Compilation of a Synthetic Judeo-French Corpus

Iglika Nikolova-Stoupak **Gaël Lejeune** **Eva Schaeffer-Lacroix**
Sens Texte Informatique Histoire, Sorbonne Université, Paris, France
iglika.nikolova-stoupak@etu.sorbonne-universite.fr,
{gael.lejeune, eva.lacroix}@sorbonne-universite.fr

Abstract

Judeo-French is one of a number of rare languages used in speaking and writing by Jewish communities as confined to a particular temporal and geographical frame (in this case, 11th- to 14th-century France). The number of resources in the language is very limited and its involvement in the contemporary domain of Natural Language Processing (NLP) is practically non-existent. This work outlines the compilation of a synthetic Judeo-French corpus. For the purpose, a pipeline of transformations is applied to Old French text belonging to the same general time period, leading to the derivation of text that is reliable in terms of phonological, morphological and lexical characteristics as witnessed in Judeo-French. A tradeoff is sought between authenticity and efficiency as the ultimate goal is for this synthetic corpus to be used in standard NLP tasks, such as Neural Machine Translation (NMT), as an instance of data augmentation.

1 Introduction

When prompted to translate a text from Old French to Judeo-French, ChatGPT offers a slightly altered and, strangely, modernised version of the source text, also written in Latin script. Asked to identify the rare language based on a short sample, it convincingly defines it as "Hebrew".

1.1 The Judeo-French Language

Judeo-French was in use between the 11th and 14th centuries by Jewish communities in the northern regions of France. In fact, its similarity to the Old French language is at times so striking as for Banitt (1963) to famously define it as "a ghost language". Despite the difference of opinions on the topic, for purposes of clarity, Judeo-French will be referred to as a "language" rather than a "variety" within this work. The key distinguishing feature of Judeo-French is its rendition into Hebrew rather than Latin script. The three main types of Judeo-French sources existent today are: isolated glosses (including those by the renowned rabbi Rashi), Biblical glossaries, and several texts compiled entirely in Judeo-French (such as "Elegy of Troyes", a lament about thirteen Jews burned in Troyes in 1288). Similarly to Old French, Judeo-French involved a number of dialects and was not uniform throughout the centuries that marked its use. Also, although both languages are written in a highly phonetic manner, not all texts reflect perfectly ongoing processes of linguistic change; in other words, the languages are "phonetic in intention, if not always in performance" (Pope 1934).

1.2 Data Augmentation

One of the main challenges in NMT and other state-of-the-art language models is their application to low-resource languages i.e. languages that lack sufficient corpora to guarantee the optimal function of models. Different solutions have been proposed to overcome this limitation, including "transfer learning" from a "parent" language model to a "child" model in a related lower-resource language. In this case, the two languages share the same vector space and, by extension, benefit from the same data used in the training process (Dabre et al 2020). Another approach to dealing with scarcely-resourced languages is the practice of data augmentation or

Proceedings of LaTeCH-CLfL 2024, pages 40–45

the enlargement of the existing corpus via a variety of methods, such as backtranslation (loop translation from the target language back to the source language) or the addition of alternative subcorpora of lower quality and relevance to the task at hand. For example, in their abstract text summarization model, Parida and Motlicek (2019) use synthetic data derived from the noisy Common Crawl corpus. Dai et al (2023) benefit from ChatGPT's state-of-the-art text generation abilities as they rephrase sentences for consequent use in text classification.

Rule-based approaches to data augmentation were especially common before the advancement of neural models; for instance, in their work on a Machine Translation system that involves minority languages, Probst et al (2002) choose to rely on "a set of human-readable rules rather than a set of statistics" in the syntactic transfer between a low-resource and high-resource language. In the current age of neural networks and large language models (LLMs), the elaboration of rules mostly comes in the face of attempts to decipher the inner workings of "black box" language models; the emphasis being on economy of labelled training data and domain expert contribution (Mishra, 2022). Yet, the preservation of historical and culturally significant languages is an example of a goal that mandates explainability, expertise, and ready application in linguistic research and education. In his work *Anaphora Resolution*, Mitkov (2014) expresses optimism about the ongoing presence of rule-based approaches in universities and academia.

2 Pipeline

2.1 Selection of Source Text

The portal "Base de français medieval" is selected as the most suitable available source in Old French to be used for the extraction of text to be converted into synthetic Judeo-French. It contains 219 texts (about 6.5 million words), composed between the 9th and 15th centuries, along with metadata (see Table 1).

