'Entrez!' she called: Evaluating Language Identification Tools in English Literary Texts

Erik Ketzan[1,*], Nicolas Werner[2]

[1]*Centre for Digital Humanities, Trinity College Dublin, Dublin, Ireland*

[3]*University of Cologne, Cologne, Germany*

Abstract

This short paper presents work in progress on the evaluation of current language identification (LI) tools for identifying foreign language n-grams in English-language literary texts, for instance, "'Entrez!' she called". We first manually annotated French and Spanish words appearing in 12,000-word text samples by F. Scott Fitzgerald and Ernest Hemingway using a TEI tag. We then split the tagged sample texts into four groups of n-grams, from unigram to tetragram, and compared the accuracy of five LI packages on correctly identifying the language of the tagged foreign-language snippets. We report that, of the packages tested, Fasttext proved most accurate for this task overall, but that methodological questions and future work remain.

Keywords

Language Identification, Computational Literary Studies

1. Introduction

This paper presents work in progress on the evaluation of current language identification (LI) tools for identifying foreign language n-grams in literary texts. The primary goal is to aid literary scholars who may wish to automatically identify, for example, bits of French in the novels of F. Scott Fitzgerald or Spanish in the novels of Ernest Hemingway. Ideally, the outcome of this project would be the creation of a single, easy-to-use Python script, by which a literary scholar could input any text, and the script would generate a table of n-grams with language candidates generated by the LI.

Automatic language identification, the task of determining the natural language that a document or part thereof is written in [7], has been extensively researched since the 1960's, and a number of advanced off-the-shelf LI packages and APIs are now available. As a richly developed NLP technology, LI systems have been the subject of large and robust evaluation methods and experiments [7], but we present this report on the precision of leading off-the-shelf LI packages for, specifically, the use case of computational literary studies, and even more specifically, texts which feature a primary language and small snippets of other languages.

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

*Corresponding author.

ketzane@tcd.ie (E. Ketzan)

0000-0003-0264-2157 (E. Ketzan)

Table 1
Selection of manually tagged Spanish in Hemingway’s *The Sun Also Rises*. Emphasis added.

He was anxious to know the English for <foreign xml:lang=”es”>**Corrida de toros**</foreign>
Drunk,” I said. ”<foreign xml:lang=”es”>**Borracho! Muy borracho!**</foreign>
They come to see the last day of the quaint little Spanish <foreign xml:lang=”es”>**fiesta**</foreign>.
At a thousand <foreign xml:lang=”es”>**duros**</foreign> apiece
You’re not an <foreign xml:lang=”es”>**aficionado**</foreign>?

Literary scholars, editors, and creators of scholarly editions may wish to have such an automatic language identification tool at hand to aid scholarly commentary on literary texts, to annotate in TEI and/or machine translate secondary languages in such texts, or to investigate code-switching (CS), which has been defined as “the ability on the part of bilinguals to alternate effortlessly between their two languages” [3]. While much literature on CS focuses on speech, “interest in written code-switching has developed more slowly” [6]. Studies on CS in literary texts often employ some form of digital tool for analysis, but identifying/tagging the second language (L2) is often performed manually [4, 1, 12, 11, 13, 15]. Anecdotally, we are aware that researchers prefer manual annotation for this step, as LI remains too unreliable. And our experimental results below support this impression.

The application of recent LI models to literary studies has so far been minimal. Pianzola et al. [14] report that in the classification of titles of texts (not the body of the texts) of fan fiction and classics in numerous languages, “CLD2 and CLD3 were not able to identify [the correct language of] 24.6% and 29.1% of the titles respectively”. As Pianzola et al. were dealing with text titles which presumably maintained language consistency throughout the title, while we look at 1- through 4-grams which can mix languages, we expect our results to be worse than the ~70-75 accuracy they report.

