The predictability of literary translation

Andrew Piper
McGill University
andrew.piper@mcgill.ca

Matt Erlin
Washington University
merlin@wustl.edu

Allie Blank
Washington University
adblank@wustl.edu

Douglas Knox
Washington University
dknox@wustl.edu

Stephen Pentecost
Washington University
spenteco@wustl.edu

Abstract

Research has shown that the practice of translation exhibits predictable linguistic cues that make translated texts distinguishable from original-language texts (a phenomenon known as "translationese"). In this paper, we test the extent to which literary translations are subject to the same effects and whether they also exhibit meaningful differences at the level of content. Research into the function of translations within national literary markets using smaller case studies has suggested that translations play a cultural role that is distinct from that of original-language literature, i.e. their differences reside not only at the level of translationese but at the level of content. Using a dataset consisting of original-language fiction in English and translations into English from 120 languages (N=21,302), we find that one of the principal functions of literary translation is to convey predictable geographic identities to local readers that nevertheless extend well beyond the foreignness of persons and places.

1 Introduction

Translation plays an important role in the international circulation of stories and ideas. Translations allow for the more widespread circulation of writing that would otherwise be hindered by global language differences. As such, translations can provide insights not only into the global commerce of ideas, but also the ways in which local regional cultures represent world cultures through their selection of works for translation. Research in corpus linguistics has consistently shown that the practice of translation is subject to producing predictable linguistic cues that distinguish translated texts from original-language texts regardless of the source or target languages (Baker, 1995; Volansky et al., 2015; De Sutter et al., 2017). From this perspective translation is understood as a particular "register" of language (called "translationese") governed by the cognitive demands of moving between languages (Liu and Afzaal, 2021; Mauranen, 2004; Xia, 2014).

At the same time, the field of literary translation studies has developed frameworks for understanding the concrete translational practices that arise in different national and historical settings. Relying mostly on smaller case studies, researchers have shown how particular cultural norms, political ideologies, and institutional contexts affect the nature and selection of literary translations within national literary markets (Reynolds, 2021; Heilbron and Sapiro, 2007; Heilbron, 1999). Heilbron (1999) and Sapiro (2010) have illustrated the asymmetry of target and source languages in international translation markets, i.e. the way translations are highly concentrated within a few core languages. Sapiro (2016) and Long (2021) have also shown how translations are often dominated by already highly reprinted canonical literature, where literary translation assumes a function of cultural consecration.

Our aim in this paper is to test the extent to which literary translations exhibit predictable traits similar to translationese but that reside at a deeper level of thematic content. Do translations function in a sense like a distinct literary genre, communicating a predictable set of themes that are otherwise less prevalent within original-language fiction? Understanding this aspect of translations' coherence will help us better understand the cultural functions that translations potentially serve. Our goal in doing so is to bring the affordances of NLP and machine learning into conversation with the work of cul-

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 155–160
November 20, 2022. ©2022 Association for Computational Linguistics

tural sociology and translation studies to further our understanding of the larger cultural function of translations in different literary contexts.

2 Data

For this paper, we follow the lead of Toury (1980) and create two equal-sized corpora of fictional texts, one consisting of works originally written in English and one of works translated into English from other languages. Our data is drawn from the NovelTM data-set of English-language fiction, which identifies 176,000 volumes of fiction located in the HathiTrust Digital Library published since the eighteenth century (Underwood et al., 2020). In order to identify a work as a translation, we use a set of regular expressions such as "translated from," "from the [language]," "tr. from," "rendered into English," etc. and match in volume metadata provided by Hathi to identify an initial list of candidates. If an author is included in this initial list, we then include all titles by that author.

In order to identify a volume as an original-language work, we use fuzzy matching against a large set of author names derived from Wikipedia and the Virtual International Authority File (VIAF), a database of author names from 69 library catalogues from around the world and their original language of publication.

We limit publication date between 1950-2008 for two reasons. The period after WWII is often considered a unique period in literary history, and thus these boundaries allow researchers to study translations as part of "post-war" literary culture. Additionally, we found that the diversity of source languages is almost exclusively European prior to this date, limiting the relevance of the data for studying questions concerning geographic space and language. Finally, we also remove all volumes where Underwood's predicted probability of being non-fiction was greater than 85%. Given that the set of original language works was larger than the set of translations, we then downsample each year of our original publications to match the number of translations.

