Multilingual Stylometry: The Influence of Language on the Performance of Authorship Attribution using Corpora from the European Literary Text Collection (ELTeC)

Christof Schöch[1], Julia Dudar[1], Evgeniia Fileva[1] and Artjoms Šeļa[2,3]

[1]Trier Center for Digital Humanities, Trier University, Trier, Germany
[2]Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland
[3]Institute of Czech Literature, Czech Academy of Sciences, Prague, Czech Republic

Abstract

Stylometric authorship attribution is concerned with the task of assigning texts of unknown, pseudonymous or disputed authorship to their most likely author, often based on a comparison of the frequency of a selected set of features that represent the texts. The parameters of the analysis, such as feature selection and the choice of similarity measure or classification algorithm, have received significant attention in the past. Two additional key factors for the performance and reliability of stylometric methods, however, have so far received less attention, namely corpus composition and corpus language. As a first step, the aim of this study is to investigate the influence of language on the performance of stylometric authorship attribution. We address this question using four different corpora derived from the *European Literary Text Collection* (ELTeC). We use machine-translation to obtain each corpus in the other three languages. We find that, as expected, the attribution accuracy varies between language-based corpora, and that translated corpora, on average, display a lower attribution accuracy compared to their counterparts in the original language. Overall, our study contributes to a better understanding of stylometric methods of authorship attribution.

Keywords

Stylometry, Authorship Attribution, Multilingualism, ELTeC, Translation

1. Introduction

Stylometric authorship attribution is the task of assigning texts of unknown, pseudonymous or disputed authorship to their most likely author, often based on a comparison of the frequency of a selected set of features that represent texts [14, 26, 6, 23].

Traditionally, the way stylometric methods approach authorship attribution is to use the frequencies of a large number of simple features, such as words, lemmas or character sequences, for a determination of the degree of similarity between texts. These similarities, in turn, are in-

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

schoech@uni-trier.de (C. Schöch); dudar@uni-trier.de (J. Dudar); fileva@uni-trier.de (E. Fileva); sela@ucl.cas.cz (A. Šeļa)
https://christof-schoech.de/en (C. Schöch); https://artjomsh.github.io/web/ (A. Šeļa)
0000-0002-4557-2753 (C. Schöch); 0000-0001-5545-9562 (J. Dudar); 0009-0006-4555-7824 (E. Fileva); 0000-0002-2272-2077 (A. Šeļa)

terpreted as an indicator of the likelihood for two texts to have been written by the same author: the more similar the feature vectors, the more likely is identical authorship. The parameters of the analysis, such as feature selection and the choice of similarity measure or classification algorithm, have received significant attention in the past (e.g. [1, 25, 27, 11]). Two additional key factors for the performance and reliability of stylometric methods, however, are corpus composition and corpus language. They are relevant not only for the results in a specific case, but also for the overall performance and reliability of stylometric methods of authorship attribution.

Therefore, the aim of ongoing research by our group is to disentangle the influence of corpus composition and language on the performance of stylometric authorship attribution: To what extent do the attribution accuracy and robustness of such approaches depend on the language of the materials? To what extent do they depend on the corpus composition, which can present more or less challenging constellations of authorship and style? How do these two factors interact with each other, and how do they interact with feature selection?

This paper tackles one part of this issue, that of language, by investigating four distinct but broadly comparable corpora in a classification scenario. The corpora are all derived from the *European Literary Text Collection* (ELTeC) and contain at least three novels each by 8-10 different undisputed authors. These corpora can therefore be used as benchmarking datasets, where the true authorship is known for all novels under investigation. Specifically, corpora in English, French, Hungarian and Ukrainian are included. While the corpora contain similar texts (fictional narrative prose from a similar time period, i.e., 1840–1920), they differ both in terms of their exact composition, which can affect overall difficulty of attribution and therefore attribution performance, and in terms of their language, which can likewise affect attribution performance.

In order to investigate the role of language independently of corpus composition, all four corpora were automatically translated into the other three languages using the DeepL machine translation system (Pro version, used in the time period July to September 2023). This allows us to vary language while keeping corpus composition stable, and in this manner, tease out effects of both in terms of attribution performance (measured as classification accuracy).

In addition, and in order not to disadvantage any one corpus or language by selecting features that may not be suitable to them, a number of further parameters have been varied, namely: the type of feature considered (word forms, lemmas, part of speech or characters); the length of the feature sequence considered as the unit of analysis (unigrams, bigrams, 3-grams, 4-grams or 5-grams); the total number of different features considered (from 50 to 2000 in several increments); and the length of contiguous textual segments considered (from 5000 words to 10000 words as well as using entire novels).

Testing all of these various parameters, corpora and languages in all of their possible combinations results in a large number of individual results. In order to make them accessible to inspection and analysis, we have developed an interactive visualization that displays the attribution quality (i.e. the classification accuracy) in a heatmap as a function of the parameters described above. While the heatmap displays the accuracy for the entire range of features and segment lengths at the same time, other parameters can be selected by the users in order for the heatmap to be updated with the corresponding accuracy values. Two such heatmaps are provided next to each other to enable convenient comparison of results between any two configurations of parameters, whether corpora, languages or other settings.

The present paper can be understood as a background publication to these heatmaps published as a publicly available, interactive online showcase that is available at https://showcases.clsinfra.io/stylometry. The target audience of this showcase are scholars and students interested in a key methodological aspect of stylometry-based authorship attribution as well as those interested in cross-lingual approaches within the digital humanities. The showcase is intended both as a convenient way to document a specific research output and as a pedagogical tool that could be used in workshop or classroom settings alike. Readers are encouraged to use the showcase online, where much more detailed results can be explored, beyond the high-level summary of results we present here.

The detailed results are discussed below, and contribute to our general understanding of the influence of language and corpus composition on stylometric results. Broadly speaking, we can show that corpora of different language and composition lead to different attribution accuracy levels and different best-performing features, an expected result. We can also show that translated corpora (at least when all texts have been translated by the same machine translation system) usually lead to a lower attribution accuracy, overall, compared to their counterpart in the original language.

In the following, we first discuss relevant prior work evaluating stylometric performance in the context of multilingual corpora (section 2). We then outline our research questions and objectives (section 3), before describing our data and methods (section 4 and section 5). We discuss the main outcomes with respect to attribution performance, detailing corpus-based and language-based variations in results (section 6). By way of a conclusion, we discuss the consequences of our findings, document some limitations of this research and outline ongoing research into the second aspect of this investigation, namely corpus composition.

2. Relevant prior work

There has been a significant body of work investigating the methodological and theoretical foundations of stylometric authorship attribution. Generally speaking, it seems fair to say that most attention has been devoted to feature selection and the choice of distance measures or classification algorithms. While much research is focused on English-language materials, a limited amount of attention has been given to the question of variation across languages. Only rarely, however, has the influence of corpus composition on attribution performance been assessed beyond general statements about the advantages of generically-homogeneous corpora. In the following, we discuss some of this research in more detail.

With respect to investigations of feature selection, John Burrows [5, 4] proposes multiple measures each focused on a different part of the frequency spectrum: Delta, Iota and Zeta. Smith and Aldridge [25] use word forms in order to focus on the best setting for most frequent words and word vector onset and have found that “a word frequency vector of between 200 and 300 words gives the most accurate results” when using Burrows’ Delta. They also show that shifting the feature vector onset, effectively skipping the most frequent features, is detrimental to attribution accuracy.

