Detecting Structured Language Alternations in Historical Documents by Combining Language Identification with Fourier Analysis

Hale Sirin
Center for Digital Humanities
Johns Hopkins University
hsirin1@jhu.edu

Sabrina Li
Center for Digital Humanities
Johns Hopkins University
sli159@jhu.edu

Tom Lippincott
Center for Digital Humanities
Johns Hopkins University
tom@cs.jhu.edu

Abstract

In this study, we present a generalizable workflow to identify documents in a historic language with a nonstandard language and script combination, Armeno-Turkish. We introduce the task of detecting distinct patterns of multilinguality based on the frequency of structured language alternations within a document.

1 Introduction

This work emerges from the goal to create a corpus in Armeno-Turkish—vernacular Turkish written in Armenian script. This historic language was actively used from the early 18th century to the early 20th century in a variety of locations in the Middle East, Europe and the US, including Istanbul, Venice, Vienna and Boston (Der Matossian, 2020). There are lists of works in Armeno-Turkish available (Stepanyan, 2005), but searching HathiTrust manually points to the existence of works that are not recorded in these lists due to challenges in the bibliographical recording of these works in library catalogues (missing titles in the original script, wrong or missing language labels). Unable to collate works in Armeno-Turkish by using the metadata, we used language identification. This process did not produce a clean dataset in Armeno-Turkish, but showed that a significant portion of these works are multilingual. Based on this observation, this study is a first attempt at modeling some of the interesting multilingual phenomena with structured language alternations that emerge in this process: bilingual translations, dictionaries, original-language text followed by commentary in a different language, language study books.

As opposed to unstructured code switching (oral interviews), structured multilingual patterns involve an organized alternation of two or more languages. These language alternations may occur at different frequencies (every sentence, paragraph, every page, every chapter). A structured language alternation may serve various purposes, including making the content available in multiple languages in a legal document to reach its target audiences, as shown in the two-column Ottoman legal code in Armenian and Armeno-Turkish in Figure 1.

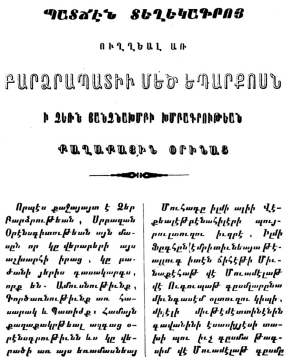

Figure 1: The first page of the Ottoman legal code, *Mejelle*, published in 1889 in a bi-column bilingual format, Armenian on the left and Armeno-Turkish on the right (mej, 1889).

Especially for historic languages, detecting structured language alternations is valuable for identifying clean monolingual segments which can then be used to create new resources for NLP tasks. This analysis also provides insight into material history of physical books and translation studies, by showcasing different formats of page segmentation in structured multilingual books (bi-column, top-bottom) (McConnaughey et al., 2017; Werner, 2012). This project introduces the task of detecting structured language alternations and makes the following contributions:

- We introduce an experiment that maps the language alternations in the time domain to the frequency domain to detect different patterns of structured language alternations in a corpus, and show that unsupervised clustering applied to the frequency spectra can be a simple and efficient first step in grouping documents with

Proceedings of LaTeCH-CLfL 2024, pages 46–50

different patterns of structured language alternations.

- We present a more comprehensive and nuanced Armeno-Turkish corpus from the HathiTrust Digital Repository, and compare the performance of a character n-gram model and a trained neural model for language identification of a historic language with a non-standard language and script combination.

2 Background

2.1 Language ID

Language identification is the task of determining the language(s) of a document and a crucial step in document classification in historical research. Character-based n-gram models are a performant statistical approach (Cavnar and Trenkle, 1994), and recent neural models offer a fast and efficient solution (Joulin et al., 2016).

While language identification of a document is mostly regarded as a solved task (McNamee, 2005), a near-perfect performance is only achieved when certain assumptions are made regarding the quantity and the quality of the data, and the monolinguality of the documents. However, historic and multilingual documents in low-resource languages motivate different approaches to language identification (Jauhiainen et al., 2018). Multilinguality of these documents is frequently overlooked, even though historic languages in non-standard scripts, such as Armeno-Turkish, represent territories that were predominantly multilingual. Despite the perceived monolingualism of 18th-and-19th-century books that are published in "national print-languages" (Anderson, 2006), multilingual activity persisted and even flourished during this period in commercial, legal, cultural and literary domains (Mende, 2023).

