.

3.2 Statistical Analysis

Average values were computed across the dataset for each language. This provided a baseline from which to compare the linguistic features of Esperanto with those of Turkish and Spanish. Statistical tests were used to determine whether the differences observed in these metrics were statistically significant, thereby providing a robust basis for any conclusions drawn from the data.

3.3 Comparative Metrics

To further refine the analysis:

- Absolute differences in character counts and word counts between the Esperanto sentences and their corresponding translations in Turkish and Spanish were calculated.
- These metrics were applied individually to each sentence pair to determine which language, Turkish or Spanish, exhibited a closer structural resemblance to Esperanto on a sentence-by-sentence basis.

3.4 Visualization

Findings were visualized using bar graphs to represent the average word lengths and character-word ratios, facilitating an intuitive understanding of the differences and similarities between the languages. Additionally, count plots were used to illustrate the frequency with which each language (Turkish vs. Spanish) aligned more closely with Esperanto based on the character and word count comparisons.

4 Results

The statistical analysis and data visualization performed on the datasets reveal significant insights into the linguistic characteristics of Esperanto, Turkish, and Spanish. The findings indicate substantial differences in word length and character-word ratio among these languages, with distinct patterns emerging between Turkish and the other two languages.

4.1 Average Word Length

The analysis shows that Turkish sentences tend to have the longest average word length at 6.122, followed by Spanish at 4.838, and Esperanto at 4.713. This difference is statistically significant, as confirmed by the t-tests, where Turkish notably differs from both Esperanto and Spanish. This suggests that Turkish, being an agglutinative language, naturally exhibits longer words due to the affixation of morphemes to form words, which is less prevalent in the more analytic structure of Spanish and the constructed simplicity of Esperanto.

4.2 Character-Word Ratio

Similarly, the character-word ratio further substantiates the morphological complexity of Turkish compared to Esperanto and Spanish. Turkish has the highest ratio at 6.81, indicating that its words, on average, contain more characters per word. This is in contrast to Spanish (5.85) and Esperanto (5.47), where the words are shorter. The t-test results between Esperanto and Turkish and between Esperanto and Spanish both show significant differences, reinforcing the distinctive linguistic structure of Turkish.

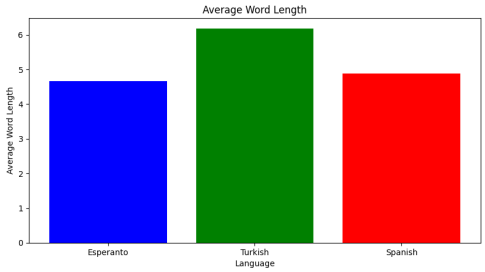

Figure 2: Average Word Length

The bar graph depicting average word lengths visually illustrates the differences observed numerically. Turkish bars stand taller than those for Esperanto and Spanish, graphically representing Turkish's tendency towards longer word forms.

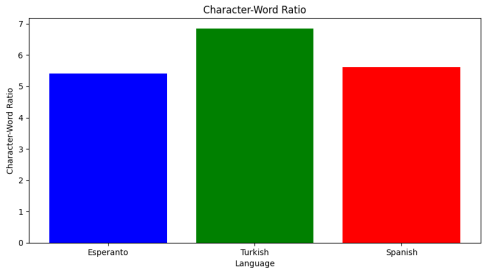

Figure 3: Character-Word Ratio

The character-word ratio bar graph mirrors the findings of the average word length graph, with Turkish again showing higher values compared to Esperanto and Spanish. This visual representation highlights the denser word structure in Turkish, which accommodates more characters within each word on average.

4.3 Comparative Analysis

By merging datasets and analyzing sentences that have been translated from Esperanto to both Turkish and Spanish, a direct comparative approach was possible. The absolute differences in character counts and word counts clearly demonstrate that Esperanto sentences align more closely with Spanish than with Turkish, both in terms of the number of characters and words used. The lesser differences in these metrics between Esperanto and Spanish suggest a closer linguistic affinity, likely due to the shared Indo-European roots and similar grammatical structures.

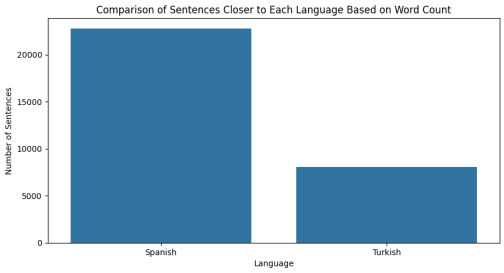

Figure 4: Comparison of Sentences Closer to Each Language Based on Word Count

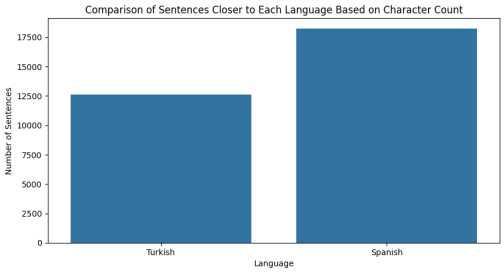

Figure 5: Comparison of Sentences Closer to Each Language Based on Character Count

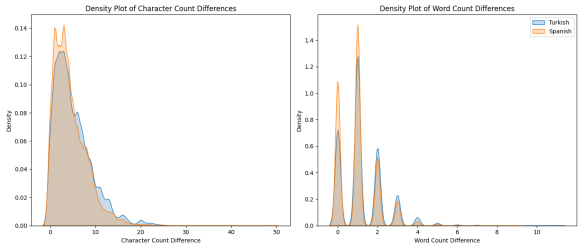

Figure 6: Density plot of character and word count difference

5 Conclusion

The results support the conclusion that Turkish, with its agglutinative characteristics, stands apart from both Esperanto and Spanish, which are more similar to each other. Despite Esperanto's constructed nature and intent as a bridge language, it retains closer structural ties to European languages, particularly Spanish, than to the distinct linguistic framework of Turkish. This finding challenges the initial hypothesis that Esperanto would show a convergence towards Turkish linguistic features over time. Instead, Esperanto maintains stronger affiliations with its European linguistic heritage.

References

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Peter Glover Forster. 1982. *The Esperanto Movement*. 32. Walter de Gruyter.

Pierre Janton. 1993. *Esperanto: Language, literature, and community*. Suny Press.

Jae Jung Song. 2014. *Linguistic typology: Morphology and syntax*. Routledge.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.