information

MDPI

Article

Morphosyntactic Annotation in Literary Stylometry

Robert Gorman

Department of Classics and Religious Studies, College of Arts and Sciences, University of Nebraska–Lincoln, Lincoln, NE 68588, USA; rgorman1@unl.edu

Abstract: This article investigates the stylometric usefulness of morphosyntactic annotation. Focusing on the style of literary texts, it argues that including morphosyntactic annotation in analyses of style has at least two important advantages: (1) maintaining a topic agnostic approach and (2) providing input variables that are interpretable in traditional grammatical terms. This study demonstrates how widely available Universal Dependency parsers can generate useful morphological and syntactic data for texts in a range of languages. These data can serve as the basis for input features that are strongly informative about the style of individual novels, as indicated by accuracy in classification tests. The interpretability of such features is demonstrated by a discussion of the weakness of an "authorial" signal as opposed to the clear distinction among individual works.

Keywords: stylometry; Universal Dependencies; authorship attribution

Citation: Gorman, R. Morphosyntactic Annotation in Literary Stylometry. *Information* **2024**, *15*, 211. https://doi.org/10.3390/info15040211

Academic Editor: Horacio Saggion

Received: 30 January 2024
Revised: 14 March 2024
Accepted: 15 March 2024
Published: 9 April 2024

1. Introduction

Stylometry is a discipline that attempts to apply rigorous measurement to the traditional concerns of stylistics. Stylistics involves the identification and evaluation of certain characteristics that may distinguish the language use of individuals, groups of individuals, genres, etc. For humanists, the objects of stylometric study are most frequently literary, historical, or philosophical texts. Recent years have seen much research in the application of stylometrics in the humanities, and this work has produced many advances in the field. However, there remain important weaknesses in the predominant methods in the field. The present study is an attempt to address aspects of these weaknesses: the lack of stylometric input features that produce results that are both (1) topic agnostic and (2) directly interpretable in traditional terms.

Generally, to be of interest to researchers, the characteristics of the "style" of a text must be distinctive enough to allow us to discriminate that text from other relevant material. Thus, from the early days of stylometrics [1], success in classification experiments has served to establish the stylometric value of the input features on which accurate classifications were based. Of course, considerations other than high accuracy must also be considered when evaluating input features. For example, the frequency profiles for a set of common words [2] or common word sequences—word n-grams [3] are generally quite effective for classification, but because these input features may include "lexical" words, they are usually avoided when the style of an individual writer is the focus. Lexical words, also called "content" words, can be strongly influenced by topic, genre, etc., and this influence may confound classification. In such a case, researchers rely upon features considered to be "topic agnostic" since they are not closely and directly dependent on the subject matter of the text in question. Chief among these topic agnostic inputs are "function" words and character n-grams. Unlike lexical words, function words (for example, prepositions, conjunctions, determiners, etc.) belong to a small, closed set. In spite of this fact, function words as a group are used more often than content words [4]. In addition, function words more closely reflect syntactic structure than semantic content. Thus, we can reasonably assume that function words are relatively free of confounding effects. Some function words may, however, be more closely dependent on genre or topic than others. Gendered

pronouns, for example, are for this reason sometimes removed from studies of function words [5]. On the other hand, while function words often allow for accurate classification and therefore clearly capture something distinctive about many texts, it is difficult to translate the frequency profile of a set of prepositions, conjunctions, etc., into a detailed understanding of the style of a text.

Character n-grams, which recently have become quite popular in textual studies, share, but to a more extreme degree, the advantages and disadvantages of function words. Consisting of character sequences without regard to their position in a word, their order in a sentence, etc., character n-grams represent a text at a sub-lexical level although, because spaces between words are usually counted as "characters", rough information about word boundaries is reflected in this input. For this reason, they are generally free of the criticism that they are closely dependent on external factors such as topic or intended audience (for reservations, see [6]). However, it should be obvious that the frequency distribution of randomized sequences of letters is practically uninterpretable in terms more of traditional approaches to style, and character n-grams are therefore uninteresting from that perspective. Thus, approaches to stylometry are closely connected to the ongoing debate in machine learning and related fields about the relative advantage of choosing, on the one hand, heuristic input features that may be difficult to interpret and, on the other hand, input that represents a symbolic structure such as syntax. (For a recent examination of the topic with a bibliography, see [7]).

Another example of a non-morphosyntactic computational analysis of texts is front-back vowel harmony testing [8,9], which tests whether there is a tendency of having words to have only front vowels or only back vowels. Front-back vowel harmony is so characteristic of certain languages that this feature can be detected even if these languages are written in an undeciphered syllabic script [8].

This paper is an introduction to the stylometric and stylistic value of the morphosyntactic information provided in the annotations of the Universal Dependency treebanks. First, using the standard criterion of text/authorship attribution, it will demonstrate that morphosyntactic input features can successfully discriminate among texts without the identification of any vocabulary items. This result indicates that these features can be effective while being topic agnostic. Second, it will show that many morphosyntactic input features can be interpreted in a relatively straightforward way that is consistent with terms and concepts long used in the precomputational study of literary style.

Advances in the field of interpretable machine learning have provided important tools for expanding the usefulness of ML by making results easier to understand, even with "black-box" algorithms (my thanks go to the anonymous reviewer for emphasizing this point). It nonetheless remains an advantage, at least when attempting to persuade researchers in literary fields of the validity of computational approaches, to select input variables that are explainable by referring to traditional stylistics. The academic study of literary style has its roots in the traditional disciplines of Poetics and Rhetoric [10]. Both approaches agree that among the most important parts of a description of the style of a text or corpus are analyses of diction and word arrangement.