In order to prepare texts for analysis, we concatenate the individual page files from each volume into a single document. We then represent each document as ten randomly selected 1,000-continuous-word samples drawn from the middle 60% of the document to avoid paratextual content in the front and backmatter. In order to avoid instances of low OCR quality, foreign-language passages, and samples that might have non-standard characters, only samples that have 90% of words in an English dictionary comprised of English-language fiction were kept. If a work did not have ten samples that met this criteria it was removed. After final review and cleaning we ended up with samples from 10,657 originals and 10,645 translations published since 1950. Our data contains 9,701 authors and translations from 120 unique languages. Fig. 1 provides the distribution of volumes by decade, while Fig. 2 provides the distribution of volumes by language region. As we can see the Hathi Trust collection is heavily biased towards translations from European languages.

To our knowledge, no existing collection of historically-matched translated and target-language fictional texts approaches the size or linguistic diversity of our corpus, and we hope that it will serve as a resource for additional research.

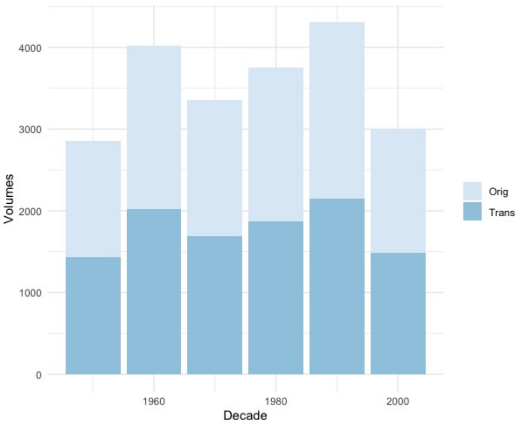

Figure 1: Distribution of translations and originals by decade

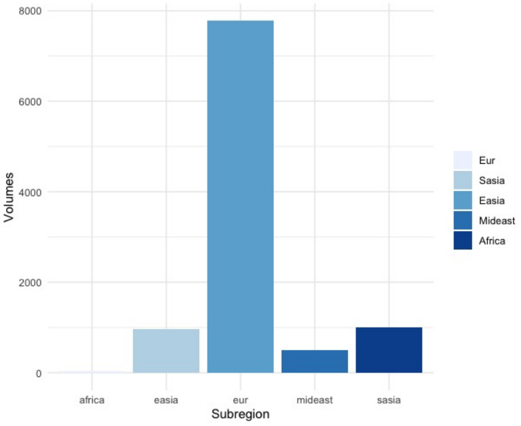

Figure 2: Distribution of volumes by language region

3 Methods

To measure the predictability of translations, we use a process of comparative supervised learning that harnesses different feature representations and different partitions of our data to better understand the conditions under which translations cohere as a distinct class of writing. We focus primarily on two scenarios, between translationese on the one hand and content-level qualities on the other. In each case, we rely on a linear SVM classifier, which has been shown to be robust for numerous text classification tasks (Colas and Brazdil, 2006).

To approximate translationese, we utilize a feature space composed of non-semantically valuable words, also known as "function words." As prior research has shown, numerous qualities of translationese – from the differential rate of pronouns (explicitation) to the recourse to shorter words (simplification) to the misalignment with expected word probabilities due to thinking in two languages simultaneously (interference) – are encoded in function words (Koppel and Ordan, 2011; Baroni and Bernardini, 2006). For our implementation, we use the 153 English stop-words included in the nltk library as our first feature space.

To capture content-level differences between translations and original-language works, we focus on constructing a feature space that is composed of semantically rich words that are not overtly culturally specific. As Volansky et al. (2015) have argued, translations are highly recognizable when compared to originals because they de facto contain foreign places and proper names. The classification task is thus often seen in this regard as a trival undertaking. Our goal is to assess the predictability of translation while masking these overt regional references.

To do so, we first construct a list of the most frequent (non-lemmatized) words that appear in our corpus not including stopwords. We then further manually remove proper names, locations, foreign words, and any obvious references to specific cultures or regions (such as "Madame" or "rupees"). After manual cleaning, we limit the number of words to a total of 2,000.

As an additional step towards masking the effects that individual and culturally-specific keywords may have, we further refine our feature space by adopting the procedure known as "authorless topic modeling" (Thompson and Mimno, 2018). Authorless topic modeling is appropriate for our purposes because it corrects for the tendency of LDA to generate overly source-specific results, especially topics that reflect key terms from a specific author or in this case language. By probabilistically subsampling words and eliminating those that are highly correlated with corpus metadata, this method helps reduce the association between particular topics and source texts, thereby producing more generalizable topics across the whole corpus. After experimentation we settle on a 30 topic model as the optimal representation. We provide samples of our topics in Table 1.

