DT-grams: Structured Dependency Grammar Stylometry for Cross-Language Authorship Attribution

Benjamin Murauer
Universität Innsbruck, Austria
b.murauer@posteo.de

Günther Specht
Universität Innsbruck, Austria
guenther.specht@uibk.ac.at

ABSTRACT

Cross-language authorship attribution problems rely on either translation to enable the use of single-language features, or language-independent feature extraction methods. Until recently, the lack of datasets for this problem hindered the development of the latter, and single-language solutions were performed on machine-translated corpora. In this paper, we present a novel language-independent feature for authorship analysis based on dependency graphs and universal part of speech tags, called DT-grams (dependency tree grams), which are constructed by selecting specific sub-parts of the dependency graph of sentences. We evaluate DT-grams by performing cross-language authorship attribution on untranslated datasets of bilingual authors, showing that, on average, they achieve a macro-averaged F1 score of 0.081 higher than previous methods across five different language pairs. Additionally, by providing results for a diverse set of features for comparison, we provide a baseline on the previously undocumented task of untranslated cross-language authorship attribution.

1. INTRODUCTION

In cross-language authorship attribution, the true author of a previously unseen document must be determined from a set of candidate authors after training a model with documents from those candidates in a different language. Previous work in single-language attribution often relies on language-specific features. Here, popular and powerful features often exploit character- and word-based measures [16, 2]. Using translation enables easy re-use of these features, and has been shown to be a useful tool for cross-language attribution [1]. However, setting up a custom machine translation system is an expensive operation in terms of time and resources. From a scientific perspective, translations from commercial and therefore, closed-source systems are difficult to explain and reproduce, as the details of the models are unknown to the customer and commercial providers will likely try to improve their models, causing different translations of the same input over time. Therefore, language-independent alternatives to traditional attribution features are crucial for cross-language attribution without translation.

Candidates for such features include high-level measurements like vocabulary or punctuation statistics [11] or features that can be mapped to a general space like universal grammar representations [1]. In this paper, our first contribution is a novel type of classification feature, DT-grams (dependency tree grams), that is based on dependency graphs and universal part-of-speech (POS) tags, making it language-independent. It calculates frequencies of substructures within a dependency graph similar to how in traditional n-grams, frequencies of character or word combinations in the original text are counted. We show that this feature is efficient for cross-language authorship attribution, a problem in which documents of bilingual authors are classified, but the language differs between training and testing documents. In our experiments, DT-grams outperform other approaches in this field consistently by an average $F1_{macro}$ score of 0.081.

For the authorship attribution experiment, we use a dataset consisting of social media comments of bilingual authors in multiple language pairs. This distinguishes this work from previous research, which used artificially constructed corpora due to the lack of data from multilingual authors [1, 7]. Thereby, classic novels by professional authors were used as training data, and human-translated versions of other novels by the same author are used as evaluation data. Although research has shown that human translation does not eliminate stylometric features [20], the original author still has only written in one language. Therefore, we argue that the classification problem is, more strictly speaking, a translation obfuscation measurement rather than an authorship attribution problem. By performing our evaluation experiments on the untranslated data from bilingual authors, we add a second contribution to this paper by providing the first baseline for true, untranslated cross-language authorship attribution.

Summarized, our contribution in this paper is twofold: (1) we present a new feature type DT-grams for cross-language authorship analysis, and (2) our evaluations represent a baseline for the novel problem of true, untranslated cross-language authorship attribution. To ensure the reproducibility of our results, all of our data and code is published online[1].

2. RELATED WORK

Cross-language authorship analysis is a significantly more difficult problem than its single-language version [16], and

32^{nd} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 25.05.2021 - 28.05.2021, Grimma, Germany.

[1] https://git.uibk.ac.at/csak8736/gvdb2021-code

in many cases, know-how learned from single-language authorship analysis can't be directly used. For example, simple syntactic features like word or character n-grams are an effective feature for stylometry [5], but are not suitable when the training and testing documents only share a few words, or even characters when given a different alphabet. Generally, using grammar features for authorship classification has been proven effective in many tasks ranging from attribution [9, 21, 4] to plagiarism detection [19]. Although these examples use language-specific grammar features in single-language settings, they show the general ability of these features to distinguish authorship, and language-independent grammar features such as universal POS tags allow for cross-language classification [1].