Research in multilingual language identification and code switching generally focuses on identifying the language, but not the relative location of these languages in each document (Lui et al., 2014). Kevers (2022) locates code switches, but primarily in multilingual documents when language diversity is unstructured. In this study, we focus on distinct patterns of structured language alternations that emerge in historical datasets (religious commentary, language study books, bilingual legal documents) in Armeno-Turkish, a low-resource language that falls outside the "national-print language" category.

2.2 Frequency Analysis

Discrete Fourier Transform (DFT) converts a time-domain sequence to a frequency domain sequence. It's defined by equation 1, mapping a sequence of N numbers $x_0, x_1 \ldots, x_{N-1}$ to a new sequence of N numbers, for $0 \leq k \leq N-1$.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i k n / N} \quad (1)$$

Fast Fourier Transform (FFT) is an efficient algorithm used to calculate the DFT, reducing the time complexity from $O(N^2)$ to $O(N \log N)$ (Cooley and Tukey, 1965). Fourier transform has a wide range of applications in NLP. In this study, we approached the probability of a language label as a discrete signal at 50-word time steps in a document. Figure 2 shows a simulation of this approach, representing an idealized alternation of one language and another language as an array of alternating 0s and 1s and plotting its frequency domain representation using the Fourier Transform.

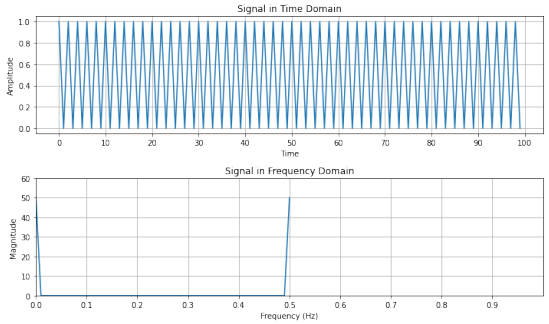

Figure 2: Time domain and frequency domain representations of an alternating discrete signal.

3 Materials and Methods [1]

3.1 Data

HathiTrust Digital Library (HT) (HathiTrust Foundation) offers unprecedented access to scholars who work with text as data in a variety of disciplines. However, corpus construction is a significant challenge when working with historic languages and multilingual documents (due to missing language and script information, OCR errors) and the overarching language label, even when it is correct, does not provide information regarding the multilingual composition of a given document.

[1] Code for the full experiment is available at https://github.com/comp-int-hum/Armeno-Turkish-Collection

The MARC 21 Bibliographic Record offers a limited structure to include script-related information, but it is not put to use systematically by cataloging librarians. The script information for the Armeno-Turkish works, if there is any, is occasionally found in the Notes section of the MARC record. This means that researchers' only recourse is to browse works in Armenian and in Turkish manually with the hopes of encountering Armeno-Turkish works.

3.2 Language ID Experimental Setup

We start by creating a training dataset of works labeled according to both language and script from the HT. For Armeno-Turkish, we use 97 expert-labeled documents. For negative examples, we first use the HT's MARC index to create a reverse mapping of languages (as assigned by librarians at the contributing institution) to documents, skipping anything dated before 1500 CE. We remove languages with less than 100 documents or whose code is not valid ISO-639. To ensure diverse temporal representation of each language, we split the range from the earliest to latest document in that language into 5 buckets covering equal time periods, and randomly select one document from each bucket. We then select an additional 5 documents at random from the overall remaining set, for a total of 10 documents per language. We split each document into sub-documents of contiguous script according to the Unicode script specification. This process results in 118 unique script and language combinations, serving as our training data. For the positive examples we only keep the Armenian script, since these are hand-annotated and does not rely on MARC metadata. At test time, we segment the document into smaller sections. By recording the segmentation offsets, the original documents can be reconstructed with the inferred language information.