2. Corpora and Tagging

As this paper presents work in progress, our text sample is small while we solicit feedback on experiment design on a larger scale. As sample evaluation texts, we chose two well-known literary texts in English. A single annotator manually annotated 49 Spanish n-grams in a 12,000-word sample of Hemingway’s *The Sun Also Rises* (1926, published under the title *Fiesta* in London),[1] and 27 French n-grams in a 12,000-word sample of Fitzgerald’s *Tender is the Night* (1934)[2] using a single TEI P5 tag, `<foreign xml:lang="es">` or `<foreign xml:lang="fr">`, which identifies a word or phrase as belonging to a language other than that of the surrounding text.[3] Named entities were excluded from manual tagging, and a sample of manually tagged foreign language strings in these texts is presented in Tables 1 and 2.

One unresolved challenge of manually tagging foreign-language strings is: should foreign

[1] In total, this was 237 unigrams + bigrams + trigrams + tetragrams.
[2] In total, this was 281 unigrams + bigrams + trigrams + tetragrams.
[3] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-foreign.html

Table 2
Selection of manually tagged French in Fitzgerald's *Tender is the Night*. Emphasis added.

<foreign xml:lang="fr">"**Entrez**!"</foreign> she called, but there was no answer
Even his <foreign xml:lang="fr">**carte d'identité**</foreign> has been seen.
The famous Paul, the <foreign xml:lang="fr">**concessionaire**</foreign>, had not arrived
At four the <foreign xml:lang="fr">**chasseur**</foreign> approached him
he was joined by a <foreign xml:lang="fr">**gendarme**</foreign>

language words widely recognizable to native speakers, as a result of language borrowing and loan words [10, 5, 2] be tagged as foreign language? Some of the many examples of borrowed words or loan words in English include *restaurant*, *table*, *bonhomie*, and *kindergarten*, and there is the additional complication with historical texts that what is considered a loan word can change over time. While loan words present a methodological challenge for the technical evaluation of LI in literary texts, this challenge is perhaps less relevant for the desired outcome of this project: a Python script which presents a literary scholar with all or many candidates of foreign language words in a text, which the scholar could then sift according to their own research goals. Future work could also, of course, consider multiple annotators and inter-annotator agreement metrics in this step.

3. Experiment: Evaluation of LI packages on Hemingway and Fitzgerald samples

In our first experiment, we evaluated five current LI packages on the two manually tagged samples of Hemingway and Fitzgerald:

- **fasttext**: includes models for LI, developed by Facebook AI Research [9, 8].[4]
- **langdetect**: port of Nakatani Shuyo's language-detection library to Python, developed by Michal Mimino Danilak.[5]
- **langid**: standalone LI tool, developed by Marco Liu and Timothy Baldwin.[6]
- **polyglot**: depends on the pycld2 library which in turn depends on Google's Compact Language Detector v2 (CLD2) library for detecting languages in plain text. Developed by Rami Al-Rfou.[7]
- **pycld3**: Python bindings to Google's Compact Language Detector v3 (CLD3), a neural network model for language identification, support for over 100 languages, developed by Brad Solomon.[8]

[4] https://fasttext.cc
[5] https://pypi.org/project/langdetect/
[6] https://github.com/saffsd/langid.py
[7] https://polyglot.readthedocs.io/en/latest/Detection.html
[8] https://pypi.org/project/pycld3/

Our manual annotation of foreign language text served as the gold standard or evaluation criterium, and we wished to see how well the different LI packages performed in matching our manual tags. For literary texts with minor amounts of secondary language, such as the bits of Spanish in *The Sun Also Rises*, our method of query is to split the texts into n-gram strings and have the LI packages make a prediction for each string. We thus transformed our manually annotated foreign language words into 1- to 4-grams, where at least 50% of the word tokens are foreign language. For instance, Hemingway's line, "We wished him 'Mucha suerte,' shook hands, and went out", generated the bigrams: "him mucha", "mucha suerte", and "suerte shook". Again, as our explored method of automatically detecting Spanish n-grams would first require a split of the original text into n-grams. But what level of n-gram would produce the most accurate result? Here, we tested the packages by splitting the texts into four different lists of n-grams, as well as a combined result, measuring whether the correct prediction was made by any level of text split (1-, 2-, 3-, or 4-gram). The texts were split using the nltk.util.ngrams method from the nltk package. For those LI libraries which offer the option to return multiple language predictions with probability values, we only used the prediction with the highest probability for evaluation.