In addition to varying the number of most frequent words, examples of using features other than words can also be found in the research literature. In particular, Stamatatos [27] focuses on the effectiveness of using character n-grams. He has conducted tests of 3-grams and frequent words on texts of different genres and topics. The researcher views as an advantage of n-grams that they capture linguistic patterns and may therefore be suitable to represent an author's style beyond word choice. He has obtained the best results with a "model based on character 3-grams in combination with an SVM classifier with linear kernel", and suggests using this as a baseline for authorship attribution. Building on the conventional use of word frequencies, Halteren et al. [13] investigate the application of syntactic features in stylometry. Their research highlights how syntactic patterns, based on word class tags, complement lexical features in finding what they call the "human stylome", i.e. individual measurable linguistic features that help distinguishing between authors. The study shows that although 100% efficiency results could not be achieved, lexical features perform slightly more efficiently than syntactic features and that in general, a greater number and/or variety of features improves the results. In a similar vein, Gorman [12] uses syntactic information for the attribution of short texts.

With respect to distance or similarity measures, Argamon [1] proposes a geometric interpretation of Burrows' Delta and derived several alternatives to this particular distance measure, among them Quadratic Delta. Smith and Aldridge [25] propose to move beyond distance-based comparisons and suggest the adoption of the Cosine similarity instead, where the directions of the vectors are compared without being influenced by their length. Evert et al. [11] follow this important new direction and propose Cosine Delta, which combines z-score normalization with the Cosine measure. They have confirmed the finding by Smith and Aldridge [25], namely that performance is more stable for longer feature vectors when Cosine Delta is used. In addition, they show that this effect holds across several languages, at least for long fiction narratives (novels). In several cases, interactions between various factors (in particular between feature vector length, distance measure and language) have been observed.

With respect to languages, many methodological investigations focus on English corpora only, as is the case with Burrows [5] or Smith and Aldridge [25]. Rybicki and Eder [19], however, show that results vary widely as a function of language when they investigated feature vector length and onset for multiple corpora in different languages and covering various literary genres. Evert et al. [11] test all their hypotheses on English, German and French corpora consisting of novels, with only minor differences between languages. However, the extent to which the differences in performance can be explained by language, on the one hand, and by the inevitable differences between corpora with respect to their genre and/or composition, is not within the scope of their research.

Stylometric research into translated works does exist, but is relatively rare. Rybicki and Heydel [22] show that in the case where several different translators translate the works of a given author, their stylistic signature can be detected by stylometric methods. In other cases, especially in corpora involving multiple authors and multiple translators, the authorial signal appears to be stronger than the translator's signal [21]. In a recent study, Rybicki [20] focuses on network analysis, using an extensive collection of translated texts from different genres and languages into Polish. He observes that the translated and original texts reveal stylistic similarities, from which it can be concluded that the translations are an extension of native literary traditions. Rybicki also notes that the distinctiveness of genres becomes visible during stylometric analysis.

There are also several relevant studies that use machine translated texts. This approach underlies Cross Language Authorship Attribution, a concept introduced by Bogdanova and Lazaridou [2]. Although the quality of automatically translated texts may distort an author's style, translation is one way of bringing texts written in different languages "into one space". Machine translation in this study is combined with a focus on lexical and higher-level features. The best results were achieved by combining machine translation and k-Nearest Neighbors. Based on this study, Mikros and Boumparis [17] apply the machine translation method for the language pair Greek-English. They conclude that when trained and tested on machine-translated texts, attribution accuracy is not strongly affected, compared to training and testing on original texts, but that the relevant features differ between the two languages, making cross-linguistic authorship attribution unreliable.

As mentioned above, the influence of corpus composition on attribution performance has so far not been investigated in any considerable detail. Smith and Aldridge [25] observe that a chronologically more diverse corpus leads to lower attribution accuracy but do not go into much detail. Kestemont et al. [15] show that cross-genre attribution is more challenging than within-genre attribution. A challenge for this kind of investigation is the availability of meaningful, high-quality metadata to assess corpus composition.

Based on this review of relevant research, we can conclude that we should expect some variation in attribution accuracy depending on language. Also, there is a clear need to investigate the role of corpus composition for the accuracy and robustness of stylometric authorship attribution.

3. Research questions and objectives

The overarching research question that ongoing work of our team addresses is to what extent, how, and under what conditions, both language and corpus influence the performance of stylometric methods of authorship attribution. The objective of our research, therefore, is to provide detailed data evaluating the accuracy and robustness of stylometric methods for corpora of different compositions and in different languages. As a first step, the present paper aims to determine how the results of stylometry are influenced by the language of the texts, when the corpus composition remains the same, that is, using corpora translated into multiple languages. How does accuracy vary across languages? How does it change when a given corpus is translated into a different language? What kind of interaction can we observe between feature selection and language, as well as between sample size?

In a second step, not covered by the present paper, we aim to determine how corpus composition affects attribution accuracy. In order to address this question, a formalization of corpus composition needs to be designed that quantifies the level of difficulty, for an attribution task, of a given corpus. How does such a measure need to be designed in order to adequately represent corpus difficulty in terms of authorship attribution? What types of metadata, for example regarding subgenre, time period or formal aspects, are most strongly correlated to differences in attribution accuracy? Does a higher corpus difficulty as measured based on metadata correlate with lower attribution accuracy? What is the effect of translation on this relationship? Our preliminary work on this issue is discussed in subsection 7.3 of this paper.

4. Data

The dataset was derived from the *European Literary Text Collection* (ELTeC; see [24, 3]) in order to build corpora in four languages: French, English, Hungarian, and Ukrainian. The choice of languages is determined by our desire to cover a number of different language groups: Romance, Germanic, Finno-Ugric, and Slavic languages. In this way, one can illustrate the effectiveness of the stylometric method and see how language affects the analysis results, for the four language groups.

The original ELTeC corpora each include one hundred novels. For our research, a selection of novels was made, retaining novels by 8 to 10 authors, each represented with three novels from ELTeC. This resulted in the following collections of texts: the English and the French corpora include each 30 novels, the Hungarian corpus 27 novels and the Ukrainian corpus consists of 24 novels (see key information on the corpora in Table 1).

Table 1
Key information on the four corpora used in this study (limited to the corpora in their original languages; we expect numbers of types and tokens to vary in the translations.)

	fra	eng	hun	ukr
number of authors	10	10	9	8
number of novels	30	30	27	24
total number of types	68,783	96,603	256,867	99,870
total number of tokens	2,414,226	4,791,903	2,635,112	824,130
median number of tokens	75,218.5	148,004.5	76,646.0	20,008.5
shortest novel, in tokens	29,460	23,459	14,590	9,517
longest novel, in tokens	149,810	348,793	285,818	102,859
earliest novel, year	1840	1844	1842	1841
latest novel, year	1900	1912	1908	1919

Each of the corpora was translated into the other three languages, thus the entire corpus includes 16 sub-corpora. The dataset also includes a metadata table for each corpus, with information about each of the novels. The metadata table includes information about the author, year of publication of digital and print editions, language, number of words in the novel, as well as subgenre (social, historical, adventure, detective or sentimental novel, bildungsroman or other) and narrative perspective of the novel (heterodiegetic, homodiegetic, epistolary, dialogue or mixed). This metadata was collected from experts in each of the relevant literary traditions who based their information on a reading of both the novels and of relevant secondary literature.

Data preparation involved two stages: translation of texts into the other languages and their linguistic annotation. The texts were translated automatically using DeepL Pro. For subsequent analysis, extracting lemmas and POS (part of speech) tags was necessary, a task we accomplished using the SpaCy library (version 3.7) for both original and translated texts. Additionally, unigrams and n-grams (from 2 to 5) were extracted using the stylo package [10].

