Style in literary machine translation

Mehmet Şahin, Ena Hodzik, Sabri Gürses, Tunga Güngör,
Harun Dallı, Olgun Dursun, Zeynep Yirmibeşoğlu

BOĞAZİÇİ UNIVERSITY

Email: mehmet.sahin5@boun.edu.tr

This study is part of a scientific research project titled "Literary Machine Translation to Produce Translations that Reflect Translators' Style and Generate Retranslations" funded by TÜBİTAK (The Scientific and Technological Research Council of Türkiye (Grant No: 121K221).

Stylistic Border Crossings in and beyond Translation: Online Conference

This online conference is jointly hosted by the British Centre for Literary Translation and the East Centre.

9-10 March 2023

Outline

1. Introduction

In this study, we focus on style in literary machine translation and examine traces of a human translator's style in the outputs generated by a MT engine trained with translations of the human translator in question. We follow Saldanha's (2011) conceptualization of "translator style" as a consistent configuration of distinctive characteristics that are identifiable across multiple translations, and which exhibit a discernible impetus that is not explicable solely in terms of authorial style or linguistic limitations.

1.1. Style in corpus-based translation studies

The use of corpus tools has provided translation studies researchers with an insight into patterns of stylistic choices rather than an analysis of isolated examples of choices.

1.2. Style in literary machine translation

Style in literary machine translation has been approached only peripherally until the last few years.

There are few studies focusing on a particular translator or a particular genre in MT research.

There have been very few studies on the affordances of a MT engine for reproducing a human translator's style involving non-literary texts.

1.3. Style in (literary translation into) Turkish

The study of style is fairly recent in Turkish literature. The late westernisation of literature in the 19th century and the language reform in 20th century resulted in radical transformations in literary style and the conception of style. These create added complications for researchers.

More recently, style in translation has been a subject-matter of several studies on Turkish and there have been attempts to create digital corpora of Turkish literary language. Our study constitutes a novel attempt for a computational stylistic analysis of a literary translator and a recreation of her style.

1.4. Present study – Leech and Short's methodology

The purpose of this study was to conduct a quantitative and qualitative analysis of stylistic features in a complete corpus of English-Turkish translations by literary translator Nihal Yeğinobalı (1927-2008) who lived through the language reforms and translated from several genres. An existing methodology of stylistic analysis in English (Leech and Short 1981; 2007) was adapted and applied to our analysis of translator style.

Table 1. Leech and Short's (1981 In Olohan 2004: 147) checklist of style markers

Lexical categories	*Grammatical categories*	*Figures of speech*	*Context and cohesion*
General Nouns Adjectives Verbs Adverbs	Sentence types Sentence complexity Clause types Clause structure Noun phrases Verb phrases Other phrase types Word classes General	Grammatical and lexical schemes Phonological schemes Tropes	Cohesion Context

By employing quantitative and qualitative corpus analysis, the present study attempted to i) determine the distinctive measurable characteristics of translator style and ii) examine to what extent a customized MT system can reflect the same translator's style when the system is trained on previous translations by that translator. To address the first question, a corpus of English-Turkish translations by Nihal Yeğinobalı was compared with a reference corpus representative of Turkish linguistic trends, more generally. This resulted in a set of stylistic features characteristic of our translator. To address the second question, the corpus of Nihal Yeğinobalı was first used to train and test a set of machine translation models. Then, the stylistic features observed in the Nihal Yeğinobalı corpus were used to compare machine translation models. The same set of features was also used to identify translators in an authorship attribution analysis.

2. Method

2.1. Translator

The subject-translator Nihal Yeğinobalı was chosen because she had worked with different trends in of American literature from bestseller novels which were adapted into movies to world classics, Nobel prize winners, and Latin American authors. She was active as an editor and also, she authored several novels.

2.2. Materials

The present stylistic investigation is based on three corpora. The main corpus contains the entire body of works by Nihal Yeğinobalı. Throughout her career timeframe, Yeğinobalı produced a total of 129 works, including 123 translations, one pseudo-translation, and five original publications. The Yeğinobalı corpus has been digitalized with the informed consent of her heirs in compliance with the pertinent copyright laws. The investigation also incorporates a reference corpus, comprising 512 texts that are reflective of the linguistic trends observed in Turkish literary translations during Yeğinobalı's s active period from 1946 to 2015. The reference corpus served as a benchmark to validate the idiosyncrasies of stylistic traits identified in the Yeğinobalı corpus.

2.3. Data analysis

This study integrates two distinct yet interdependent methodologies: qualitative analysis via close-reading and quantitative analysis via distant-reading. Drawing upon Youdale's (2019) hybrid methodology, this study acknowledges the limitations of both approaches, and synthesizes close- and distant-reading techniques.