Evaluating the accuracy of the five LI packages on our sample, the LI package Fasttext performed overall best when considering 1- through 4-grams (Figures 1 and 2). The Langid package outperformed Fasttext on French 4-grams and Spanish 3-grams. An intuitive hypothesis would be that all LI packages would perform better, given a longer text sample. Our method, however, of performing the evaluation on n-grams that mix languages, complicates this process. By splitting the texts into n-grams that mix lexis from two languages, this no doubt "confused" the LI packages, which in turn resulted in longer text snippets (as long as 4-grams) not necessarily generating better LI accuracy. This also strikes at the core of the problem in this particular research: while LI presumably performs best with longer texts in a single language, foreign language words and phrases in English literary texts seem to most often be very short, down to the unigram level. It must be noted that at these small sample sizes, even a few correct or incorrect classifications can alter the results significantly.

With those caveats, what emerges from this small experiment is that Fasttext performs best on the task, with around 40% - 60% accuracy when 1- 4-grams are all considered together. Is a tool based on this method actually useful for literary scholars who wish to automatically pull out the French in Fitzgerald or Spanish in Hemingway? At this juncture, it is better than nothing, but certainly short of the goal.

4. Conclusion and Future Work

We suggest that LI holds great potential to assist a variety of computational literary studies. Some ideas for future work are discussed above, but certainly include larger evaluation samples, more languages, texts from more time periods, perhaps multiple annotators and inter-annotator agreement metrics, and a resolution of methodological questions such as the classification and possible exclusion of loanwords. Future work must devote significant attention to the different approaches of the packages to better understand and improve results. For instance, Fasttext is a (sub)word embedding library, while langid employs a Naive Bayes classification