2.2 Preprocessing

Standard preprocessing is applied to a concatenated version of the texts, including the removal of capitalisaton and special symbols. The text is tokenised into sentences and the sentences are shuffled. A sample size is defined and extracted based on user input in function of the amount of augmented data that may be required by the NLP task at hand.

2.3 Transliteration

2.3.1 Into IPA Notation

As mentioned, the main difference setting apart Judeo-French text from Old French text is the script in which it is written. Therefore, the pipeline follows elaborate steps to guarantee the systematic transliteration of Latin to Hebrew letters. As an intermediary stage, the Old French text is converted into international IPA notation via the Python tool *epitran*. Specifically, the *fra-Latn-np* model for transliteration from French is applied, as it is highly based on the values of written letters as opposed to pronunciation as observable in the modern French language. To illustrate, the sentence "entre ses femmes appella cellui que elle avoit plus chiere" is rendered as "entrɛ sɛs fɛmɛs apɛla sɛlyi kɛ ɛlɛ avwat plys ʃirɛ".

2.3.2 Into Hebrew Script

The issuing text in IPA notation is then transliterated into Hebrew script on the basis of hand-crafted rules, derived from historical information about Judeo-French (see Figure 1).

When applicable, decisions are taken to reduce ambiguity (e.g. פ is taken to correspond to the sound *p*, although a value of *f* is also possible, in order to differentiate it from its *rafe* version, ֿפ). Where a symbol can be transliterated into multiple

Table 1: An overview of the source corpus.

	id	auteur	titre	siècle	dialecte	domaine
2	adgar	Adgar (dit Guillaume)	Collection de miracles	12	anglo-normand	religieux
3	AlexisProlRaM	anonyme	Prologue de la Vie de saint Alexis	09-11	normand	religieux
4	AlexisRaM	anonyme	Vie de saint Alexis	09-11	normand	religieux
5	aliscans1	anonyme	Aliscans	12	picard	littéraire
6	aliscans2	anonyme	Aliscans	12	picard	littéraire

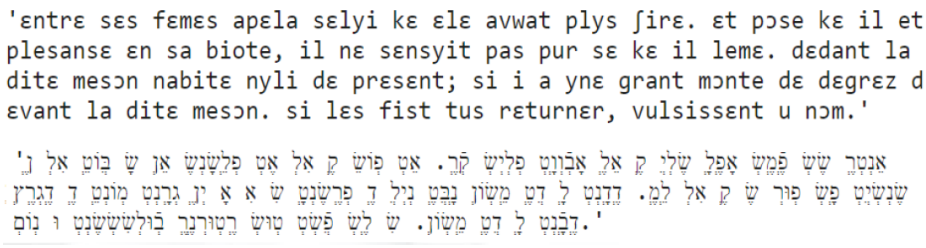

Figure 1: Transliteration from IPA notation into Hebrew letters.

Hebrew letters (e.g. *v* into ו, ב, בּ or ו), the most frequent mapping is used (in this case, ב). Given a larger sample, it is expected that it would be a better decision to also include the alternative renditions in a pre-defined proportion.

The automatised conversion pipeline includes the following steps: 1) vowels with IPA values that have the same Hebrew letter equivalents are made identical; 2) vowels are replaced with wildcards and consonants are replaced with their Hebrew equivalents while more wildcards are introduced for consonants that are interpreted as multiple symbols (e.g. those containing the *dagesh* diacritic); 3) where applicable, consonants are replaced with their *sofit* (end-of-word) versions; 4) initial vowels are replaced with א and the respective diacritic; 5) the *sheva* (vowel-less) diacritic is added to remaining consonants; 6) finally, remaining vowels are also replaced with א and the respective diacritic.

2.4 Simulation of Lexical Features

2.4.1 Lexical Borrowing

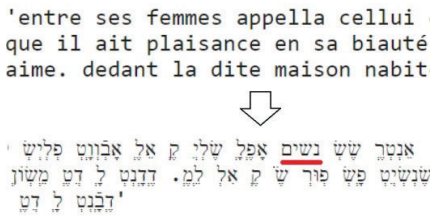

Figure 2: The French word “femmes” is replaced by the Hebrew word “נשים”

Another distinctive feature of the Judeo-French language is its occasional borrowing of Hebrew vocabulary. This phenomenon concerns particularly nouns and lexical fields associated with Jewish lifestyle and worship. Six out of the 80 nouns (i.e. 7.5 %) in “Elegy of Troyes” are such lexical borrowings: *torah* (Law), *tosafot* (additions, commentary), *hatan* (son-in-law), *sofer* (scribe), *cohen* (priest), and *qedushah* (holiness).