Because we are interested in assessing the extent to which content-level distinctions are potentially geographically dependent, we rerun the above two steps only on the translation data (i.e. we generate new lists of most frequent words, clean and re-apply authorless LDA). We then partition our translation data according to two different scenarios. The first is based on assumptions in the field of world literary studies that models the literary sphere into a European "centre" and non-European "periphery" (Casanova, 2004; Heilbron and Sapiro, 2007). The second subsets texts by each major geographic region as listed in Table 2.

Overall, this results in a total of five prediction tasks, four binary and one multiclass (see Table 2). The binary models allow us to compare the predictive accuracy of our two feature spaces (translationese v. content words) for translations and originals as well as our two larger global regions ("centre" and "periphery"). The multiclass model allows us to assess the regional predictability of translations across four major global areas according to topical distributions. For our binary models, we generate fifty models using a random sample of the data with replacement. For our multiclass model, we use ten-fold cross validation. We report mean F1, Precision, and Recall.

4 Results

We present our results in Table 2. While function words provide a strong level of accuracy, as expected, when predicting translations, surprisingly, our 30-feature LDA model outperforms the translationese model. Despite our efforts to create a set of general-language terms and topics, translations exhibit distinct topical behavior that is independent of proper names, places, or overt cultural references. Such topics also have predictive power for accurately identifying sub-regions according to our

Topic	Mean Coeff	Top Words
17	10.9	house time day village eyes felt face away mind people days started towards body home place water looked today rice
16	1.8	money old good day master hundred thousand pay make business buy shop house time told wife days year men head
0	1.2	woman wife mother husband house old girl daughter father home young child son women family children married sister years day
12	2.1	village old time work horse house good day home road away land long horses farm round men night fields yard
15	2.6	good great time replied day hand make found soon indeed moment order place friend house dear certain cried began long
20	2.9	young eyes old face round girl voice moment hand look white towards hair good smile head room table woman evening

Table 1: Most distinctive topics for non-European and European translations

data.

Translations are thus notably different at the level of content and not just in their reliance on low-level linguistic cues. Indeed, our models suggest that these content-level differences are meaningfully stronger than those indicated by translationese. When we break down our translations by sub-region, we also see that they exhibit very high levels of predictability (with the exception of our Middle Eastern texts though still well above chance). This suggests that translations from different regions are communicating thematically coherent and historically consistent information about those regions that extends beyond superficial markers of places or persons.

Corpus	Feature	F1	Prec	Recall
T/O	function	0.8235	0.8435	0.8267
T/O	LDA	0.8701	0.8707	0.8701
Eur/Non	function	0.7827	0.8163	0.7896
Eur/Non	LDA	0.8752	0.8763	0.8753
Europe	LDA	0.9572	0.9316	0.9844
Sasia	LDA	0.9128	0.9221	0.9048
Easia	LDA	0.7242	0.7893	0.6737
Mideast	LDA	0.3919	0.6566	0.2833

Table 2: Results of classification tasks

5 Discussion

Our paper provides the first ever attempt to use natural language processing to assess 1) whether literary translations exhibit categorically different behavior at the level of content when compared to original-language literature, and 2) whether these differences can be reliably mapped onto specific geographical regions while masking geographic information. We have found that while literary translations do indeed exhibit predictable qualities of translationese, they register even stronger stylistic differences at the level of content. Most notably, this holds even when explicit references to cultural contexts have been removed. Literary translation is distinctive as a class of writing because it talks about different kinds of experiences in different ways than original language literature.

This insight should motivate a good deal of future research into further understanding the particular nature of these differences. While prior work has suggested that literary translation plays a largely hierarchizing function – i.e. reproduces cultural hierarchies by conditioning on already highly reproduced (canonical) works – we find that literary translations are also distinctive because they introduce alternative subject matter into a target language that is geographically predictable even without overt geographical and cultural identifiers.

This suggests to us that one of translation's cultural functions is to encode geographic space, not simply through proper names or locations, but through a more extensive semantic field of references. Translations, in other words, make foreign spaces predictable and familiar to readers.

