Exploring Variation Across English, Finnish, Greek, and Portuguese Using Reality TV Captioning and Legal Documents as a Basis for Analysis

John Speaks
jspeaks2@illinois.edu

Abstract

This paper investigates register variation in web data across four languages—English, Finnish, Greek, and Portuguese—by comparing it to two well-defined registers: legal documents and reality TV subtitles. Building on the work of Li et al. (2022), this study utilizes legal texts from the MultiEURLEX corpus and reality TV captions as standardized corpora for comparison. Using Jaccard and cosine similarity measures, web documents are analyzed against these registers. The results confirm the presence of register-specific patterns across all languages, with legal and reality TV data forming identifiable clusters. This method offers a more consistent framework for cross-linguistic comparison of web data, with implications for future studies on linguistic variation.

1 Introduction

The aim of this exploration was to gather data about register variation in web data in different languages with as standardized of a methodology as possible. This was inspired by the paper "Register variation remains stable across 60 languages" (Haipeng Li and Nini, 2022) where they evaluated corpora across 60 languages to detect variation. One way that they looked at variation was by looking at similarity of web documents to Wikipedia and Twitter corpora in the same language. A problem with this approach however is that Twitter and Wikipedia are necessarily equivalent contexts across languages. In some languages Twitter may only be used for certain informal communities, in others it could be used for news and political rhetoric. The same could be said for Wikipedia, where only certain languages have a large variety of articles and others are specific to local histories or topics. This then means that comparing web corpora to these sources across languages, it may be a comparison to non-equal standards and an ineffective approach. This paper attempts to recreate the same analysis, but with more concrete standards for comparison.

Instead of Twitter and Wikipedia, this paper uses a comparison to Legal documents and reality TV captions. These are both datasets that have a more specific and limited scope and register which remains consistent across languages. Using these other datasets, web corpora for four different languages (English, Finnish, Greek, and Portuguese) were plotted in comparison to samples of reality TV subtitles and legal documents and then evaluated using both Jaccard and cosine similarity.

Similarly to the Li et al. paper, this exploration found that register variation is universal and identifiable. In every case reality TV and legal data were distinctly identifiable as different registers regardless of language and similarity measure. More information was also found of how each of the explored languages web presence varies relative to the standards of reality TV and legal documents.

2 Background on the Corpus

Three datasets were used in this exploration. In addition to the web scraped data to be evaluated, two datasets were used to anchor the comparison of web data in two highly contrasted, but very distinct categories: legal documents and reality TV subtitles. These two were chosen as anchors because both domains are very similar across languages and cultures, where other accessible corpora vary in use much more in different cultures. Unfortunately, due to the narrower nature of the datasets available for reality TV captions and legal data, data was only available across all datasets for a small number of western languages. Reality TV captions were the main limiting factor as there was not much data for many non-western languages, and when there was it was generally harder to find legal data. For that reason, the scope of this exploration was limited to four EU countries (chosen

to standardize with EU law), varying across European language groups. The languages chosen were English, Finnish, Greek, and Portuguese.

Reality TV show data for this experiment came from opensubtitles.com for a number of international shows like Too Hot to Handle, Love Island, and more. These shows were chosen because of their global reach, readily available captions, and spinoff shows in other languages. Due to lack of paid access to opensubtitles.com, the quantity of data that was collected for this exploration was not very large (only a few megabytes for each language), but enough was captured to explore similarity.

The Legal data was collected from the MultiEURLEX corpus available through Hugging Face. This corpus comprises over 65 thousand laws that are translated into each of the EU's 23 official languages. For a study limited to a selection of western languages, this was an ideal dataset because it offered standardization across languages and had all the languages needed for comparison. The dataset was much more than sufficiently large and provided a great metric for analyzing the web corpora.

The web scraped corpora that were being compared were collected from Professor Jonathan Dunn and the EarthLINGS dataset. This collection of corpora is a collection of web data separated by country and language for countries around the world. For this exploration, web data for the native speaking countries of the languages explored were used, namely English data from the UK, Finnish data from Finland, Greek data from Greece, and Portuguese data from Portugal.

3 Preparing the Data

Before starting any analysis, the data used had to be cleaned to varying degrees. The code detailing how the data is cleaned is in the GitHub repository associated with the paper[1]. The web data from EarthLINGS was pre cleaned and required little modification to a central standard. The reality TV show captions, by contrast, were not standardized between languages as they were written by different authors, and had many extraneous marks used to denote audible elements of the shows that had to be removed. This cleansing was done using regex patterns and the re library in Python. The legal documents had a very standard format but had significant metadata and header information about what the law was and other clerical information. By accessing the data through Hugging Face and not the original researchers' files, much of this data was removed. The rest was taken out through data manipulation using regex expressions as well.