In order to mimic the phenomenon, Python’s *spacy* library is used to derive words’ part-of-speech tags. Then, a list of all nouns is produced and translated into Hebrew with the *googletrans* library. A set percentage of the produced Hebrew nouns are incorporated in the text in place of the original Old French nouns (see Figure 2).

Due to the scenario of data scarcity, it is recommended for informative features to be emphasised in the sample. For instance, in their recent article, Bansal and Sharma (2023) demonstrate the efficiency in selecting the most representative domain-specific data to annotate and consequently use in a language model, thus encouraging generalisation. For this reason, the percentage of instances of the feature is initially doubly increased in the synthetic sample (to 15%), with the ready possibility for modification based on performance of the sample in specific NLP tasks.

2.4.2 Words with Specific Spelling

Some commonly used words in Judeo-French tend to be spelled in a uniform way across dialects and time frames. A distinctive example is the word “God”, which is counter-intuitively spelled as גֹי (in contrast with the common Latin-based spellings “Dé” or “Dieu”), thus demonstrating sensitivity to current linguistic processes.

2.5 Simulation of Morpho-Syntactic Features

2.5.1 Interrogative Particle

Occasionally, Judeo-French texts use the word “si” as a question particle, calquing the Hebrew equivalent, הֲ. A ratio of the questions in the synthetic sample are set to follow this pattern.

2.5.2 Graphical Separation

Another discernible feature of Judeo-French is that definite articles, the conjunction “and” and several prepositions are typically connected to the word that follows, mimicking the behaviour of their Hebrew counterparts. The feature is reflected in the entire synthetic text.

2.5.3 Nominal Expressions

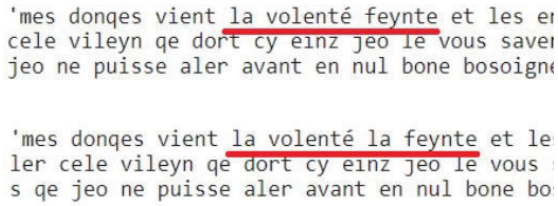

Figure 4: An example of a modified nominal expression.

Occasionally, Judeo-French nominal expressions follow the Hebrew structure of the definite article being repeated before both the noun and its attributive adjective (a necessity in the Hebrew language, as it does not feature the verb “to be” in the present tense, as a result of which it would otherwise be impossible to tell apart attributive from predicative adjectives).

In the compiled sample, combinations of consecutive *determinant* + *noun* + *adjective* and *determinant* + *adjective* + *noun* are sought and for a ratio of them, a second definite article is added accordingly (see Figure 3).

2.5.4 Plural Nouns

Commonly used nouns which are plural in Hebrew, such as “sky” and “water”, are usually pluralised in Judeo-French. These nouns, along with possible articles that precede them, are specifically sought in the source text in all of their common spellings as found in Old French during the examined time period (e.g. “water” could be spelled as “eue”, “eve” or “ewe”) and then pluralised.

2.5.5 Feminine Nouns

The unpronounced consonant ה- is often used in Judeo-French to mark feminine nouns, similarly to its role in the Hebrew language. In “Elegy of Troyes”, 3 out of 12 feminine nouns (25%) display the feature. A defined ratio (e.g. 50%) is made to comply to this rule in the assembled synthetic corpus. Firstly, a general assumption is made (and then verified manually) that nouns ending in the *e* or *ɛ* sound in the source text are feminine. A portion of these nouns are marked with a wildcard prior to transliteration into Hebrew letters, and it is eventually replaced by the letter ה.

2.6 Simulation of Scribal Errors

Scribal errors were a rather common occurrence in texts issuing from the discussed time period. Although their number varies significantly from text to text, they are all the more prominent when Hebrew script is involved due to the close resemblance of some letters. Consonants that were commonly confused include: ד and ז; ר and ץ; ב and ם; ן and ס. Coincidently, mistakes (a.k.a. noise) are often regarded as a positive addition in machine learning models, as they ensure that the system does not overgeneralise the text it encounters during training. 10% of occurrences of each of the involved letters are set to be erroneous in order to for the tendency to be emphasised (see Figure 4). This step takes place before the consonants and the vowels’ wildcards are replaced with Hebrew consonants carrying diacritics.