5. Method(s)

Fundamentally, we used an authorship attribution classification task with leave-one-out cross-validation. Inspired by Rybicki and Eder [19], we used grid search over the features space to assess the general performance of authorship attribution and quantified the attribution performance in terms of its accuracy (mean attribution accuracy score for each condition as well as Cohen's Kappa for each condition). The goal was not to optimize for performance, but to cover the space of reasonable approaches that work for different languages and to understand the nature of differences in performance across languages and corpora.

5.1. Features

While in many scenarios, simple word form frequencies still prove effective, other kinds of features can in some settings be advantageous, for example for short texts or agglutinative languages [18]. Therefore, we did want to vary the feature types to some extent, in order not to miss conditions which work well in one or the other of the less-frequently investigated languages. There are three main levels of variation: types of features, sample size and feature vector length.

In terms of types of features, we used frequencies of word forms, lemmas, part-of-speech (POS) tags and characters. Each feature was cut to n-grams of different size: 1-3 for words and lemmata, 2-5 for character and POS n-grams (because the number of POS and character unigrams is too limited to be useful). Each n-gram length was tested independently.

With respect to sample size, frequency-based approaches to authorship attribution naturally depend on the available size of the text. There is a considerable variation in text sizes in ELTeC (as shown in Table 1); to mitigate this, we draw a random sample of consecutive tokens (a "chunk") for each text based on the shortest text across all corpora (10,000 words). We use test sizes of 5,000 to 10,000 tokens for word-based features, and 10,000 to 50,000 for character-based ones. The number of available tokens per text differs dramatically between words and characters; so sample sizes mean different things for n-grams based on these features. To account for the variability that is introduced with taking only one limited sample out of sometimes very large novels, at each step we take a random consecutive sample out of all available 'chunks' and record performance 100 times. As the last step, we perform classification on full-length texts.

In terms of feature vector length, vectors that represent texts are constructed based on most frequent feature length cutoffs. We used 50 to 2000 features with incrementally increasing step sizes.

5.2. Classification task

For author-based text classification, we use the Support Vector Machine classifier and perform leave-one-out cross-validation. This means that in each step, one text is removed from the data, the model is then trained on the remaining texts, and then the authorship of the left-out text is predicted and the result is recorded. This process continues until all texts have been left out once. For each language combination, we run 100 iterations of leave-one-out cross validation classification, taking a sample from each text (chunks of consecutive sequences of size n) at random. Additionally, we run a single full-text analysis for each corpus.

We report two performance measures: simple accuracy (proportion of correctly predicted authors) and Cohen's kappa (typically used for measuring inter-annotator agreement [7]). The latter measure, while highly correlated to accuracy, allows us to partly offset the different amount of classes across corpora, since it provides an inter-rater agreement score adjusted for random classification.

5.3. Visualization

To create the interactive visualizations of the showcase, we used Bokeh, a Python library offering extensive capabilities in crafting interactive and dynamic data visualizations. The visualization – which is available online at https://showcases.clsinfra.io/stylometry – consists of two heatmaps, enabling users to compare and contrast two sets of results simultaneously. In this article, we focus on the results for the Hungarian corpus, which are presented in subsection 6.2. For results from other corpora, please see section 8 (appendix) and/or visit the online showcase.

The x-axis represents various settings of the most frequent features (MFF) used in a specific analysis, while the y-axis denotes distinct sample sizes, ranging from shorter text snippets to entire novels ("full novel"). Words and characters are treated differently and analyzed across diverse sample sizes (see subsection 5.1 above). Each heatmap cell correlates MFF and sample size, with the color intensity indicating accuracy levels, offering a visual approach to the accuracy obtained in the analysis. A mouseover provides further information, such as the features used as well as numerical indications of accuracy and Cohen's Kappa. When there is no data corresponding to selected feature combinations, the plots display a "No data available for these selectors" message. The range of MFF values depends on the features; in some cases, as for character and POS bigrams, the theoretical maximum of features is limited.

Users can engage with the data through several selectors, enhancing the exploration and analysis process. The available choices include:

1. Corpus selector: Enables the selection from a range of available corpora, facilitating comparative studies;
2. Feature level selector: Permits users to toggle between "words" and "chars" (characters) feature types;
3. Feature type selector: Allows for the choice between plain text, lemmas, or POS tags;
4. Ngram size selector: Offers options to choose n-gram sizes from 1 to 5, allowing for detailed linguistic patterns examination.

A color scale on the right side of each graph delineates the accuracy levels, ranging from 0 to 1. Here, cooler tones like greens, blues and purples signify lower values, whereas warmer hues, like red and orange, indicate higher values. This color coding serves as a primary indicator of stylometric effectiveness across various parameter settings, vividly illustrating the success rate of stylometric analysis as a function of the chosen corpus and the selected criteria.

6. Results

In relation to the main outcomes, we aim to present a concise overview of some of our key findings. First, we focus on evaluating the overall performance of authorship attribution, based on mean accuracy scores across multiple settings, for all corpora. Second, we provide a detailed description of the main outcomes for the Hungarian corpus and its translations. We focus on one language for reasons of space, but readers can find a detailed description of the performance of the other three languages in the appendix (section 8) to this paper.

6.1. Attribution performance across all corpora

Regarding overall performance, we discovered several parameters that consistently yield the best results across all corpora and translations. At the word level, the following parameters showed the highest scores: unigram for both feature types (plain word forms and lemmas), with only minimal differences. Meanwhile, at the character level, very high performance was observed using 5-grams based on both plain word forms and lemmas as feature types and using full novels as sample size.

In Figure 1 and Figure 2, the average classification accuracy at the word level is illustrated, employing unigram plain word forms, with sample sizes ranging from 7,000 to 9,000 and from 10,000 to full novel size, respectively. As observed in the figures, the clear winner among the original texts is the Hungarian corpus. Translations from Hungarian into three other languages also demonstrate very high performance, slightly decreasing to 0.6 and 0.7 from 0.86 compared to original texts when full texts and sample size of 10000 are taken into account. Concerning smaller sample sizes, the decrease is more noticeable, achieving an accuracy of 0.4-0.5, while the accuracy of original texts is 0.8. The second-best performance comes from the original Ukrainian corpus; however, translations from Ukrainian have low performance, falling to 0.5 and 0.4 from the original scores of 0.76 and 0.65 for large and smaller sample sizes respectively, indicating that translations from Ukrainian are particularly challenging for attribution tasks. The best performance among translations from Ukrainian, surprisingly, is shown by the Hungarian language.

The French and English corpora have almost similar performance when the original texts are analyzed. However, translations from English at the word level perform lower compared to the original texts, and the difference in performance is more pronounced with smaller sample sizes, falling from 0.5 to 0.3-0.4. On the other hand, at the character level, translations into Hungarian outperform original texts with the following settings: 5-gram based on both lemmas and plain word forms.

Remarkably, the smallest difference in performance between translated texts and the original is observed in the French corpus, where the difference is smaller than 0.1 for large sample sizes and only slightly greater than 0.1 for smaller sample sizes. It appears that the authors' style is best preserved in translations from French. All three translations from French demonstrate similar performance; however, translations into English slightly outperform the other translations.