2.4. Authorship attribution

Authorship attribution is the study of detecting the author of a given text. In the context of our work, we use authorship attribution in the sense of determining the translator of a target text, based on the features applied in the data analysis step. We have incorporated computationally feasible features from Leech and Short's (1981) proposed methodology, added some traditional NLP metrics, and included morphological analysis, due to its particular importance in Turkish.

2.5. Machine translation

In this study, we fine-tuned a pre-trained OPUS Transformer model (Tiedemann and Thottingal, 2020) to capture the stylistic features of Yeğinobalı. This process amounts to adapting a general model to the works of the translator. 48 of the manually aligned books (referred as *Manual-large* corpus) have been used for training (training the MT system) and validation (adjusting hyperparameters of the system), and 3 manually aligned books (5,550 sentences) have been used for testing. We have also constructed six parallel corpora with varying sizes (50K, 100K, 150K, 200K, and 250K) to observe the effect of training set size on machine translation quality (measured by the BLEU score) and the success of capturing the translator's style.

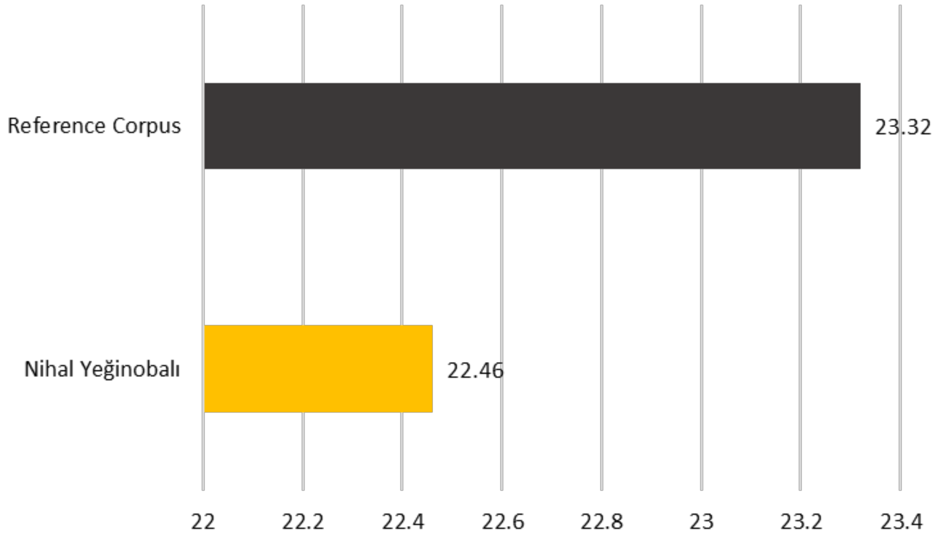
3. Results

3.1. Quantitative and qualitative analysis

Combining close- and distant-reading methods, we have identified a multitude of idiosyncratic lexical features that exhibit higher incidence rates than the reference values.

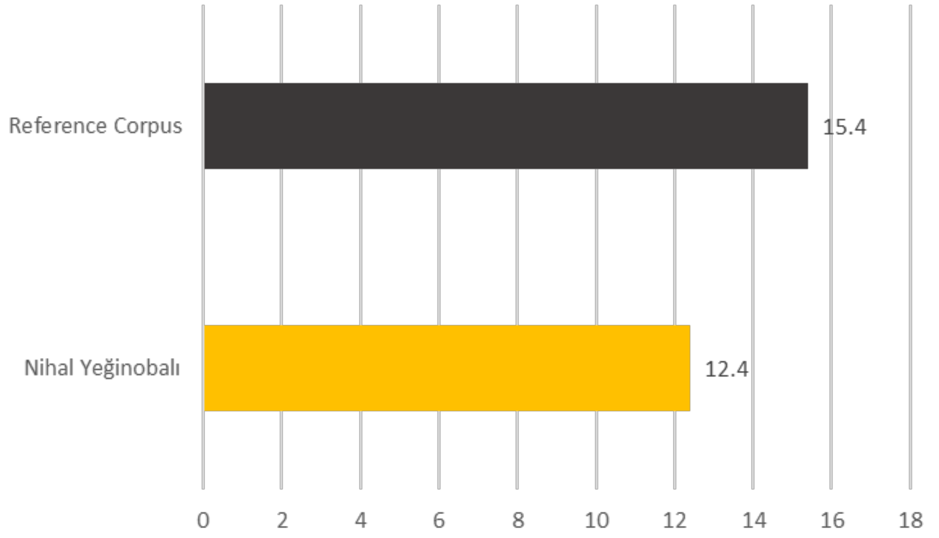
At the sentence level, the Yeğinobalı corpus features fewer morphemes per sentence than is typical of the reference values.