Using different combinations of words by leveraging the dependency of sentences rather than the original word order has lead to increased classification performance [15]. However, this study does not make use of language-independent features but rather changes how word n-grams are constructed by providing an alternative measure of which words neighbor each other. Nevertheless, their findings suggest that the dependency relationships between words within sentences hold valuable information for authorship analysis.

Our proposed feature, DT-grams, leverages key findings of previous observations by combining language-independent universal POS tags in combination with dependency graphs.

Previous attempts at cross-language attribution define the task itself inconsistently and different approaches to this term are taken, including datasets of monolingual authors of different languages [17] or comparing the performance of feature families in mono-lingual attribution problems for different languages [2]. When refining the definition of cross-language attribution as the task of attributing authors that have written documents in multiple languages, and training and testing documents must be written in different languages, few existing studies remain: [1] use a variety of different features including the frequency of universal POS tags on attribution, but conclude that machine-translation followed by traditional attribution techniques provides the best results. [7] use differently sized windows in which vocabulary richness measurements are aggregated. However, in both works, the datasets that were used contain human-translated novels, where the original author only wrote in one language and the source of the other languages was added by using translations of these works. Although it has been shown that translation keeps stylistic features mostly intact [20], we claim that the setup by these studies more likely measures the extent to which the authorship was obfuscated by the translator rather than the authorship itself. We state that authors writing in multiple languages are likely to do so in different styles, and we distinguish this problem as a different type of task.

Therefore, in this paper, we use social media texts that have been written by bilingual authors [10]. While this change in text type makes it more difficult to compare the results directly to previous work, it also allows us to analyze a more comprehensive set of language pairs that are available within this resource, and have not been included in previous studies due to the lack of data. More importantly though, by using this resource, our evaluations of the DT-grams feature along with several previously established baseline features provide first reference results for untranslated authorship attribution in five different language pairs.

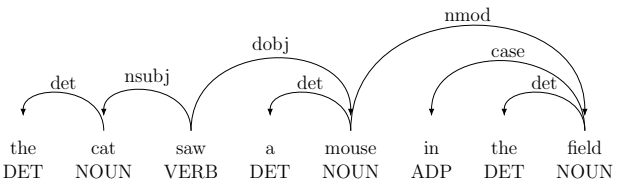


Figure 1: Dependency graph representation of the sentence 'the cat saw a mouse in the field'.

3. DT-GRAMS CONSTRUCTION

To construct the proposed DT-grams feature, we parse textual data to obtain dependency relationships between the words within sentences, which are then mapped to a tree structure. Then, differently sized substructures are selected from those trees to produce sequences of DT-grams. Finally, while some classification models used in our experiments use these sequences directly, we also reduce them to tf/idf-normalized frequencies to form a bag-of-DT-grams for other models used in the evaluation. In the following section, these steps are explained in detail.

3.1 Grammar Representations

In the first step, the raw text is parsed by a dependency parser. For this, we use the *stanza*[2] python library. This produces graphs as depicted in Figure 1. Along with the dependency graph, the parser also provides additional information for each word, including its lemma and universal POS tag. The latter is a mapping from the more fine-grained language-dependent POS tag to a coarse, but language-independent universal tag [12], and we use it as a supplemental representation of the word itself and by discarding the original word. This way, we construct a language-independent tree from the graph of each sentence, and encode both the relationship between the words as well as their grammatical role.

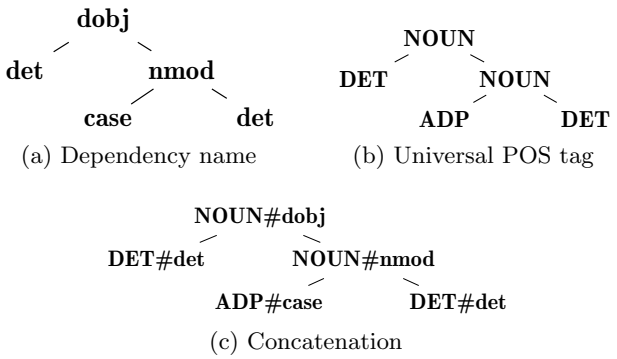
We test three different representations of the nodes within the tree which are depicted in Figure 2: (1) the name of the incoming dependency (Figure 2a), (2) the universal POS tag of the word (Figure 2b), and (3) both (Figure 2c). This way, we hope to gain insight into which parts of the dependency graph are more important for authorship stylometry. The resulting influence of these choices is discussed in Section 5.