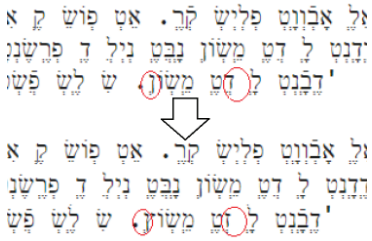

Figure 3: Simulation of scribal errors

3 Conclusion and Future Directions

The current study provides a basis for major extension in terms of both breadth and depth. On one hand, a number of rare Jewish languages share in their specificities as seen in relation to more common languages and language varieties spoken in the same time and geographical area. That is to say, Judeo-French relates to Old French in a similar manner that, for instance, Judeo-Italian relates to Italian or Judeo-Greek relates to Greek. Minor modifications in the presented pipeline can therefore allow for the derivation of synthetic corpora in these languages. From an even broader perspective, the authors hope that through its detailed documentation and shared

code, the study can encourage similar work with rare languages that are not related to the discussed one.

On the other hand, this work's artefact in the face of a sample for data augmentation is only the beginning of what can become a larger and more sophisticated NLP system, such as a Machine Translation or summarisation model, whose usability would in turn be exponentially larger as authentic documents in Judeo-French and related languages become translatable or otherwise more easily accessible in today's digital context.

Limitations

To underline the Judeo-French language's uniqueness and linguistic unpredictability, Kiwitt (2015) notes that "[c]ette transposition en graphie courante ne peut pas être mise en œuvre en appliquant des règles de substitution de manière mécanique" ("This common graph transposition cannot be implemented by applying substitution rules mechanically"). However, whilst synthetic Judeo-French text cannot reach the point of having authentic value, a simulation of the language's distinguishing characteristics can enable its active participation in contemporary NLP tasks.

The proposed pipeline can clearly benefit from the involvement of more elaborate NLP tools. For instance, topic modelling may be applied in order for nouns to be associated to relevant lexical fields before being replaced by Hebrew translations.

Although the described system's output corresponds to the authors' expectations and is subjectively judged as resembling Judeo-French text, its quality can be estimated best in the framework of its involvement in NLP tasks.

Ethics Statement

The synthetic Judeo-French corpus presented in this work has no claims of authenticity or full plausibility. Instead, it is meant to be used as a tool that would allow for the integration of authentic Judeo-French text into the framework of contemporary NLP tools, such as Machine Translation systems.

Acknowledgments

This work has benefitted immensely from the lectures provided by the Oxford School of Rare Jewish Languages (OSRJL) and, in particular, the Judeo-French lectures of Dr. Sandra Hajek.

References

Menahem Banitt. 1963. Une langue fantôme: le judéo-français. *Revue de linguistique romane*, 27:245–294.

Parikshit Bansal and Amit Sharma. 2023. Large Language Models as Annotators: Enhancing Generalization of NLP Models at Minimal Cost.

David Simon Blondheim. 1926. Poésies judéo-françaises. *Romania LII*, 17-36.

Raj Dabre, Chenghui Chu and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys*, 53(5).

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu and Xiang Li. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation, arXiv:2302.13007.

Stephen Dörr and Marc Kiwitt. 2016. Judeo-French. In Lily Kahn and Aaron D. Rubin (eds.): *Handbook of Jewish Languages*. Brill, Leiden: 138–177.

Kirsten A. Fudeman. 2008. Restoring a vernacular Jewish voice: The Old French Elegy of Troyes. *Jewish Studies Quarterly*, 15(3): 190-221.

Marc Kiwitt. 2015. L'ancien français en caractères hébreux. In David Trotter (ed.): *Manuel de la philologie de l'édition.* De Gruyter, Berlin: 219–236.

Ruslan Mitkov. 2014. *Anaphora Resolution*. Routledge.

Pradeepta Mishra. 2022, Model Explainability for Rule-Based Expert Systems. In *Practical Explainable AI Using Python*. Apress, Berkeley, CA: 315-326.

Shantipriya Parida and Petr Motlicek. 2019. Abstract Text Summarization: A low resource challenge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5994-5998.

Heinz Pfaum. 1933. Deux hymnes judéo-français du moyen âge. *Romania*, 59: 389-422.

Mildred K. Pope. 1935. From Latin to Modern French with a special consideration of Anglo-Norman:

Phonology and morphology. *Modern Language Review*, 30: 385.

Katharina Probst, Lori Levin, Erik Peterson, Alon Lavie and Jaime Carbonell. 2002. MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17: 245-270.