Figure 1. Average morphemes per sentence



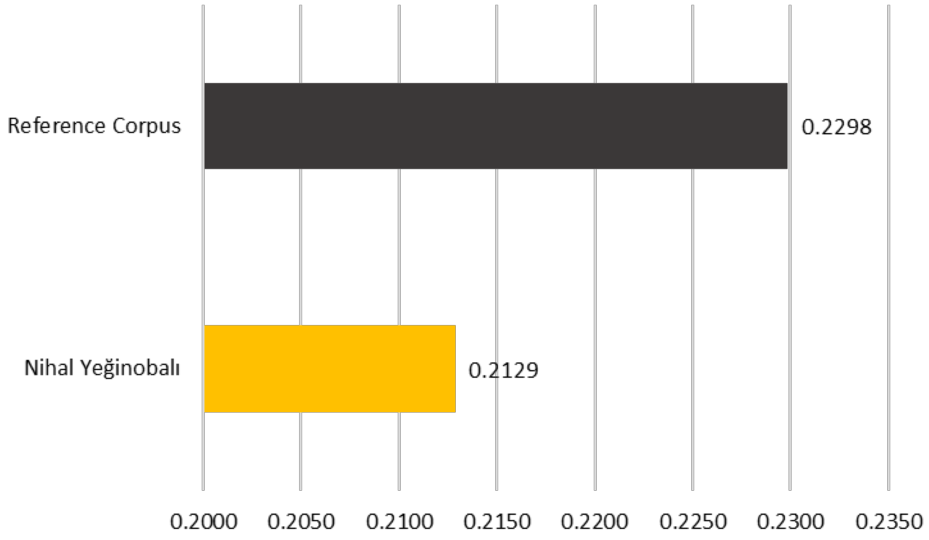
This is substantiated by the shorter sentence lengths observed in the Yeğinobalı corpus compared to the reference corpus:

Figure 2. Average words per sentence



Integrating the type-token ratio with average morphemes and words per sentence can enable a more holistic comprehension of the complexity of the lexicon within the Yeğinobalı corpus. While this graph may be difficult to interpret in isolation, it offers intriguing insights when considered alongside other variables.

Figure 3. Type-token ratio



Another recurring feature is the predominance of particular morpheme combinations within both the Yeğinobalı and reference clauses. In Turkish computational linguistics, morphemes are often represented in a generalized form when there are allomorphs, as is the case with the verbal adjective suffix ("-dik," "-dık," "-duk," "-dük," "-tik," "-tık," "-tuk," "-tük," "-diğ", "-dığ", "-duğ," etc.) that is commonly represented as "+DHk."

The 5 most common morpheme combinations in the reference corpus are as follows:

1. +(s)H+nA (e.g., ev+i+ne, kapı+sı+na)
2. +(s)H+nDA (e.g., ev+i+nde, kapı+sı+nda)
3. +(s)H+nH (e.g., ev+i+ni, kapı+sı+nı)
4. +Hyor+(y)DH (e.g., gel+iyor+du, yap+ıyor+du)
5. +DHk+(s)H (e.g., gel+diğ+i, yap+tığ+ı)

The Yeğinobalı corpus, on the other hand, features the following frequently occurring morpheme combinations:

1. +(s)H+nDA (e.g., ev+i+nde, kapı+sı+nda)
2. +(s)H+nA (e.g., ev+i+ne, kapı+sı+na)
3. +(s)H+nH (e.g., ev+i+ni, kapı+sı+nı)
4. +Hyor+(Y)DH (e.g., gel+iyor+du, yap+ıyor+du)
5. +lar+(s)H (e.g., ev+ler+i, kapı+lar+ı)

A notable observation is that Yeğinobalı's frequent use of the +(s)H+nDA combination contrasts with the established pattern +(s)H+nA within the reference corpus.

3.2. Translation Quality

The MT models fine-tuned on different portions of the Yeğinobalı corpus in the English-Turkish direction have been used to make predictions on the test set that contains 3 books and 5,500 sentences. The predictions were compared against Yeğinobalı's translation and evaluated with the BLEU score to measure the quality of the translation. It was observed that increasing the size of the training set almost consistently improves translation quality, and the best translation performance (9.04 BLEU score) was obtained from the *Manual-large* corpus that contains 48 books.