A similar representation of sentences can be achieved by using constituency parsers, which we refrained from using for two reasons: firstly, the availability of parser models for non-English languages is limited, and secondly, the resulting constituents are not language-independent and a global mapping must be used in order to perform cross-language classification. While such mappings exist for POS tags [12], no similar resources for constituents are available to our knowledge.

3.2 Tree Substructure Representations

Along the lines of [19], we use patterns of tree structures representing parts of the dependency tree. We propose several patterns, which we collectively call DT-grams and which are displayed in Figure 3. The intention behind choosing these specific structures is as follows: We first extract node combinations from direct ancestors (DT_{anc}, Figure 3a) and

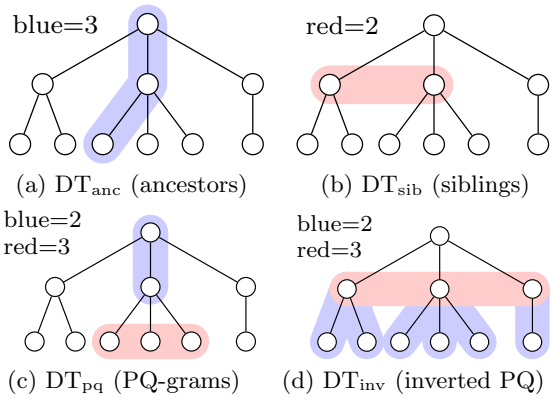
[2] `https://github.com/stanfordnlp/stanza`



(a) Dependency name (b) Universal POS tag

(c) Concatenation

Figure 2: Three node representations of the dependency graph of the subphrase "mouse in the field" from Figure 1 containing the name of the dependency (a), the universal POS tag (b), and both (c).



(a) DT_{anc} (ancestors) (b) DT_{sib} (siblings)

(c) DT_{pq} (PQ-grams) (d) DT_{inv} (inverted PQ)

Figure 3: DT-grams. Substructures are based on simple tree building blocks (a, b), PQ-grams by [19] (c) and an inverted form thereof (d).

siblings (DT_{sib}, Figure 3b), representing the most basic building blocks of a tree. In Figure 3c, DT_{pq} is displayed, based on the PQ-grams used by [19]. Finally, we add DT_{inv} that use a different order of sibling/ancestor relationship (Figure 3d) compared to PQ-grams.

While character and word-based n-grams only have one dimension to scale (namely, n), these tree substructures can have more. In general, two parameters control the number of siblings (red) and ancestors (blue) taken into account for each pattern, whereas DT_{anc} and DT_{sib} both only have one of those parameters each. For DT_{anc} and DT_{sib}, setting the parameter to 1 results in calculating POS tag unigramsd.

To get instances of the DT-gram patterns from a tree, the substructure patterns are moved across the tree similar to a sliding-window, generating an instance of the substructure at every step. Thereby, one has to define an order in which the DT-grams are parsed from the trees (i.e., depth-first or breadth-first). If a substructure does not fit onto a certain position of a tree, the empty spots in the pattern are filled with a wildcard element `X`. Thereby, an instance is generated for every step as long as at least one of the substructure's positions is filled with a non-wildcard node.

This way, the sequence of DT-grams can either be used directly as input for a sequence-based model (e.g., a recurrent network), or the frequencies of the parsed instances can be used analogously to those of character or word n-grams.

For example, applying DT_{anc} shown in Figure 3a with its parameter set to 3 to the tree in Figure 2a results in 11 substructures: *X-X-dobj*, *X-dobj-det*, *dobj-det-X*, *det-X-X*, *X-dobj-nmod*, *dobj-nmod-case*, *nmod-case-X*, *case-X-X*, *dobj-nmod-det*, *nmod-det-X*, *det-X-X*,

Finally, the frequency of each produced instance is counted over the entire document, and these frequencies are then tf/idf-normalized over the entire dataset.