We compare a trigram character language model with FastText neural language ID model (Joulin et al., 2016), trained on our labeled dataset in Table 1. We create a dataset of all documents in the HTC tagged as Turkish (tur), Ottoman Turkish (ota), or Armenian (arm), a total of 18367 records.

We apply the trained FastText model with a 0.91 fscore on the test set, to all documents in the HTC tagged as Turkish (tur), Ottoman Turkish (ota), or Armenian (arm) in 50-word windows. This process yields 95 works with Armeno-Turkish as the majority language label.

Model	Fscore (macro)
Multinomial NB	0.67
Char N-gram	0.83
Trained FastText	0.91

Table 1: Lang ID Model Performance Comparison

3.3 Frequency Analysis

The output of the FastText model is a grid of probability distributions over all possible language_script labels for each 50-word window in a document. In order to reach our goal of bringing out the periodic phenomena in these 95 works, we simplify this fairly noisy information. Since it is unclear how well-calibrated the model is, we calculate the majority language label of the whole document, and use the probability value of that language for each chunk. This language-agnostic approach radically simplifies the initial big grid of the full probability distribution, into a sequence of probabilities of that majority language for each 50-word chunk.

We transform each probability distribution array into a frequency domain, using FFT, and cluster each frequency spectrum using k-means clustering. Frequency domain transformation allows us to compare signals of different lengths.

4 Results and Discussion

The clustering of the frequency spectra yielded three coherent groups:

1. Works that are predominantly in Armeno-Turkish (59 documents)

2. Works that are bilingual, alternating between Armeno-Turkish and another language (10 documents: Language textbooks: alternating every few sentences, Bilingual editions: alternating every column or every page)

3. Works that are multilingual in languages other than Armeno-Turkish (13 documents: Language textbooks in Armenian-Greek, Armenian-Russian, Armenian-French)

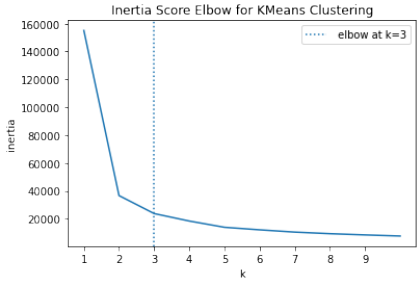

Figure 3: Visualization of k-elbow inertia metric for optimal k in k-means clustering.

As visualized in Figure 3, we selected the number of clusters based on the inertia metric for optimal k. While the clusters are relatively coarse-grained, this is a fast and efficient approach to historical datasets with unreliable metadata and high variation in genre and language composition. The frequency analysis combined with segmented language identification is a promising venue to explore documents in historic languages, since it lets us divide the corpora automatically into more distinct categories, revealing a variety of genres. Preserving the diversity of genres is valuable for low resource situations in which there is a risk of certain genres dominating the fine-tuning material.

This experiment identified 30 new records in Armeno-Turkish that were not in the training set, including translations of the Bible, dictionaries, textbooks for learning foreign languages, and legal documents.

4.1 Error Analysis

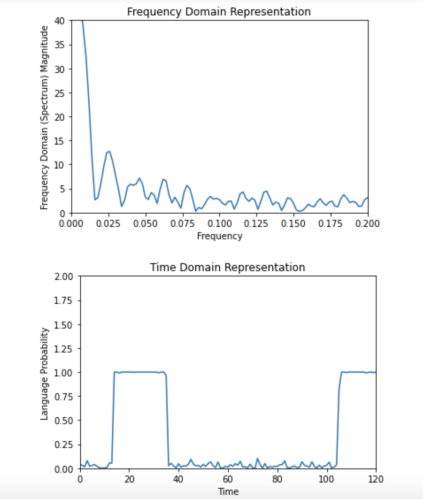

Figure 4: Time domain and frequency domain representations of the alternating language probability signal in a section of the monolingual book with page segmentation shown in Figure 4.

For example, Figure 4 shows a book that is entirely in Armeno-Turkish, but is segmented into four parts, with the lower two segments in a smaller font and occasionally in a different typeface, resulting in a significantly worse OCR output periodically. This creates a falsely identified language alternation pattern. Figure 5 shows the frequency and time domain representations of the Armeno-Turkish probability in the same book. In comparison, figure 6 shows the time-domain and frequency-domain signal representations of an actual bilingual book with an alternation pattern every sentence.