3.3. Authorship attribution

Our complete list of features are average morphemes per sentence, median morphemes per sentence, average morphemes per word, median morphemes per word, TTR (type-token ratio), number of unique words, number of unique words that occur at least 10 times, Mean word length, Standard deviation of word lengths, reduplications, ellipsis, questions, exclamations, mean sentence length, standard deviation of sentence lengths, median of sentence lengths, mode of sentence lengths, and normalized frequencies of the unigrams *gelgelelim, gelgeldim, maamafih, gene, ki, ve, pek, hem, derken, acaba, sahiden, doğallıkla*.

For each book in the reference and the translator corpora, we first create a book feature vector *bookv* and normalize each index for all books to fit the values between 0 and 1.

We then apply KNN (3 neighbors), SVC (linear kernel), Gaussian Naïve Bayes, Decision Tree, Random Forest, and Gradient Boosting (Alpaydın, 2020) classifiers to the dataset with 90 NY and 521 reference books. We obtain scores between 86.9% and 97.1% accuracy of determining the translator of a given book.

3.4. Stylistic analysis of MT output

Our machine translation outputs are evaluated through the authorship attribution methods, as they are successful at classifying whether a translation is made by a specific translator. We found out that, generally the outputs of our fine-tuned models are classified as belonging to Nihal Yeğinobalı, and the pre-trained model is classified as reference.

4. Discussion

The Yeğinobalı corpus demonstrates marked differences compared to the reference corpus. A solid stylistic indicator is the morpheme combination outlined in the preceding section, which serves to distinguish Yeğinobalı from other literary translators. Syntactic implications that arise from morpheme combinations can have a profound impact on the process of meaning-making. As an illustration, the following excerpts sourced from two separate Turkish translations of *Great Expectations* feature two distinct morpheme combinations. Specifically, the Yeğinobalı corpus exhibits the locative case “+DA” as in “evinde” (‘at her house’), while the reference corpus includes the dative case “+(y)A” as in “evine” (‘to/towards her house’):

1. Gene de, Miss Havisham'ın **evinde** oyun oynamaya ne diye gönderildiğim,
2. However, Miss Havisham's at her house game to play what for was sent
3. orada ne gibi bir oyun oynamak zorunda olduğum sorusuna hiçbir ışık
4. there what like a game to play I had to to the question no light
   tutmadılar. (Nihal Yeğinobalı)
5. they did not shed.

6. Sonra birer birer parlamaya başladılar, ama Bayan Havisham'ın **evine** neden
7. After one by one to shine they started but Miss Havisham's to her house why
8. gidip oyun oynamam gerektiği hususunda beni aydınlatamadılar. (A. E. İyidoğan)
9. go game I to play should in regards to to me they did not clarify

Original English version:
But they twinkled out one by one, without throwing any light on the questions why on earth I was going to play at Miss Havisham's, and what on earth I was expected to play at.
(Charles Dickens)

This differentiation substantially suggests that it is feasible to discern the stylistic traits of a translator based on morphological patterns. The Yeğinobalı corpus uses the "+(s)H+nDA" locative form to generate a sequence of phrases within a single sentence divided by commas. In contrast, the reference corpus maintains a uniform, more complex structure. The aforementioned example reveals another compelling attribute, whereby Yeğinobalı initiates a new sentence using the adverbial conjunction "gene de," thus corroborating the hypothesis that she favours brevity and simplicity in her sentence structures. As such, a heuristic exchange between the morphological patterns and lexical observations yields a nuanced understanding of the distinctive discourse produced by Yeğinobalı.

On top of these, our analyses of authorship attribution show that there is at least some degree of stylistic difference between Nihal Yeğinobalı and other translators in our reference corpus, and we are computationally capturing some of these through our features. Whether there are more features that help our classifiers get a better performance or help us pinpoint the exact occurrences within the natural flow of reading is an open question.

5. Conclusions

Our stylistic analysis reveals features on lexical and morphological levels that distinguish our translator corpus from a reference one. In addition, our authorship attribution analyses suggest that our observed stylistic features can be used to distinguish between MT models fine-tuned on our translator's stylistic features from a pre-trained MT model. Since our pre-trained model does not belong to the literary domain, it is difficult to say whether the observed differences between MT models are due to a particular translator's style or a characteristic of style in literary translation, more generally. To be able to draw more valid conclusions, our future work will include stylistic analysis of a second translator's corpus in order to determine stylistic features that can be attributed to a particular translator (but not to others). Moreover, a more-fine grained analysis of stylistic features based on time periods will be carried out to determine whether our observed features are particular to a translator or representative of other factors, such as the time periods in which the translations were produced.