4. EVALUATION

To evaluate the DT-grams feature, we perform cross-language authorship attribution using data from multiple language pairs and different classifiers, and we compare the results to different baseline features.

4.1 Datasets

Since there are no untranslated cross-language corpora available to our knowledge, we use the framework by [10] to generate several datasets by bilingual authors in different languages. It collects user comments from the social media site Reddit and allows us to set minimum requirements for document count, length, and language. We use this resource to evaluate the performance of DT-grams for different language pairs and generate bilingual datasets for the combinations presented in Table 1. We choose five different language pairs which all contain English, which represents the largest portion of text in Reddit comments. The other languages were chosen as they represent the largest non-English text sources for this corpus. We set the parameters of the generation framework to produce corpora with at least 10 authors for each pair, where each author has at least 20 documents for both languages. To increase the quality of the text documents, we also required a minimum document length of 3,000 characters. The tools that generate these corpora perform preprocessing including replacing URLs with a tag `<URL>` or filtering messages that mainly consist of punctuation. For a full list of preprocessing steps, we refer to the original publication by [10]. We performed no additional preprocessing. The resulting corpora are shown in Table 1 and we provide them publicly for download[3].

Languages	A	Docs	L_{doc}	D/A_{min}
EN + DE	10	2,790	3,055	22 + 20
EN + DeepL	10	2,790	3,055	22 + 20
EN + ES	20	3,402	3,148	20 + 21
EN + PT	37	4,481	2,996	20 + 20
EN + NL	11	2,056	3,225	20 + 20
EN + FR	45	7,374	3,142	21 + 20

Table 1: Datasets used for evaluation. A denotes the number of authors. L_{doc} denotes the average document length in characters. D/A_{min} denotes the minimum number of documents written by each author in the respective languages in the first column. "DeepL" corresponds to the German documents machine-translated to English with DeepL.

In previous work, mono-lingual attribution techniques on machine-translated documents outperform cross-language

[3] `https://git.uibk.ac.at/csak8736/gvdb2021-code`

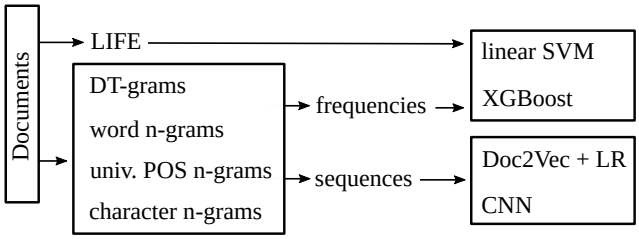


Figure 4: Models used in the experiments.

techniques [1]. We therefore provide data to calculate such a baseline by using the commercial translation service DeepL[4] to translate the German documents to English, creating a mono-lingual version of the German documents for comparison. However, due to budgetary reasons, we only perform this step for one randomly picked language (German).

For each language pair (A, B), we conduct all experiments both with training on A and testing on B, as well as the other way around.

4.2 Evaluation Strategy

Since the parameterized datasets only define lower limits for the number of documents per author and the size of these documents, the resulting datasets have varying amounts of documents and authors. We ensure that results from experiments using these datasets can be easily compared to each other by only selecting 10 random authors of each dataset, and selecting 10 random documents of each language from those authors.

To reduce bias, each of these evaluations is repeated 10 times, and the selected authors and documents are randomized in each repetition. For each of these repetitions, all combinations of features and classifiers are tested, and the mean value of each combination across all repetitions is used as a representative for that combination. This also functions as a supplement for traditional cross-validation, which is impossible for cross-domain classification as documents in the training set can't be used interchangeably for testing, which would break the cross-domain nature of the setup. We are aware that this results in some datasets having a larger overlap between the repetitions than others, which is a flaw that might be mitigated in the future if more comprehensive corpora of bilingual authors become available, or direct comparison between results originating from differently sized datasets is not important.

4.3 Models and Baselines

We test several different text classification models by following previous approaches in authorship attribution tasks. These are summarized in Figure 4.