[1]https://github.com/JTSIV1/register-variation-using-legal-and-reality-tv-data/

4 Calculating and Plotting Similarity

To explore the variation in web data across each language, for each language samples were taken from the datasets and compared for similarities. For each language, 40 samples were taken from the reality TV, legal, and web corpora. It was decided to take 40 samples because it captured a broad enough snapshot of the language and register, but was also computationally feasible with limited resources. Next, for each of these samples a similarity value was calculated for each of the three corpora. This was done by further sampling 10 items from each corpus and averaging the similarities relative to the original sample. It was chosen to calculate similarity using 10 samples to ensure that if an outlier was chosen it would not impact the data, and that instead of one document the similarity score was relative to the population as a whole. These 40 samples were plotted using the relative similarities to the reality TV and legal documents for each corpus. All of this was done once using Jaccard similarity and once using cosine similarity to explore different aspects of the data relationships. These calculations were all done in Python using the Pandas package to manipulate and access the corpora. Jaccard and cosine similarity was calculated using the Scikit Learn library. All of the code for these calculations is available in the paper's GitHub repository [1].

Error analysis was also performed to ensure that outliers in the data were real documents. With some of the web data, samples contained multiple languages and not just the target which was leading to unexpected and confusing results. These were manually removed when they came up and were not part of any final results presented.

5 Results and Evaluation

After similarity was calculated and plotted for a sample of the documents in the corpora for each language a number of patterns emerged. The graphs for each language and their cosine similarity are shown in figure 1. The same data but calculated with Jaccard similarity and displayed in figure 2.

In addition to each document plotted as a point and labeled by color, each plot also has a dashed line indicating a 1:1 ratio of similarity. This was included to highlight the boundary where a document is more similar to the reality TV captions or legal texts. To the left of the line, documents are more similar to the reality TV captions and to the right, the legal texts.

To explore the implications of the generated graphs, it is first important to understand the difference between the Jaccard and cosine similarity scores. Cosine scores capture the relative frequency of terms. Two documents might be considered similar even if they differ greatly in length, as long as the terms they share have similar relative frequencies. Jaccard similarity scores focus on the overlap between sets, helping it detect cases where the exact presence or absence of terms is important rather than their relative frequency.

Across all languages included in the study, with Jaccard similarity both web and reality TV data are relatively dissimilar to both reality TV and legal data, while the legal data stands out as fairly similar to itself. This is probably because legal texts would have fairly specialized vocabulary that are identified more readily by Jaccard similarity. When looking at the Jaccard approach, web data seems to lie mostly along the line of equality, meaning that most web documents are equally similar to legal or reality TV data, and clusters mostly in the same way as reality TV, indicating that the web data likely has a similar vocabulary to the reality TV and not the legal documents.

The cosine similarity in figure 1 shows a different picture however. In Greek and Portuguese, the web data varies and is spread between the reality TV cluster and the legal cluster showing that the data in the corpus has a variety of registers between the two anchor uses. In Finnish, the web data lies much closer to the TV data meaning that web data may be more like reality TV captions and less like legal documents as a whole. English had interesting results, with overall similarity being higher, but a small chunk of web data that is extremely dissimilar to both scales, suggesting the presence content that is of a completely separate form.

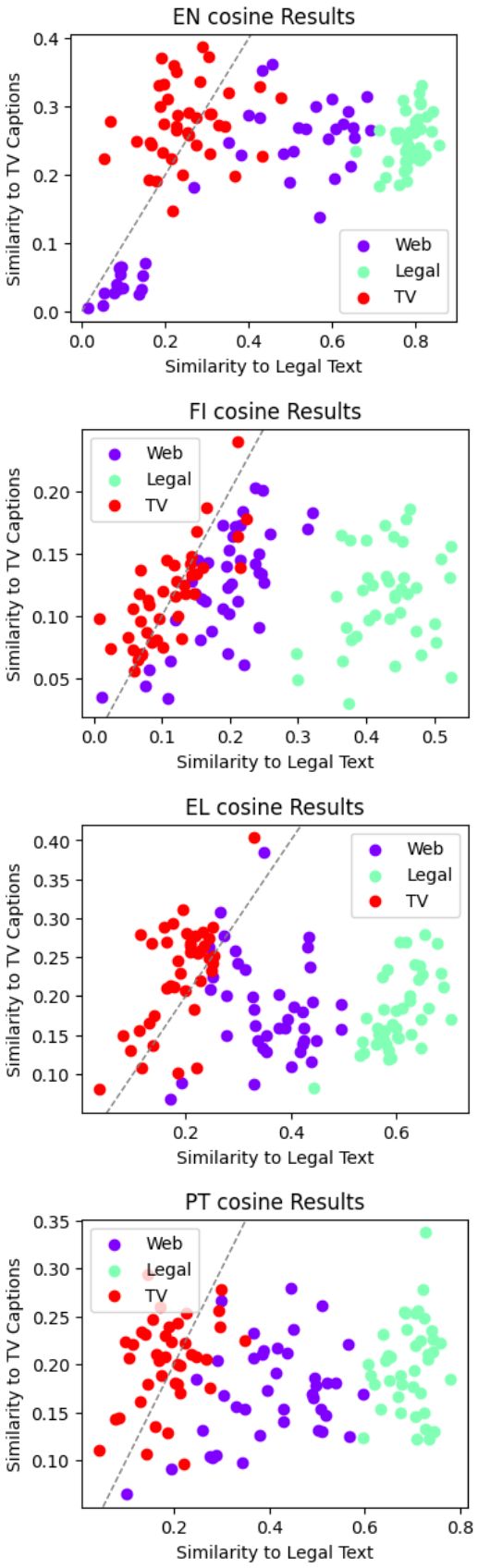

Figure 1: Cosine Similarity Scores for English (EN), Finnish (FI), Greek (EL), and Portuguese (PT)