An exploration of our topic models suggests that translations may indeed be capitalizing on longstanding cultural associations with various geographic regions. One can see this on a superficial level in Table 1 by comparing topic 17 (distinctive

Region	Topic	Z-Score	Top Words
Sasia	17	1.75	house time day village eyes felt face away mind people days started towards body home place water looked today rice
Sasia	24	1.11	doctor read letter book day room years write time books work school name paper written writing wrote professor hospital reading
Easia	16	1.65	money old good day master hundred thousand pay make business buy shop house time told wife days year men head
Easia	6	1.36	right good maybe want time think make tell look old started things sure bit better kind else bad mean anyway
Mideast	28	1.3	god father priest church good people men holy world soul son great poor words heaven tell death devil prayer heart
Mideast	1	1.29	eyes black old world night body light life city white woman death sun people women earth sky time dead men
Africa	7	1.73	people work new party men country government young old women children workers war life meeting city office group political power
Africa	22	1.67	away saw water began day people dog told head old tree eat time men found night ran heard dead boy

Table 3: Most distinctive topics for each region

for non-European translations) with topic 12 (distinctive for European translations). Both of these focus on what we might term “village life,” but they include culturally specific elements: farms and horses versus rice. The very stereotypicality of these distinctions reveal how deeply culture is encoded into these texts. One a more interesting level, Topic 0 provides strong evidence of a focus on kinship relations in the non-European translations, possibly one that lines up with conventional narratives of asymmetrical global modernization (Dussel, 1993). Translations into English from non-European languages represent these worlds as shaped by more traditional, kin-driven social structures.

To further unpack the relationship between particular topics and translations from different regions, we use a Z-score calculation to determine which topics were more distinctive for individual regions, as shown in Table 3. The Z-scores are calculated by subtracting the mean of a topic’s average probability for all five regions combined from the score for a particular region and then dividing the difference by the standard deviation of the topic’s probability across all five regions. While additional research is necessary before any definitive conclusions can be drawn, the top words from the top two topics for each region provide a basis for some preliminary hypotheses.

Topic 17 turns out to be most distinctive for South Asia in particular, suggesting that works from this region that are translated often feature depictions of traditional village life. The East Asian topics prove difficult to parse without additional investigation but suggest a focus on morality (topic 6) and merchantry (topic 16). The relatively high representation of topic 28 in Middle Eastern texts indicates a predictable emphasis on religious matters. And finally, the “African” topic 7 suggests a concern with war and politics, possibly reflecting the postcolonial concerns of post-WWII African fiction. Topic 22 seems rather diffuse, but a glance at the texts in which it has strong representation reveal that it is associated with folk and fairy tales, which again suggests a stereotypical approach to translations from African languages.

Measuring the predictability of translation at the level content allows us to better understand the ways in which different regions and languages are represented in English. Studying translation at this level of scale can offer insights into how different regions consume and portray the world beyond their borders. While our work offers an initial insight into the function of translations into English, future work will want to compare these results with other regional and linguistic contexts. How do different regions represent world cultures differently when compared to each other? Our work offers a framework that can be applied to future parallel datasets to further understand the role that translation plays in shaping the global literary marketplace.

References

Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, 7(2):223–243.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Pascale Casanova. 2004. *The world republic of letters*. Harvard University Press.

Fabrice Colas and Pavel Brazdil. 2006. Comparison of svm and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 169–178. Springer.

Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere. 2017. *Empirical translation studies: New methodological and theoretical traditions*, volume 300. Walter de Gruyter GmbH & Co KG.

Enrique Dussel. 1993. Eurocentrism and modernity (introduction to the frankfurt lectures). *boundary 2*, 20(3):65–76.

Johan Heilbron. 1999. Towards a sociology of translation: Book translations as a cultural world-system. *European journal of social theory*, 2(4):429–444.

Johan Heilbron and Gisèle Sapiro. 2007. Outline for a sociology of translation. *Constructing a sociology of translation*, pages 93–107.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Kanglong Liu and Muhammad Afzaal. 2021. Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *Plos one*, 16(6):e0253454.

Hoyt Long. 2021. Chance encounters: World literature between the unexpected and the probable. *Journal of Cultural Analytics*, 6(3):25525.

Anna Mauranen. 2004. Corpora, universals and interference. *Translation universals: Do they exist*.

Matthew Reynolds. 2021. *Prismatic translation*. Legenda.

Gisèle Sapiro. 2010. Globalization and cultural diversity in the book market: The case of literary translations in the us and in france. *Poetics*, 38(4):419–439.

Gisèle Sapiro. 2016. How do literary works cross borders (or not)?: A sociological approach to world literature. *Journal of World Literature*, 1(1):81–96.

Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914.

Gideon Toury. 1980. *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University.

Ted Underwood, Patrick Kimutis, and Jessica Witte. 2020. Noveltm datasets for english-language fiction, 1700-2009. *Journal of Cultural Analytics*, 5(2):13147.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Yun Xia. 2014. *Normalization in translation: Corpus-based diachronic research into Twentieth-century English–Chinese fictional translation*. Cambridge Scholars Publishing.