Firstly, calculating tf/idf-normalized frequencies of different types of n-grams has been used widely in the authorship analysis field, including character, word, or part-of-speech tag n-grams. This approach can be used analogously by counting the frequencies of the parsed DT-grams and normalizing them using tf/idf. We then test two commonly used classifiers: linear SVMs [16, 11, 6] and extreme gradient boosting [7]. As comparison baselines of this category, we include results from character, word, and universal POS tag n-grams, whereby n ranges from 1 to 5.

Secondly, we utilize the Doc2Vec document embedding technique in combination with a logistic regression classifier, as proposed by [3]. For this solution, we have to define what a document is in terms of DT-grams, as their order is no longer well-defined. We interpret each document as the sequence of DT-grams that is returned by the parser, which in our case uses a depth-first approach. We include baselines for comparison along the lines of [3], which consist of character, word, and universal POS n-grams ranging from n=1 to 5.

Thirdly, we use a convolutional neural network proposed in [14] by interpreting each DT-gram as a unique token used in the embedding layer of the network. Thereby, we use the same parameters and network layout as in [14], except for an increased embedding layer size to fit the larger documents. We utilize the same depth-first order as in the second approach to define a sequence of tokens. The baseline for this model uses character, word, and universal POS tag unigram representations of the documents.

As a further comparison baseline, we compute the vocabulary richness feature LIFE from [7], which counts the vocabulary frequency over differently sized windows and calculates various aggregated measures. We refrain from using other language-agnostic features presented in related cross-language research [1], which depend on language-specific resources like sentiment databases, which are difficult to collect and even harder to compare. Additionally, in their research, these approaches showed inferior performance compared to character-based features from machine-translated text. We use the same linear SVM and extreme gradient boosting classifiers as the tf/idf frequency feature category to classify the documents with LIFE features (see Figure 4).

Parameter	Values
n-gram size	1 - 3
DT-gram structure	DT_{anc}, DT_{sib}, DT_{pq}, DT_{inv}
DT-gram dim. sizes	1 – 4, 1 – 4
C-value of SVM	0.1, 1, 10
Doc2Vec emb. size	50, 100, ..., 250
CNN batch size	5, 10, 20

Table 2: Hyperparameters optimized by grid search. All n-gram sizes were tested individually for word, character and universal POS-tag n-grams.

5. RESULTS AND DISCUSSION

We run the classification experiment for each model, each language pair in both directions, and every parameter combination shown in Table 2, generating an exhaustive grid of results. In this section, different aggregations and selections of this entire result set are used to extract the key findings for this paper.

5.1 Performance per Model

Table 3 shows that the linear support vector machine with tf/idf frequency features outperforms all other models in every language combination and for most of the feature categories. In the case of the vocabulary richness feature LIFE, we can confirm the results of the original work that the random forest-based approach outperforms the support vector machine [8].

We suspect that the CNN model underperforms because we

[4]`https://www.deepl.com/`, translation performed in November 2019

Model	EN/DE	EN/ES	EN/FR	EN/NL	EN/PT	EN/DeepL
svm	**0.375**	**0.291**	**0.310**	**0.277**	**0.246**	**0.479**
xgb	0.268	0.207	0.229	0.209	0.175	0.332
cnn	0.112	0.108	0.104	0.102	0.119	0.133
d2v	0.261	0.180	0.179	0.193	0.213	0.344

(a) Max. $F1_{macro}$ score of the models across all datasets. "DeepL" denotes the German documents machine-translated to English with DeepL.

Model	LIFE	Word n-grams	Char. n-grams	Uni. POS n-grams	DT-grams
svm	0.110	**0.396**	**0.479**	**0.385**	**0.453**
xgb	**0.157**	0.189	0.332	0.282	0.328
cnn	-	0.092	0.075	0.133	0.102
d2v	-	0.143	0.341	0.344	0.336

(b) Max. $F1_{macro}$ score of the models across different features.

Table 3: $F1_{macro}$ of the models across different datasets (a) and features (b).

DT_g	EN/DE	EN/ES	EN/FR	EN/NL	EN/PT	EN/DeepL
DT_{anc}	0.33	0.21	0.23	0.23	0.18	0.42
DT_{sib}	0.29	0.24	0.24	0.25	0.25	0.42
DT_{pq}	0.35	0.26	0.28	**0.28**	**0.29**	**0.43**
DT_{inv}	**0.37**	**0.30**	**0.29**	0.23	0.27	0.43

Table 4: Max. $F1_{macro}$ score of each DT-gram type.

have significantly less training documents than in the original paper, in which case network models have been shown to have trouble capturing the style of authors [5].