6. References

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, ., Demirhan, U., Yilmazer, H., Atasoy, G., \"Oz, S., Yildiz, I., & others (2012). Construction of the Turkish National Corpus (TNC).. In *LREC* (pp. 3223–3227).

Alkım, B.E. (2002). Re-creating H.P. Lovecraft's Distinctive Style in Turkish [Unpublished master's thesis]. Boğaziçi University.

Alpaydın, E. (2020). Introduction to Machine Learning (4th ed.). The MIT Press.

Bada, V., Blanchon, H., Esperança-Rodier, E., & Hansen, D. (2022). La traduction littéraire automatique: Adapter la machine à la traduction humaine individualisée. *Journal of Data Mining & Digital Humanities*.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target. International Journal of Translation Studies*, *12*(2), 241-266.

Boase-Beier, J. (2011). *A critical introduction to translation studies*. Bloomsbury Publishing.

Bosseaux, C. (2001). A study of the translator's voice and style in the French translations of Virginia Woolf's The Waves. *CTIS occasional papers*, *1*, 55-75.

Caballero, C., Calvo, H., & Batyrshin, I. (2021). On explainable features for translatorship attribution: Unveiling the translator's style with causality. *IEEE Access*, *9*, 93195-93208. doi: 10.1109/ACCESS.2021.3093370.

Dinçel B.(2010) "From Motherless Brooklyn To Öksüz Brooklyn: Translating The Style Of Jonathan Lethem," İstanbul Üniversitesi Çeviribilim Dergisi 1 (1), 51-65.

Durmuş, H. E. (2021). Reframing an author's image through the style of translation: The case of Latife Tekin's Swords of Ice. *Babel*, *67*(6), 683-706.

Frankenberg-Garcia, A. (2022). Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart?. *Target*, *34*(2), 278-308.

Guerberof-Arenas, A., & Toral, A. (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, *9*(2), 255-282.

Kenny, D., & Winters, M. (2020). Machine translation, ethics and the literary translator's voice. *Translation Spaces*, *9*(1), 123-149.

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Leech, G. N., & Short, M. (1981). *Style in fiction: A linguistic introduction to English fictional prose* (No. 13). London, New York: Longman.

Leech, G. N., & Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose* (No. 13). Pearson Education.

Li, D., Zhang, C., & Liu, K. (2011). Translation style and ideology: A corpus-assisted analysis of two English translations of Hongloumeng. *Literary and linguistic computing*, *26*(2), 153-166.

Malmkjær, K. (2003). What happened to God and the angels: An exercise in translational stylistics. *Target. International Journal of Translation Studies*, *15*(1), 37-58.

Mastropierro, L. (2018). Key clusters as indicators of translator style. *Target*, *30*(2), 240-259.

Michel, P., & Neubig, G. (2018). Extreme Adaptation for Personalized Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers,* 312–318, Melbourne, Australia. Association for Computational Linguistics.

Mikhailov, M. (2021). Corpus-based analysis of Russian translations of Animal Farm by George Orwell. In *Corpus Exploration of Lexis and Discourse in Translation* (pp. 56-82). Routledge.

Munday, J. (2008). The Relations of Style and Ideology in Translation: A case study of Harriet de Onís. In *Actas del III Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación. La traducción del futuro: mediación lingüística y cultural en el siglo XXI. Barcelona* (pp. 22-24).

Olohan, M. (2004). *Introducing corpora in translation studies*. Routledge.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É., (2011). "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, 12, p. 2825-2830.

Şahin, M., & Gürses, S. (2019, August). Would MT kill creativity in literary retranslation?. In *Proceedings of the Qualities of Literary Machine Translation* (pp. 26-34).

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings* (pp. 417-427). Springer Berlin Heidelberg.

Saldanha, G. (2011). Translator style: Methodological considerations. *The Translator*, *17*(1), 25-50.

Saldanha, G. (2014). Style in, and of, Translation. *A companion to translation studies*, 95-106.

Tiedemann, J., & Thottingal, S. (2020, November). OPUS-MT--Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Toral, A., & Way, A. (2015). Machine-assisted translation of literary text: A case study. *Translation Spaces*, *4*(2), 240-267.

Toral, A., & Way, A. (2018). *What level of quality can neural machine translation attain on literary text?* (pp. 263-287). Springer International Publishing.

Vajn, D. (2009). *Two-dimensional theory of style in translations: an investigation into the style of literary translations* (Doctoral dissertation, University of Birmingham).

Wang, Y., Hoang, C., & Federico, M. (2021, June). Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1193-1199).

Youdale, R. (2019). *Using computers in the translation of literary style: challenges and opportunities*. Routledge.