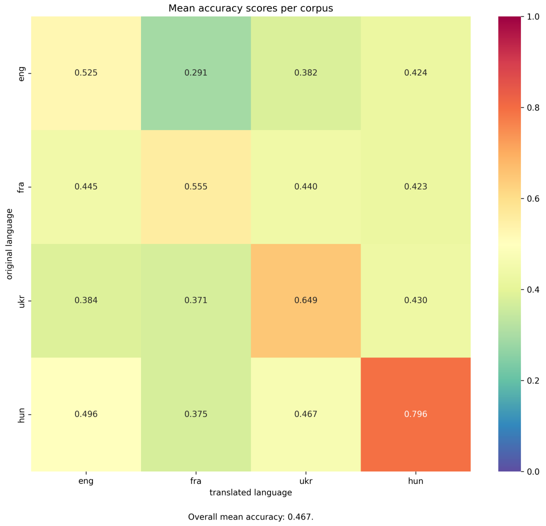


Figure 1: Mean accuracy across each corpus with the following parameters: word forms, word level, ngram = 1, sample size = 7000, 8000, 9000. Number of most frequent features: 100–1000.

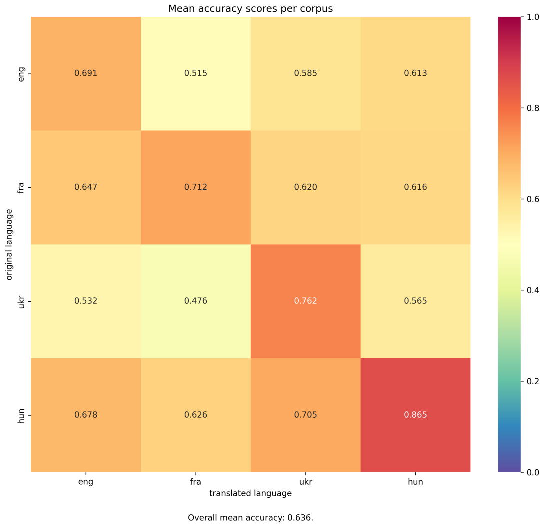


Figure 2: Mean accuracy across each corpus with the following parameters: word forms, word level, ngram = 1, sample size = 10000 and full novels. Number of most frequent features: 100–1000.

6.2. Hungarian corpus attribution performance

As already discussed earlier, the Hungarian corpus (Figure 3) demonstrates the highest accuracy scores overall, reaching approximately 80–95% across many sample sizes; starting with samples of 8000 words, results are reliably very accurate. This achievement is realized by employing the following settings: the number (N) of the most frequent words ranges from 50 to 700, with the analysis based on unigram plain word forms. Although slightly less effective, classification based on unigram lemmas still produces high scores. Regarding the analysis using character n-grams as features, the performance remains remarkably high, fluctuating between 70-93% for sample sizes starting from 4000, with the most remarkable performance observed when 5-gram character sequences are used with full novels (Figure 4). In this scenario, the attribution based on characters, derived from word forms with the number of MFF from 50 to 500, maintains a constant accuracy rate of over 90% for full novels.

When evaluating the accuracy of translations from Hungarian into other languages, it becomes evident that the performance is noticeably lower compared to original novels. One of the few configurations displaying high accuracy (over 90%) is observed in translations into Ukrainian utilizing 5-gram characters based on plain word forms. For lemmas, the Ukrainian translation, in some cases, outperforms the original texts (Figure 5). As for the attribution at the word level, high accuracy is observable only for full novel sample sizes. Translations into English and Ukrainian exhibit very similar performance, whereas French displays lower accuracy.

7. Conclusion

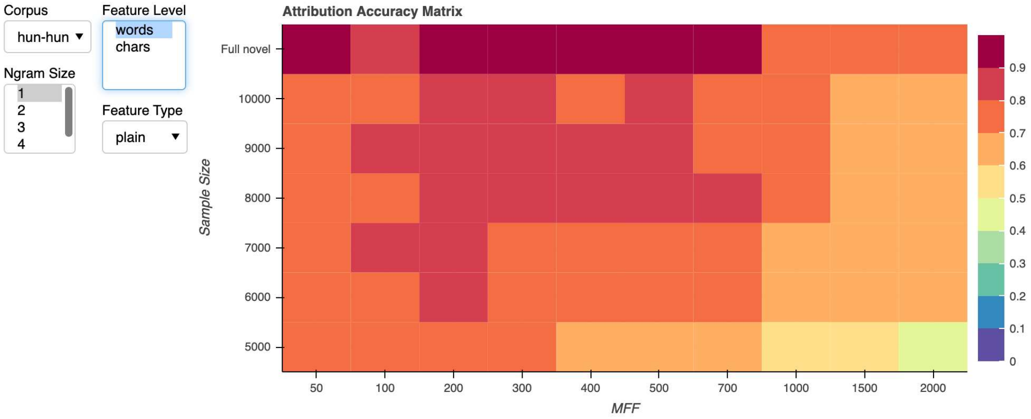
In conclusion, we would like to summarize some of the key findings of this study, but also discuss a number of limitations and next steps for research in this area.

7.1. Findings

We can discern consistent patterns across all languages and translations when analyzing their performance. These trends remain mostly consistent whether we are examining analyses based on words, lemmas or character n-grams.

Optimal performance consistently emerges when the analysis is based on entire novels rather than shorter samples. When using samples, results tend to improve as the size of the analyzed samples increases. This finding is consistent with previous studies [4, 8, 9].

Conversely, increasing the number of features used for attribution beyond a certain value tends to lead to a deterioration in classification scores in most cases. The decline in performance becomes noticeable at full novels and word unigrams beyond roughly 300-800 most frequent words for French, English, and Hungarian (but this is not the case for Ukrainian). Furthermore, the impact is considerably less pronounced for word bigrams and trigrams. This finding is consistent with results from distance-based attribution studies when using Burrows' Delta, but not when using Cosine or Cosine Delta [25, 11]. This phenomenon may be caused either by an increasing proportion of features driven by topic rather than authorship, or by an increased amount of noise in the signal, when using higher numbers of features.



Corpus: hun-hun
Feature Level: words
Ngram Size: 1
Feature Type: plain

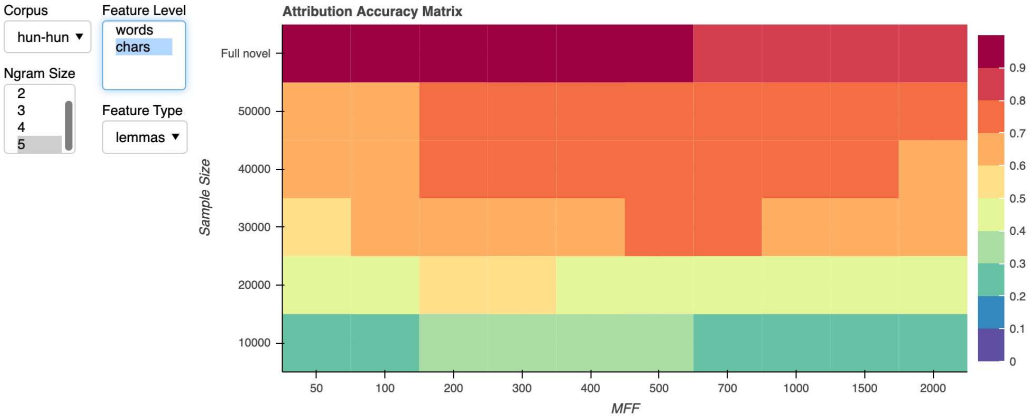
Figure 3: Mean accuracy scores for the Hungarian corpus in the original language, for different settings of MFF and sample size. Parameters: Plain word form unigrams.

Performance for corpora is always better when they are analyzed in their original language, rather than in a translation into another language. However, this might be due to the fact that we have used a unique translator, as it were, this fact being a major limitation of our study (see below).