While the document embedding model (d2v in the table) outperforms the frequency-based features with the extreme boosting trees in some cases, it does not reach the support vector machine's F1 scores in any language or feature set.

5.2 Performance per Feature Category

Figure 5 displays the highest $F1_{macro}$ score for each frequency feature category and dataset. It becomes clear that the vocabulary richness feature LIFE is not able to model the authors effectively. An explanation for this is found in the basic principle behind the feature itself, which counts aggregated vocabulary richness measures across sliding windows over the document. Being originally developed for classifying entire novels from professional authors allowed these window sizes to be large and carry more information than is the case with shorter texts. Likewise and unsurprisingly, the word n-grams are not able to model authorship except for the machine-translated dataset, which is the only case where a significant intersection between training and validation vocabulary can be expected.

Confirming the results of [1], we observe that traditional features are effective in classifying machine-translated text, outperforming all other features. We can also confirm their finding that machine-translation increases the performance of language-independent features. Interestingly, the character n-gram features perform well above the 10% random baseline also for the non-translated datasets. This suggests a measure of similarity between these languages, but we leave the interpretation of these results to the field of linguistics. Future

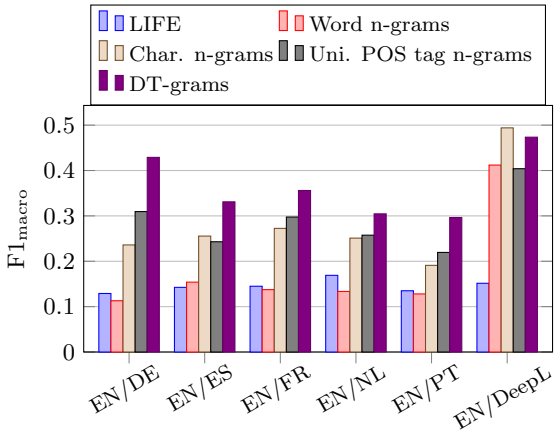


Figure 5: Comparison of the highest $F1_{macro}$ scores for different feature types. The different datasets are plotted on the x-axis, where "DeepL" stands for the documents that have been machine-translated from German to English. For layout reasons, experiments that differ only in classification direction (e.g., en → de and de → en) are averaged, whereas the difference in $F1_{macro}$ between the directions was below 0.02 for each pair. The DT-gram feature outperforms the next best feature by 0.081 $F1_{macro}$ averaged over all untranslated language pairs.

experiments including datasets from less related language families such as Japanese or Arabic may provide further insights into this relationship.

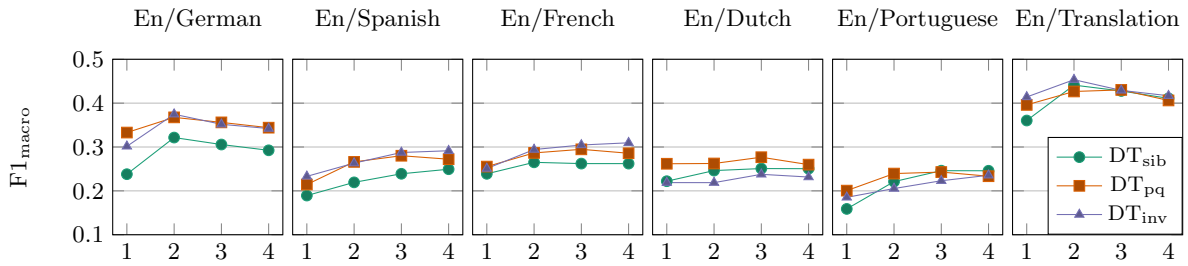
The proposed DT-gram feature is the most effective feature for the untranslated scenarios, outperforming the next best feature across the language pairs by an average of 0.081 $F1_{macro}$.

This suggests that the grammatical characteristics of multilingual authors are kept across languages. The performance of these features consistently outperforms n-grams constructed from the universal POS tag-based on the original word order, we conclude that the dependency relationships between the words and therefore, a grammatical style contribute to an author's stylometric fingerprint.