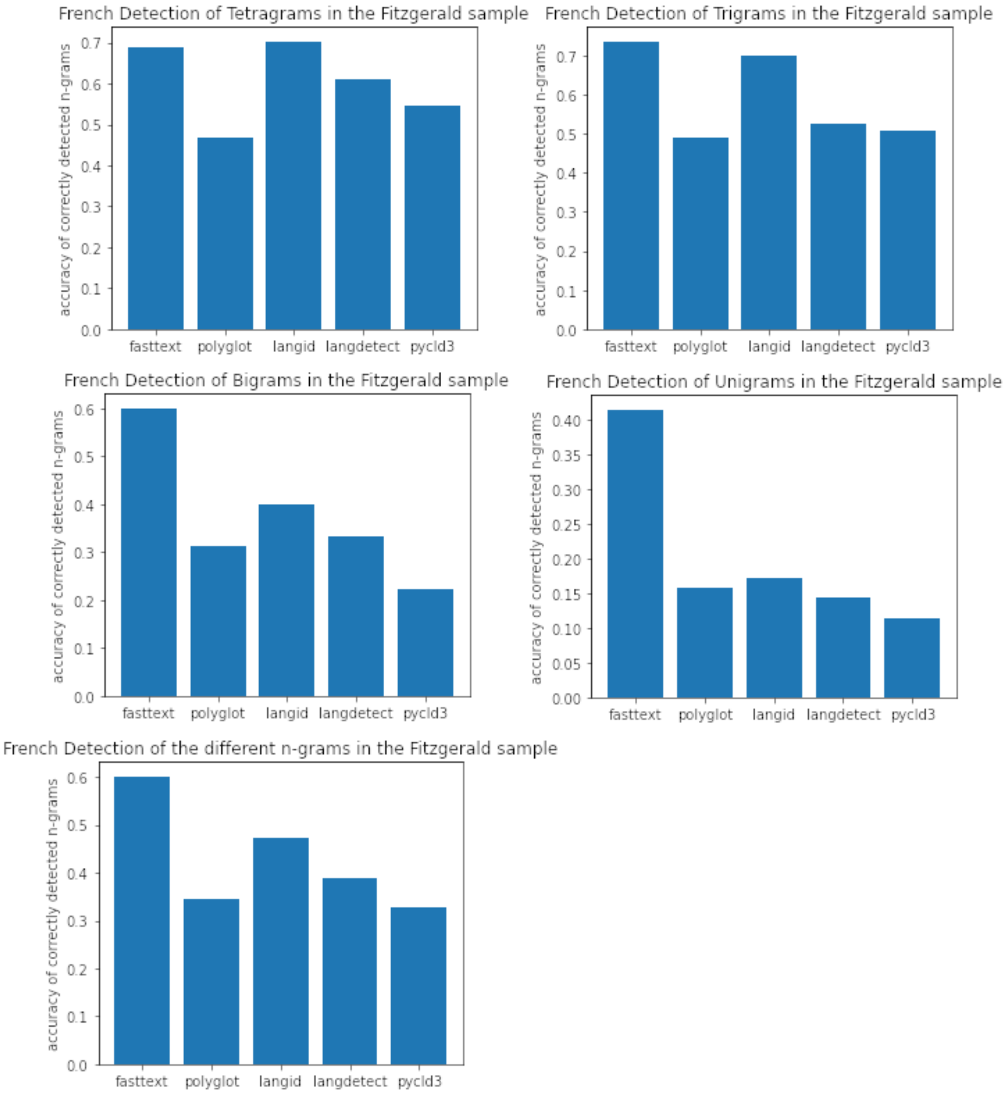

Figure 1: Accuracy of five LI packages in detecting French n-grams in a 12,000 word sample of Fitzgerald's *Tender is the Night*.

approach. Errors in the classification task should also be analyzed closely.

Another unexplored method is to simply restrict the languages that an LI can guess for each text. Our experiments allowed the LI packages to predict any language (leading to incorrect predictions from German to Filipino), but undoubtedly results would improve if, for instance, we restricted the LI to predict only English or Spanish in the Hemingway text. While this would improve results, it would require manual inspection of each literary text, which would become a greater burden for researchers, the larger the target corpus grows.

Different approaches other than LI should be attempted for this task. A dictionary-