7.2. Limitations

The finding, mentioned above, that the translated corpora always perform less well, in our experiments, than the original corpora, may very well be an artefact of our method of translating these texts. Given that we used DeepL for all translations, the translations could be said to have all been produced by the same 'agent', leading to a certain homogenization of their style and lexicon, which we would expect to be a challenge for stylometric authorship attribution. In contrast, translations of literary texts are usually performed by a range of different translators, some of which maintain privileged relationships with certain authors, in a constellation where authorial and translatorial signal may reinforce each other. At the same time, the question of the (in)visibility of the translators in stylometry has been addressed several times [21, 16].

Another limitation with respect to the corpora is that for each language, we have only one particular corpus available. This means that, given how important the genre of a corpus as well as the corpus composition are likely to be, we cannot generalize our results based on these particular corpora to the languages in general.



Corpus
hun-hun ▾

Feature Level
words
chars

Ngram Size
2
3
4
5

Feature Type
lemmas ▾

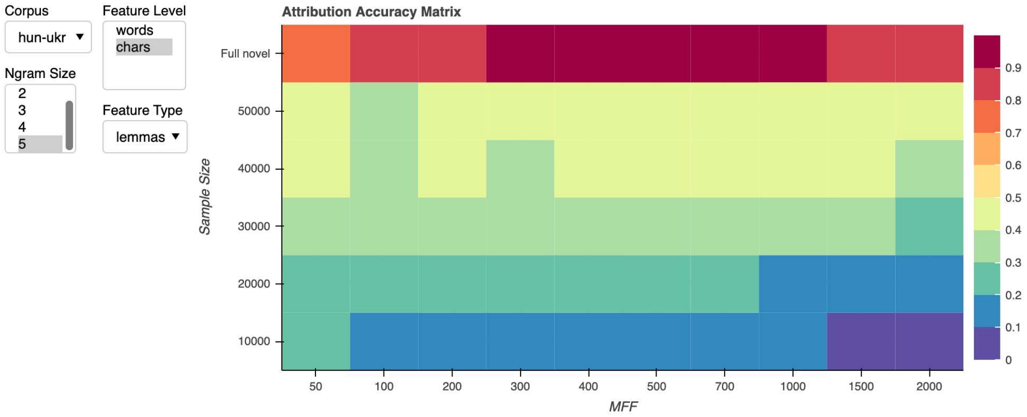
Figure 4: Mean accuracy scores for the Hungarian corpus in the original language, for different settings of MFF and sample size. Parameters: Character 5-grams of lemmas.

7.3. Next steps

As mentioned above, the research presented here only covers one of two aspects that we consider relevant to our overall research question. Our corpora differ between each other both in language and in composition. Therefore, the investigation of the influence of language and of translation described here needs to be complemented by an analogous investigation into the influence of corpus composition on the results.

We have already made significant progress in this investigation, notably by collecting relevant metadata about all four corpora and by computationally modeling the relationship between corpus composition and attribution performance. With respect to the metadata, we have collected information about the year of publication, the narrative form and the subgenre for each novel in each corpus. To enable cross-corpus application of the subgenre labels, they come from a small, closed set of eight different, relatively broad categories, including one category called 'other'. With respect to the computational model of the relation between corpus composition and attribution accuracy, we have fitted a multiple generalized Bayesian linear model that takes into account within-author and between-author variability and allows us to assess the positive or negative influence of metadata differences on textual distinctiveness (and, therefore, stylometric performance). However, further work on this issue is needed and presenting this work warrants a separate publication.

Apart from this, our research opens up several avenues for future research, in order to attempt to close some of the gaps just described in the section on limitations. For instance, the same kind of analysis could be performed using several corpora of different genres, size and composition for each language, in order to assess the degree of within-language variation and



Corpus: hun-ukr ▾

Feature Level: words, chars (chars selected)

Ngram Size: 2, 3, 4, 5 (5 selected)

Feature Type: lemmas ▾

Figure 5: Mean accuracy scores for the Hungarian corpus translated into Ukrainian, for different settings of MFF and sample size Parameters: Character 5-grams of lemmas.

the robustness of differences between languages that we have observed in this study. In addition, a study could be conceived that varies corpus composition deliberately and systematically, based on a significantly larger corpus with document-level metadata (at least) on subgenre and narrative perspective.

8. Data and code

Data and code are available online at https://gitlab.clsinfra.io/cls-infra/d33 (DOI: https://doi.org/10.5281/zenodo.11080205).

Acknowledgments

The research presented here as been performed as part of CLS INFRA, a project supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984.

References

[1] S. Argamon. "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations". In: *Literary and Linguistic Computing* 23.2 (2008), pp. 131–147. DOI: 10.1093/llc/fqn003.

[2] D. Bogdanova and A. Lazaridou. “Cross-Language Authorship Attribution”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik: Elra, 2014, pp. 2015–2020. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/145%5C%5FPaper.pdf.

[3] L. Burnard, C. Odebrecht, and C. Schöch. “In Search of Comity: TEI for Distant Reading”. In: *Journal of the Text Encoding Initiative* 14 (2021). DOI: 10.4000/jtei.3500.

[4] J. Burrows. “All the Way Through: Testing for Authorship in Different Frequency Strata”. In: *Literary and Linguistic Computing* 22.1 (2007), pp. 27–47. DOI: 10.1093/llc/fqi067.

[5] J. Burrows. “The Englishing of Juvenal: Computational Stylistics and Translated Texts”. In: *Style* 36.4 (2002). URL: https://www.jstor.org/stable/10.5325/style.36.4.677.

[6] J. Byszuk. “Analysis in authorship attribution”. In: *Survey of Methods in Computational Literary Studies*. Ed. by C. Schöch, J. Dudar, and E. Fileva. Trier: Cls Infra, 2023. DOI: 10.5281/zenodo.7782363.

[7] J. Cohen. “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[8] M. Eder. “Does size matter? Authorship attribution, small samples, big problem”. In: *Digital Scholarship in the Humanities* 30.2 (2015), pp. 167–182. DOI: 10.1093/llc/fqt066.

[9] M. Eder. “Short Samples in Authorship Attribution: A New Approach”. In: *Book of Abstracts of the Digital Humanities Conference 2017)*. Montréal: Adho, 2017. URL: https://dh2017.adho.org/abstracts/341/341.pdf.

[10] M. Eder, J. Rybicki, and M. Kestemont. “Stylometry with R: A Package for Computational Text Analysis”. In: *The R Journal* 8.1 (2016), p. 107. DOI: 10.32614/rj-2016-007.

[11] S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, and T. Vitt. “Understanding and explaining Delta measures for authorship attribution”. In: *Digital Scholarship in the Humanities* 32.2 (2017). DOI: 10.1093/llc/fqx023.

[12] R. Gorman. “Universal Dependencies and Author Attribution of Short Texts with Syntax Alone”. In: *Digital Humanities Quarterly* 16.2 (2022). URL: https://www.digitalhumanities.org/dhq/vol/16/2/000606/000606.html.

[13] H. v. Halteren, H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. “New Machine Learning Methods Demonstrate the Existence of a Human Stylome”. In: *Journal of Quantitative Linguistics* 12.1 (2005), pp. 65–77. DOI: 10.1080/09296170500055350.

[14] D. I. Holmes. “The Evolution of Stylometry in Humanities Scholarship”. In: *Literary and Linguistic Computing* 13.3 (1998), pp. 111–117. DOI: 10.1093/llc/13.3.111.

[15] M. Kestemont, K. Luyckx, W. Daelemans, and T. Crombez. “Cross-Genre Authorship Verification Using Unmasking”. In: *English Studies* 93.3 (2012), pp. 340–356. DOI: 10.1080/0013838x.2012.668793.