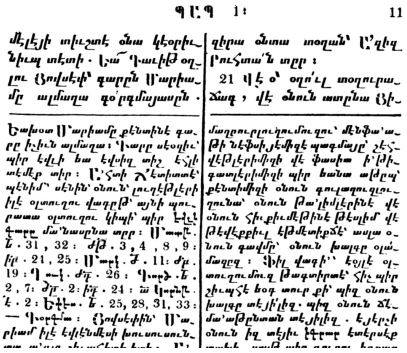

Figure 5: Page segmentation in the book, *Commentary On the Gospel of Matthew*, in Armeno-Turkish. (Goodell, 1851)

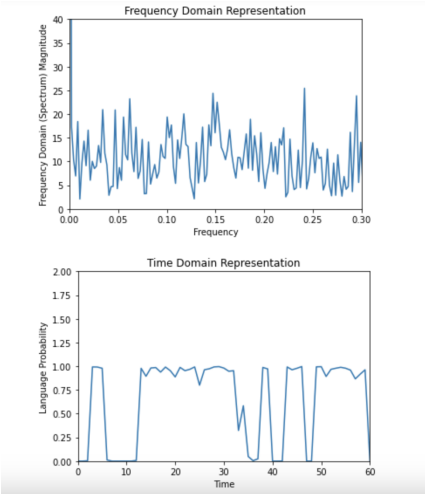

Figure 6: Time domain and frequency domain representations of the alternating language probability signal in a section of a bilingual book that alternates between Armenian and Armeno-Turkish every sentence. (Erg, 1881)

The patterns of language alternation that emerge are not very fine-grained, due to the high degree of noise in the HathiTrust corpus. In some cases, this leads to alternations in the lang ID probability

results, not due to a change in the language, but due to a periodic noise in the post-OCR text (footnotes in smaller font, highly segmented pages).

5 Future Work

We plan to expand the frequency analysis approach by using a cleaner dataset and including different languages with the goal of reaching a more nuanced classification at the word or sentence level (such as dictionaries). This process has a potential to be applied as a feature extractor for a downstream classification task. Training a classifier on clean data, where the patterns of structured language alternations are known, could lead to a specific classifier (dictionary, bilingual, annotated edition). This process would require clean data, but we hypothesize that a model trained on carefully annotated high-resource data could then be used on a low-resource language, since the periodic signal would be the same across languages.

References

1881. *Erger Tghayots' Hamar*. Konstandnupolis: Tpagrut'iwn Aramean.

1889. *Mejelle [Ottoman Civil Law.]*.

B. Anderson. 2006. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. ACLS Humanities E-Book. Verso.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Bedross Der Matossian. 2020. The development of armeno-turkish (hayatar t'rk'erēn) in the 19th century ottoman empire: Marking and crossing ethnoreligious boundaries. *Intellectual History of the Islamicate World*, 8(1):67–100.

William Goodell. 1851. *Madtéos Injilinin Tefsiri: [Commentary on the Gospel of Matthew]*. Smyrna: William Griffith.

HathiTrust Foundation. 2023. HathiTrust Digital Library.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic language identification in texts: A survey.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Laurent Kevers. 2022. CoSwID, a code switching identification method suitable for under-resourced languages. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 112–121, Marseille, France. European Language Resources Association.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

Lara McConnaughey, Jennifer Dai, and David Bamman. 2017. The labeled segmentation of printed books. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 737–747, Copenhagen, Denmark. Association for Computational Linguistics.

Paul McNamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *J. Comput. Sci. Coll.*, 20(3):94–101.

Jana-Katharina Mende, editor. 2023. *Hidden Multilingualism in 19th-Century European Literature*. De Gruyter, Berlin, Boston.

Hasmik Stepanyan. 2005. *(Armenian-Turkish-French) Bibliographie des livres et de la presse armeno-turque, 1727–1968*. Istanbul: Turkuaz Yayınları.

Sarah Werner. 2012. Where material book culture meets digital humanities. *Journal of Digital Humanities*, 1.