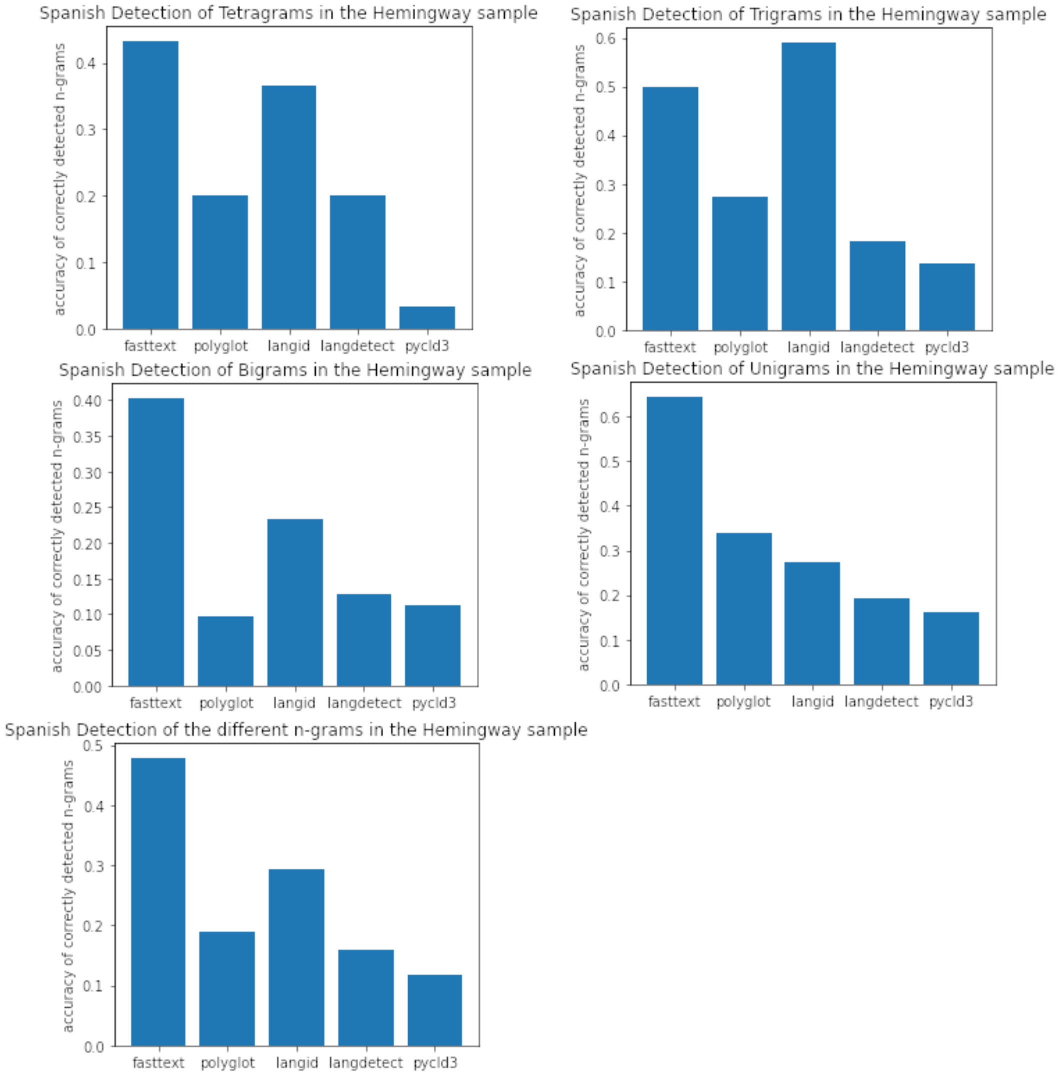

Figure 2: Accuracy of five LI packages in detecting Spanish n-grams in a 12,000 word sample of Hemingway's *The Sun Also Rises*.

based method could reveal possible foreign words. Certain part-of-speech taggers tag foreign/unknown words. A predictive language model may be able to identify foreign words candidates; for instance, given “he was joined by a”, the model could predict the next word, and if the actual word (here “gendarme”) is not in the top 50/100/500, then it may be a foreign word. A document-based language distribution may also provide candidates for foreign words; as a text introduces most of the words that are used throughout at the beginning of text (a Zipf distribution), if a new word appears later in the text, it could possibly be foreign.

It should be noted that there are already many practical solutions for readers and scholars in locating and understanding snippets of foreign languages in literary texts. Canonical novels are often published in popular or critical editions with translations in footnotes. Professional or fan-made websites may contain lists of translations of foreign language texts in popular novels. Ebook readers such as Amazon's Kindle contain machine translation of snippets as a built-in feature. Users may point their mobile phones at foreign language texts and obtain on-the-fly machine translation from such applications as Google Translate. Improved foreign language identification in literary texts could, however, supplement or improve upon some of these tasks.

A future goal of our research is to potentially unearth macro trends in foreign language text usage in literary history, for example the presumed decline of Latin or French in English-language literature. Such distant reading or cultural analytics of large literary corpora could perhaps reveal long-term trends in the rise and fall of foreign language usage over decades and centuries. And indeed, we have already performed such exploratory experiments with large corpora, but the low accuracy of LI packages in our experiments thus far and the tremendous amount of false positives in the results have not allowed meaningful results. But as LI packages improve at the n-gram level, and our methodological application of them to literary texts improves, such trends may become observable.

5. Acknowledgments

The authors wish to thank the peer reviewers for their extremely helpful comments.