[16] C. Lee. “Do language combinations affect translators’ stylistic visibility in translated texts?” In: *Digital Scholarship in the Humanities* 33.3 (2018), pp. 592–603. DOI: 10.1093/llc/fqx056.

[17] G. Mikros and D. Boumparis. “Cross-linguistic Authorship Attribution and Author Profiling. Machine translation as a method for bridging the language gap”. In: *Digital Scholarship in the Humanities* 39.3 (2024), pp. 954–967. DOI: 10.1093/llc/fqae028.

[18] A. Nini. *A Theory of Linguistic Individuality for Authorship Analysis.* Cambridge University Press, 2023. DOI: 10.1017/9781108974851.

[19] J. Rybicki and M. Eder. “Deeper Delta across genres and languages: do we really need the most frequent words?” In: *Literary and Linguistic Computing* 26.3 (2011), pp. 315–321. DOI: 10.1093/llc/fqr031.

[20] J. Rybicki. “A Third Glance at a Stylometric Map of Native and Translated Literature in Polish”. In: *Retracing the History of Literary Translation in Poland.* New York: Routledge, 2021, pp. 247–261. DOI: 10.4324/9780429325366-20.

[21] J. Rybicki. “The great mystery of the (almost) invisible translator: Stylometry in translation”. In: *Studies in Corpus Linguistics.* Ed. by M. P. Oakes and M. Ji. Vol. 51. Amsterdam: John Benjamins, 2012, pp. 231–248. DOI: 10.1075/scl.51.09ryb.

[22] J. Rybicki and M. Heydel. “The stylistics and stylometry of collaborative translation: Woolf’s Night and Day in Polish”. In: *Literary and Linguistic Computing* 28.4 (2013), pp. 708–717. DOI: 10.1093/llc/fqt027.

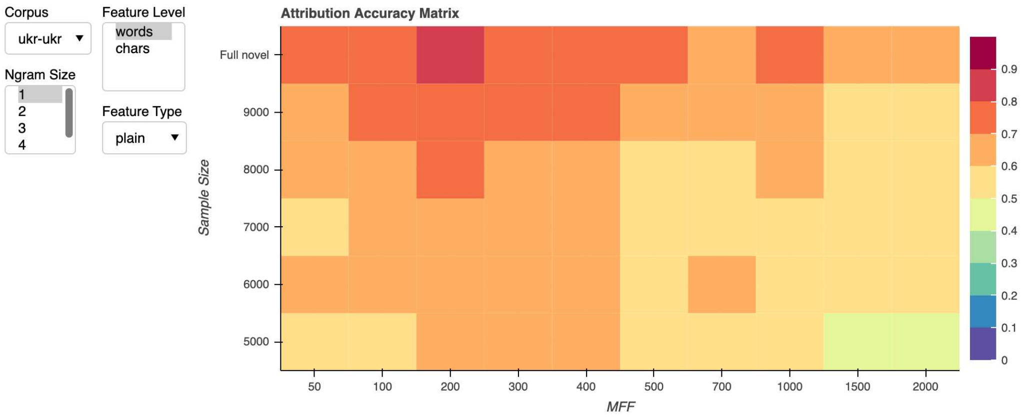
[23] C. Schöch, J. Dudar, and E. Fileva, eds. *Survey of Methods in Computational Literary Studies.* Cls Infra, 2023. DOI: 10.5281/zenodo.7892112.

[24] C. Schöch, R. Patraș, D. Santos, and T. Erjavec. “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives”. In: *Modern Languages Open* (2021). DOI: 10.3828/mlo.v0i0.364.

[25] P. W. H. Smith and W. Aldridge. “Improving Authorship Attribution: Optimizing Burrows’ Delta Method”. In: *Journal of Quantitative Linguistics* 18.1 (2011), pp. 63–88. DOI: 10.1080/09296174.2011.533591.

[26] E. Stamatatos. “A survey of modern authorship attribution methods”. In: *Journal of the American Society for Information Science and Technology* 60.3 (2009), pp. 538–556. DOI: 10.1002/asi.21001.

[27] E. Stamatatos. “On the robustness of authorship attribution based on character n-gram features”. In: *Journal of Law and Policy* 21.2 (2013). URL: https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/7/.


Corpus: ukr-ukr

Feature Level: words, chars

Ngram Size: 1, 2, 3, 4

Feature Type: plain

Figure 6: Ukrainian corpus, results for word form unigrams.

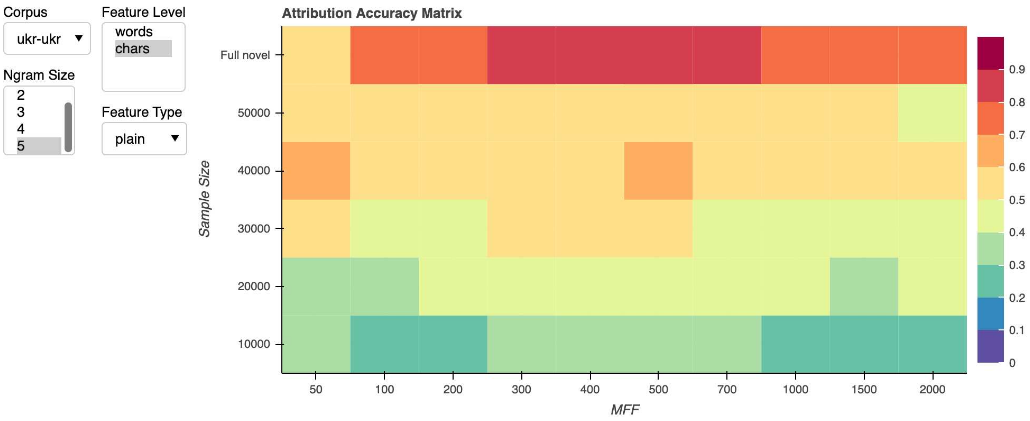
Appendix

Note that all result data can be inspected online in our interactive visualization at https://showcases.clsinfra.io/stylometry. A selection of findings is documented here for convenience.

A. Ukrainian

The Ukrainian corpus demonstrates lower performance compared to the Hungarian original corpora, achieving accuracies ranging from 50-80% when using unigram lemmas. Analyses based on word forms yields higher scores (Figure 6, 60-85%), particularly when the number of features is not higher than 400-500. An interesting observation is that variation in accuracy is a bit less across parameters in the Ukrainian original texts than for other corpora. When evaluating the accuracy of character-level analysis, high scores (70-85%) are observed using 3-5-gram characters with full novel sample size in word forms and lemmas settings (Figure 7).

Translations of the Ukrainian corpus into other languages demonstrate significantly lower results than the original texts, with accuracies surpassing 50% only for full novel sizes in word-level analysis with unigram word forms and lemmas. Among the translations from Ukrainian, the English translation shows the highest results, reaching up to 75% for full novel sample size (Figure 8). This observation holds true for analyses based on full novel sample size using character n-grams on translations as well. In most cases, the performance of translation into English achieves 80% accuracy for 5-gram characters in both word forms and lemmas. Concerning the mean performance, the translation into Hungarian has the highest mean accuracy in most settings.



Corpus: ukr-ukr

Ngram Size: 2, 3, 4, 5 (5 selected)

Feature Level: words, chars (chars selected)

Feature Type: plain

Figure 7: Ukrainian corpus, results for character 5-grams.