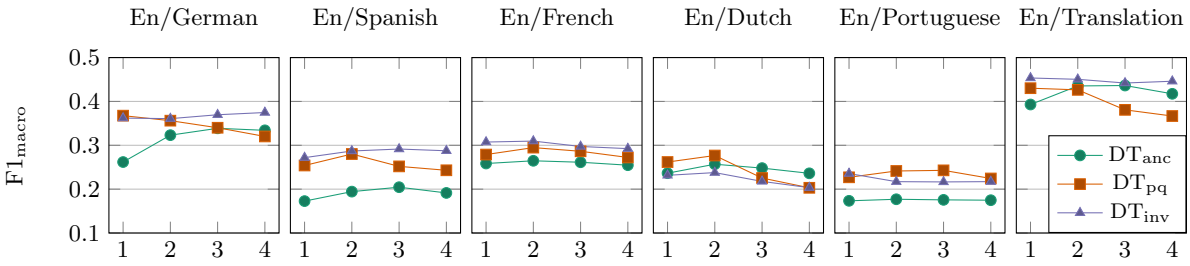
When comparing the different languages, we can see a clear difference in classification performance. For the two grammatical feature types, namely universal POS tag n-grams and DT-grams, the results of the German dataset show better F1 scores compared to the other languages. One possible explanation for this result the overall higher grammar complexity of German compared to the other languages [13], which would, in turn, suggest that either (1) classification across languages with grammars of different complexity, or (2) classification across languages with general high complexity improve the usefulness of grammar features themselves.

However, to answer these questions, additional language combinations must be analyzed, which may prove difficult for low-resource languages given the already small amount of available data from bilingual authors for languages that are not considered low-resource.

In summary, no approach is able to beat traditional methods performed on machine-translated texts, but our proposed DT-gram feature outperforms all other tested features on untranslated cross-language scenarios, especially on German documents. It represents a promising start for future de-



(a) Influence of the horizontal (red) parameter value (x-axis) on the $F1_{macro}$ score (y-axis).



(b) Influence of the vertical (blue) parameter value (x-axis) on the $F1_{macro}$ score (y-axis).

Figure 6: Influence of the horizontal (a) and vertical (b) DT-gram parameter sizes. Note that DT_{sib} is only included in (a) as it lacks a vertical parameter, and likewise, DT_{anc} is only included in (b).

Node	EN/DE	EN/ES	EN/FR	EN/NL	EN/PT	EN/DeepL
Dep.	0.366	0.239	0.274	0.257	0.218	0.445
U.POS	**0.375**	**0.291**	**0.310**	**0.277**	**0.246**	0.450
both	0.368	0.232	0.294	0.262	0.235	**0.453**

Table 5: Max. $F1_{macro}$ scores of different internal node layouts for the dependency tree.

velopment and research of true cross-language authorship attribution.

5.3 Performance by Tree Node Structure

As described in Section 3.1, we tried different representations of the internal nodes of the dependency tree structure. In Table 5, the best results for each of these can be found. Interestingly, the type of dependency which is used in the graph does not seem to have a large impact on the classification performance, but rather using only the structure of the graph along with the universal POS tag of each word shows the biggest advantage.

5.4 Tree Substructure Performance Analysis

As we demonstrated the general efficiency of the dependency tree-based features, Table 4 shows how the different DT-grams perform on each language combination. In general, the substructures that combine ancestor and sibling nodes (DT_{pq} and DT_{inv}) outperform the more simple patterns for each language and suggest that complex structures in grammatical style are a valuable stylometric feature for bilingual authors across languages.

Figure 6 shows a more detailed analysis of how the sizes of the two parameters influence this result. For both the vertical and horizontal parameters, the optimal value is between 2 and 3, depending on the language and substructure, which is similar to reported optimal values for character n-grams [16]. Only DT_{anc} benefits from a higher vertical parameter size, especially in German documents, which may benefit from even higher values of the respective parameter. While Spanish shows the least difference in classification performance across the different parameter sizes, it is difficult to draw conclusions from the other languages, indicating that more data is required for further experiments.