References

[1] M. Albakry and P. H. Hancock. "Code switching in Ahdaf Soueif's The Map of Love". In: *Language and Literature: International Journal of Stylistics* 17.3 (2008), pp. 221–234. DOI: 10.1177/0963947008092502. URL: http://journals.sagepub.com/doi/10.1177/0963947008092502.

[2] K. Brown and J. Miller. *The Cambridge Dictionary of Linguistics.* 1st ed. Cambridge University Press, 2013. DOI: 10.1017/cbo9781139049412. URL: https://www.cambridge.org/core/product/identifier/9781139049412/type/book.

[3] B. E. Bullock and A. J. Toribio. "Themes in the study of code-switching". In: *The Cambridge Handbook of Linguistic Code-switching.* Ed. by B. E. Bullock and A. J. Toribio. 1st ed. Cambridge University Press, 2009, pp. 1–18. DOI: 10.1017/cbo9780511576331.002.

[4] L. Callahan. *Spanish/English Codeswitching in a Written Corpus.* Vol. 27. Studies in Bilingualism. Amsterdam: John Benjamins Publishing Company, 2004. DOI: 10.1075/sibil.27. URL: http://www.jbe-platform.com/content/books/9789027295378.

[5] D. Crystal and D. Crystal. *A dictionary of linguistics and phonetics.* 6th ed. The language library. Malden, MA ; Oxford: Blackwell Pub, 2008.

[6] P. Gardner-Chloros and D. Weston. “Code-switching and multilingualism in literature”. In: *Language and Literature: International Journal of Stylistics* 24.3 (2015), pp. 182–193. DOI: 10.1177/0963947015585065. URL: http://journals.sagepub.com/doi/10.1177/0963947015585065.

[7] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, and K. Lindén. *Automatic Language Identification in Texts: A Survey.* 2018. URL: http://arxiv.org/abs/1804.08186.

[8] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. *FastText.zip: Compressing text classification models.* 2016. URL: http://arxiv.org/abs/1612.03651.

[9] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. *Bag of Tricks for Efficient Text Classification.* 2016. URL: http://arxiv.org/abs/1607.01759.

[10] P. H. Matthews. *The concise Oxford dictionary of linguistics.* 2nd ed. Oxford paperback reference. Oxford [England] ; New York: Oxford University Press, 2007.

[11] C. Montes-Alcalá. “Code-switching in US Latino literature: The role of biculturalism”. In: *Language and Literature: International Journal of Stylistics* 24.3 (2015), pp. 264–281. DOI: 10.1177/0963947015585224. URL: http://journals.sagepub.com/doi/10.1177/0963947015585224.

[12] A. Mullen. “‘In both our languages’: Greek–Latin code-switching in Roman literature”. In: *Language and Literature: International Journal of Stylistics* 24.3 (2015), pp. 213–232. DOI: 10.1177/0963947015585244. URL: http://journals.sagepub.com/doi/10.1177/0963947015585244.

[13] K. B. Müller. “Code-switching in Italo-Brazilian literature from Rio Grande do Sul and São Paulo: A sociolinguistic analysis of the forms and functions of literary code-switching”. In: *Language and Literature: International Journal of Stylistics* 24.3 (2015), pp. 249–263. DOI: 10.1177/0963947015585228. URL: http://journals.sagepub.com/doi/10.1177/0963947015585228.

[14] F. Pianzola, S. Rebora, and G. Lauer. “Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers’ comments in the margins”. In: *Plos One* 15.1 (2020). Ed. by D. Orrego-Carmona, e0226708. DOI: 10.1371/journal.pone.0226708. URL: https://dx.plos.org/10.1371/journal.pone.0226708.

[15] M. J. Sánchez and E. Pérez-García. “Acculturation through Code-Switching Linguistic Analysis in Three Short-Stories: “Invierno”, “Nilda” and “The Pura Principle” (Díaz 2012)”. In: *Miscelánea: A Journal of English and American Studies* 61 (2020), pp. 59–79. DOI: 10.26754/ojs_misc/mj.20205139. URL: https://papiro.unizar.es/ojs/index.php/misc/article/view/5139.