B. French

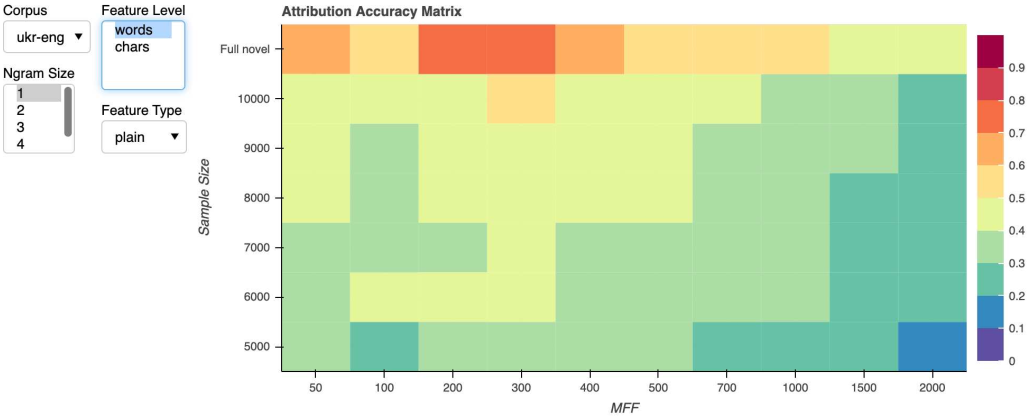
The French corpus shows a bit lower average performance compared to Ukrainian and Hungarian, with accuracy ranging from 40-70%, and in some cases, up to 80-95% for original texts based on unigram word forms and lemmas with N features up to 550 and a sample size starting from 8000 (Figure 9). Notably, the mean accuracy of original texts based on unigram word forms with sample size of 10000 and full novels is very high and reaches 83%, which approximates the performance of the Hungarian original corpus.

When discussing character n-gram features, we observe high scores (up to 80-95%) exclusively for full novels, with n-gram sizes ranging from 3-5 (Figure 10). This observation holds true for both lemmas and plain word forms.

Translations from French show a moderate decrease in performance compared to the original texts in both lemma- and word-form-based analyses, maintaining high accuracy only for the full novel sample size. Regarding the character n-grams, the scores of translations are almost identical to those of the original novels, with high performance achieved only for the full novel sample size.

C. English

The English corpus (Figure 11) yields the lowest performance among all corpora for original texts, with accuracy occasionally reaching 93% only on the full novel sample size in unigram lemma-based analysis at the word level. Apart from these instances, the corpus's performance displays high variability, including the lowest results around 20% for both lemma-based and



Corpus: ukr-eng

Feature Level: words

Ngram Size: 1

Feature Type: plain

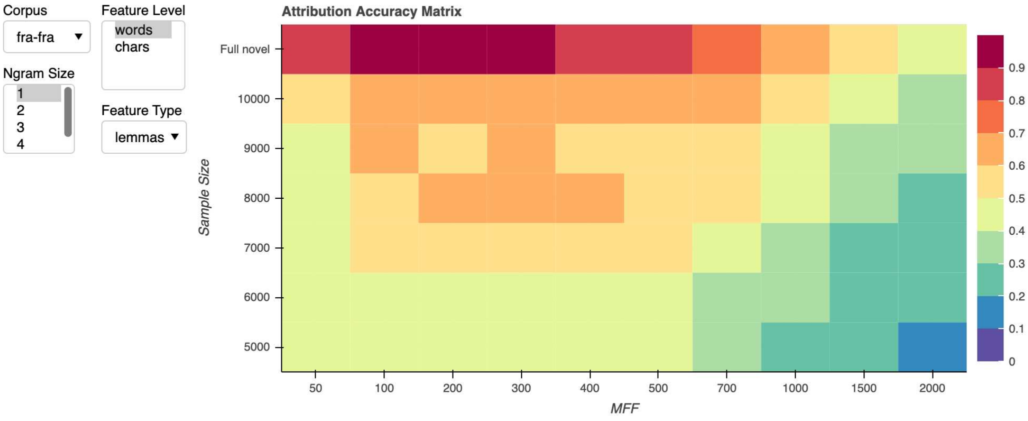
Figure 8: English translations of Ukrainian texts, performance for word form unigram.

plain word form-based analyses at the word level.

Regarding the classification performance when using character n-grams, we observe a similar trend as seen in the case of the French original corpus, where high accuracy is evident only for full novels. However, for the English corpus, the performance is slightly lower, reaching 86% in the best cases (Figure 12).

Importantly, the Hungarian translation of the English corpus (Figure 13) approximates the performance of the original novels, and in some cases even outperforms the original novels. The highest accuracy of up to 96% (lemma-based analysis) and 90% (word forms-based) is achieved for full novel sample size at the word level. However, the performance decreases rapidly with an increase in the number of features and decrease in sample sizes.

The translation into Ukrainian (Figure 14) also demonstrates high performance, up to 86% for full novels, although the average accuracy is lower compared to the original corpus and the translation into Hungarian (Figure 13). On the other hand, the translation into French achieves best performances at around 84% for lemma-based analysis, with most cases hovering around 20-30% (Figure 15). Regarding the analysis of translations at the character level, Hungarian translation outperforms the original corpus, in full novels sample size setting, while French translations exhibit lower accuracy even for full novels, reaching a score of 80% only for the 1500 and 2000 features setting for word forms and lemma-based analysis.



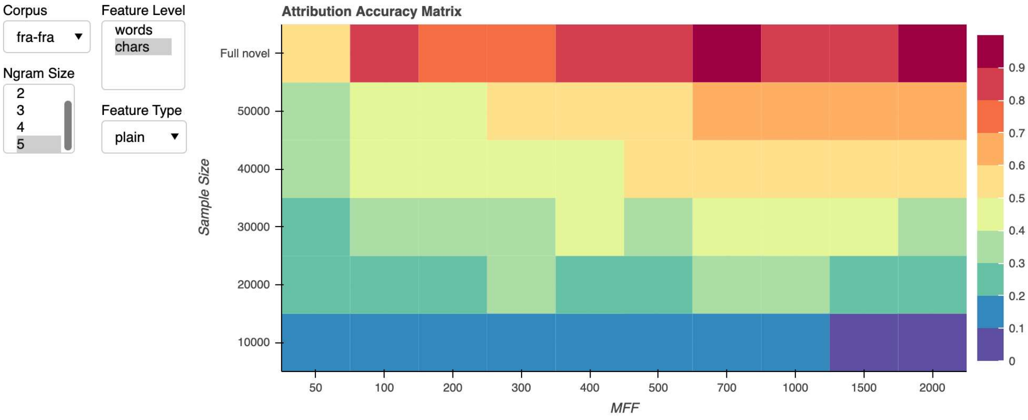
Corpus
fra-fra ▼

Feature Level
[✓ words]
chars

Ngram Size
[✓ 1]
2
3
4

Feature Type
lemmas ▼

Figure 9: Results of French original novels, based on lemma unigrams.