6. CONCLUSION

In this paper, we have presented a novel type of classification feature called DT-grams, based on dependency graphs and universal POS tags. We have shown in experiments that DT-grams able to efficiently model stylometric fingerprints of bilingual authors across languages, premiering authorship analysis even in cases where machine-translation is unavailable, with an average lead of 0.081 $F1_{macro}$ to the next best approach tested in our experiments. Additionally, we have expanded the field of cross-language authorship attribution by providing baseline results for the previously undocumented problem of untranslated cross-language authorship attribution of bilingual authors and analyzed results of 5 different language pairs. Finally, we have collected findings including unexpectedly good performances of language-dependent features applied to cross-language settings as well as significant differences across language pairs.

The most important limitations of our approach are the dependency on the performance of the external parsing tools used, which may differ in quality across languages, as well as the superior performance of approaches based on machine-translation.

In future work, we want to investigate on using more specialized syntax classification models like tree-LSTMs [18] or more complex syntactic networks [4], as well as combining multiple feature categories to further improve classification results in both cross- and single-language experiment settings.

7. REFERENCES

[1] D. Bogdanova and A. Lazaridou. Cross-language authorship attribution. In *Proceedings of the 9th International Conference on Language Ressources and Evaluation (LREC'2014)*, pages 2015–2020, 2014.

[2] M. Eder. Style-markers in authorship attribution : a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6(1):99–114, 2011.

[3] H. Gómez-Adorno, J.-P. Posadas-Durán, G. Sidorov, and D. Pinto. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756, 2018.

[4] F. Jafariakinabad and K. A. Hua. Style-aware neural model with application in authorship attribution. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 325–328. IEEE, 2019.

[5] M. Kestemont, M. Tschugnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, and M. Potthast. Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, editors, *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, 2018.

[6] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660. ACM, 2006.

[7] M. Llorens and S. J. Delany. Deep level lexical features for cross-lingual authorship attribution. In *Proceedings of the first Workshop on Modeling, Learning and Mining for Cross/Multilinguality*, pages 16–25. Dublin Institute of Technology, 2016.

[8] M. Llorens-Salvador. *Lexical rIchness Feature Extraction method (LIFE) for Multilingual and Cross-lingual Authorship Attribution*. Dissertation, Dublin Institute of Technology, 2018.

[9] K. Luyckx and W. Daelemans. Shallow Text Analysis and Machine Learning for Authorship Attribution. In *Proceedings of the 15th meeting of Computational Linguistics in the Netherlands*, pages 149–160. LOT, 2005.

[10] B. Murauer and G. Specht. Generating cross-domain text classification corpora from social media comments. In *Working Notes of the Conference and Labs of the Evaluation forum (CLEF'2019)*, pages 114–125. Springer, 2019.

[11] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314. IEEE, 2012.

[12] J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, 2016.

[13] M. Sadeniemi, K. Kettunen, T. Lindh-Knuutila, and T. Honkela. Complexity of european union languages: A comparative approach. *Journal of Quantitative Linguistics*, 15(2):185–211, 2008.

[14] P. Shrestha, S. Sierra, F. Gonzalez, M. Montes, P. Rosso, and T. Solorio. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 2017.

[15] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández. *Syntactic Dependency-Based N-grams as Classification Features*, volume 11 of *Mexican International Conference on Artificial Intelligence (MICAI'2012)*, pages 1–11. Springer Heidelberg Berlin, 2013.

[16] E. Stamatatos. On the Robustness of Authorship Attribution Based on Character N-Gram Features. *Journal of Law & Policy*, pages 421–439, 2013.

[17] L. M. Stuart, S. Tazhibayeva, A. R. Wagoner, and J. M. Taylor. Style features for authors in two languages. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 459–464. IEEE, 2013.

[18] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks, 2015.

[19] M. Tschuggnall and G. Specht. Countering Plagiarism by Exposing Irregularities in Authors' Grammar. In *Proceedings of the European Intelligence and Security Informatics Conference, (EISIC'2013)*, pages 15–22. IEEE, 2013.

[20] L. Venuti. *The translator's invisibility: A history of translation*. Routledge, 1995.

[21] R. Zhang, Z. Hu, H. Guo, and Y. Mao. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.