6 Limitations and Future Explorations

This study opened paths for both improvements and further investigation. With a relatively limited size of reality TV corpus, issues with length

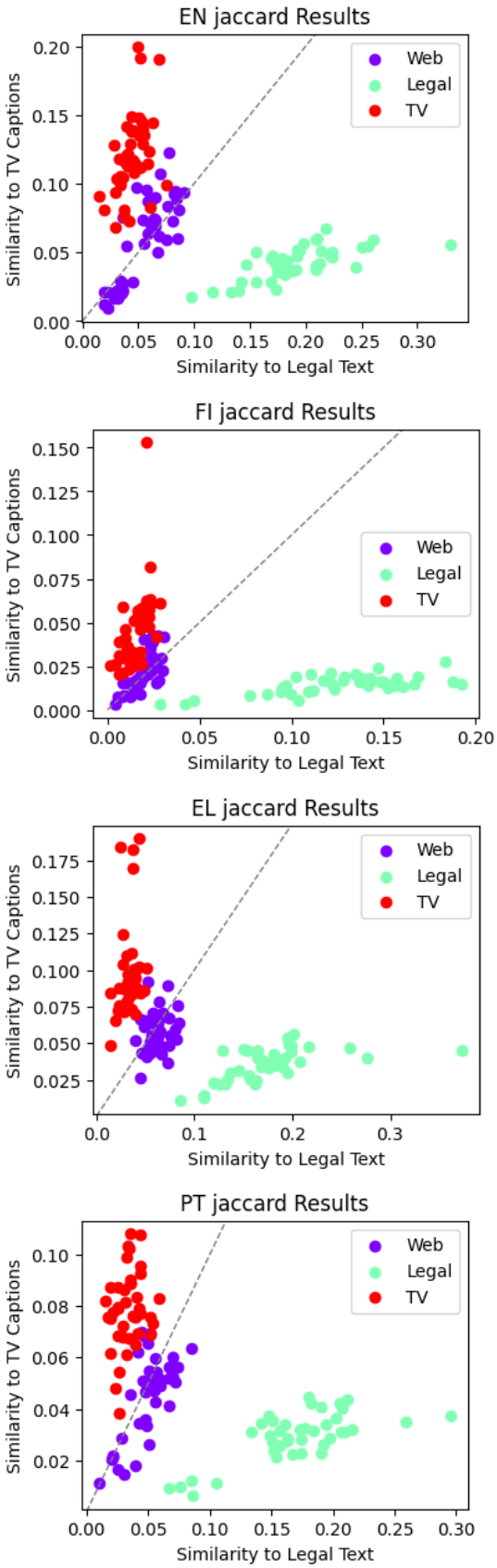

Figure 2: Jaccard Similarity Scores for English (EN), Finnish (FI), Greek (EL), and Portuguese (PT)

of documents or quality could not be completely eliminated because it got rid of too much data. Ensuring that documents could be longer and more standardized could certainly lead to better results.

Further studies could explore anchoring their comparisons in distinct register data in more countries. With a better dataset of specific but universal registers for more countries, more insight could be gained on how web data varies relative to those registers around the world. Additionally, this study assumed little to no variation in the register of reality TV and law across the world which may not be true with a broader scope. This could also be explored in future papers.

7 Conclusions

This study investigated register variation in web data across multiple languages by anchoring the analysis to two consistent and distinct registers: legal documents and reality TV subtitles. With this approach, and using similarity measures such as Jaccard and cosine, the paper identified distinct linguistic patterns across English, Finnish, Greek, and Portuguese, building upon prior research in register variation. The analysis showed that register variation is universal and identifiable, as legal and reality TV data were clearly distinguishable regardless of the language, offering a more standardized means of comparing web corpora than previous methods.

Additionally, the use of two distinct similarity metrics, Jaccard and cosine, provided insights into the variation present in web data. Jaccard similarity captured the overlap of specific vocabulary items, showing how legal texts are more self-similar across languages, while cosine similarity revealed the underlying frequency patterns within different registers, highlighting variation in web data, especially for Greek and Portuguese.

Although this study was limited by the scope of the datasets, especially the size of the reality TV corpora, the findings offer a solid foundation for future research. Expanding analysis to non-Western languages, obtaining larger standardized corpora, and exploring more nuanced registers in different contexts would further enrich our understanding of how language varies across web domains.

References

Jonathan Dunn Haipeng Li and Andrea Nini. 2022. Register variation remains stable across 60 languages. *Journal of Corpus Linguistics and Linguistic Theory*.