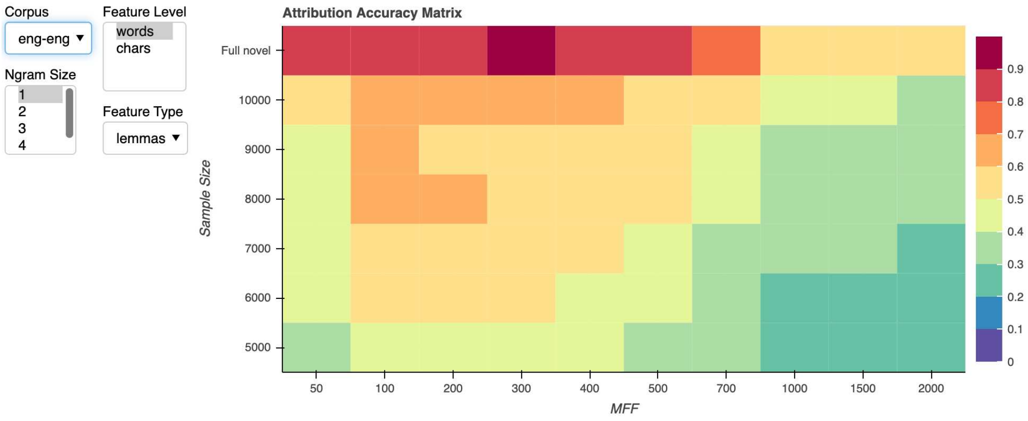
Corpus
fra-fra ▼

Feature Level
words
[✓ chars]

Ngram Size
2
3
4
[✓ 5]

Feature Type
plain ▼

Figure 10: Results of French original novels, based on character 5-grams.



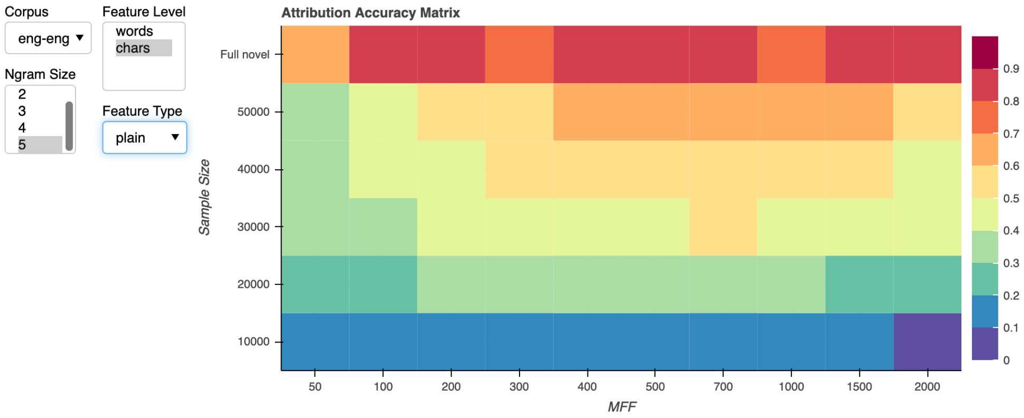
Corpus
eng-eng ▾

Feature Level
words
chars

Ngram Size
1
2
3
4

Feature Type
lemmas ▾

Figure 11: Results of the English original corpus, based on lemma unigrams.



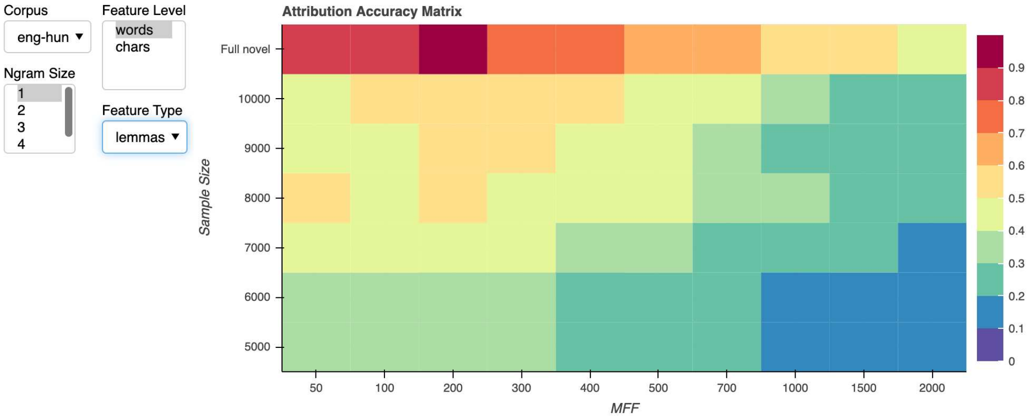
Corpus
eng-eng ▾

Feature Level
words
chars

Ngram Size
2
3
4
5

Feature Type
plain ▾

Figure 12: Results of the English original corpus, based on character 5-grams.



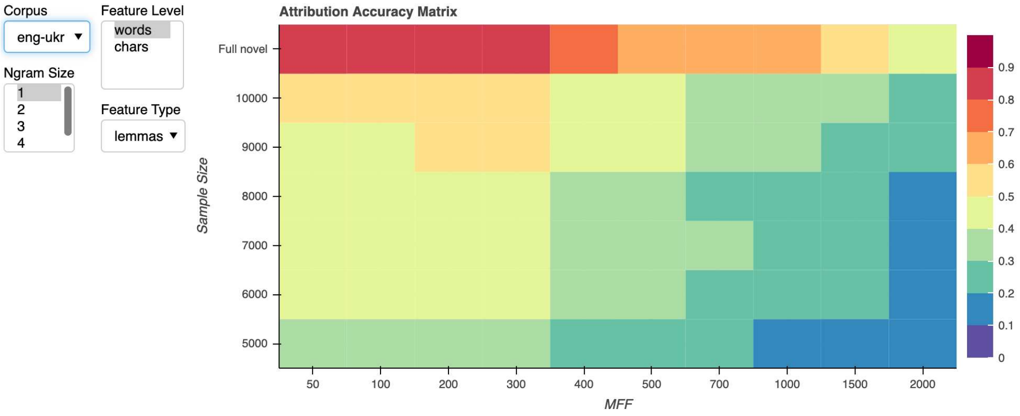
Corpus
eng-hun ▾

Feature Level
words
chars

Ngram Size
1
2
3
4

Feature Type
lemmas ▾

Figure 13: Performance of translations from English into Hungarian, lemma unigrams.



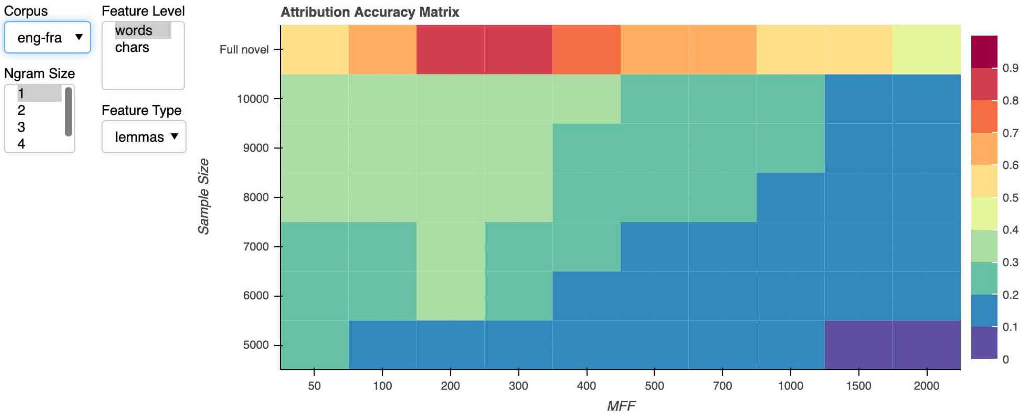
Corpus
eng-ukr ▾

Feature Level
words
chars

Ngram Size
1
2
3
4

Feature Type
lemmas ▾

Figure 14: Performance of translations from English into Ukrainian, lemma unigrams.



Corpus: eng-fra ▾

Feature Level: words (selected), chars

Ngram Size: 1 (selected), 2, 3, 4

Feature Type: lemmas ▾

Figure 15: Performance of translations from English into French, lemma unigrams.