Diction is essentially word choice or vocabulary. This traditional focus is also central to our investigations, in that information about every word in every text analyzed is included in our data. At the same time, morphological annotation allows us to abstract away from individual vocabulary choices. Each word is included in our input features not as a token of a particular lexical item but rather as a representation of the relevant morphosyntactic categories (part-of-speech, singular or plural, subject or direct object, etc.). Thus, in accordance with the traditional importance of diction in stylistic research, words remain the basic unit of analysis in this study, but in a way that seeks to be topic agnostic and avoid the confounding effects often introduced by a consideration of vocabulary.

The traditional importance of word arrangement to stylistic research in the humanities is also reflected in the input features chosen for this study. Information about the syntactic annotation of every word in the corpus is reflected in the input features. Syntactic

annotation explicitly encodes the relationship between words, and therefore, its use as a dimension of analysis can be seen as a natural expansion of a traditional approach.

Thinking of the morphosyntactical features used in these experiments as computational "enhancements" of the traditional pillars of literary style, diction and arrangement, will, it is hoped, promote a broad understanding of the approach. Interpretability should also be increased by our use of traditional terminology. Terms used for morphological categories and values are known to any serious researcher in literary style. The widespread adoption of dependency grammar is, admittedly, relatively recent, but generally, the protocols of dependency grammar are closely related to the traditional concepts and terminology of the humanistic study of language and literature.

Thus, a stylometric analysis based on the features presented here is well suited to contribute to a thoroughgoing investigation of the style of a work or author in a way that should be interpretable to a wide range of readers. In addition, in the course of our discussion, it will become clear that this morphosyntactic approach is effective, with minimal adjustments, across a range of languages. This quality is beneficial in a field where English texts have been the predominant source of, and testing ground for, stylometric methods [11]).

The organization of the remainder of this paper is as follows. First, in Section 2 the various corpora are described and a step-by-step construction of effective input features from morphosyntactic annotation is described. Section 3 explains the classification experiment used to demonstrate the stylometric value of the input features. In Section 4, the results of the classification are briefly discussed. In Section 5 morphosyntactic frequencies are used as the basis of an investigation into the interrelations between the "local" characteristics of individual novels and the more general "authorial" signature. A short conclusion rounds off the article.

2. Corpora and Morphosyntactic Input Features

The input features used in this study are derived from texts annotated according to the framework used in the Universal Dependencies Treebank Collection [12,13]. The Universal Dependencies (UD) project is an open community effort that has been growing rapidly in recent years. The project has given impetus to the development and publication of software implementing pipelines for tokenization, tagging, lemmatization and dependency parsing of texts in a wide range of languages. These invaluable programs—called UDPipes—cover a wide range of languages and are available for the R and Python environments as well as through a convenient web interface (https://lindat.mff.cuni.cz/services/udpipe/, accessed 1 January 2024).

Because the focus of this paper is the advantages of morphosyntactic features as stylometric tools for the humanities, corpora consisting of a selection of novels have been chosen. While much recent stylometric work has concentrated on social media texts and the like, this material is less central to the interests of humanists than more traditional literary writing. Our corpus includes novels in English, French, German, and Polish. A more diverse set of languages would have been preferable, but such works meeting the requirements of our study were not readily available. The design of our investigation calls for literary works that are similar in genre and chronology. As many authors as possible should be represented, and for each author, the set should include three separate novels. These criteria could be met by combining reference corpora freely available at github.com. The English, German, and Polish novels were made available by the Computation Statistics Group (https://github.com/computationalstylistics, accessed 1 January 2024). The French novels were a resource provided by *Computerphilologie Uni Würzburg* (https://github.com/cophi-wue/refcor, accessed 1 January 2024). Because we will compare the performance of morphosyntactic input variables for works in the different languages, all corpora were limited to the size dictated by the smallest set (German). As a result, each language corpus contains 15 authors, each represented by three different

works. In order to facilitate comparisons between the languages, only the first 20,000 tokens (excluding punctuation) of each work in each corpus are considered.

After the collection of a suitable set of texts, the next step is to generate the basic annotations from which the input features will be assembled. This processing is carried out with the appropriate UDPipes through the "udpipe" package for the R Software Environment [14] (R version 4.2.1). Raw text (.txt files) provided to the udpipe produces output in the CONLL-U format. An example of this output is given below (Tables 1–4).

Table 1. "Shallow" annotation output by UDPipe. Sentence: "It gives us the basis for several deductions" (Doyle, *The Hound of the Baskervilles*, 1901).

Token	Lemma	Upos	Feats
It	it	PRON	Case = Nom | Gender = Neut | Number = Sing | Person = 3 | PronType = Prs
gives	give	VERB	Mood = Ind | Number = Sing | Person = 3 | Tense = Pres | VerbForm = Fin
us	we	PRON	Case = Acc | Number = Plur | Person = 1 | PronType = Prs
the	the	DET	Definite = Def | PronType = Art
basis	basis	NOUN	Number = Sing
for	for	ADP	NA
several	several	ADJ	Degree = Pos
deductions	deduction	NOUN	Number = Plur

Table 2. "Deep" annotation output by UDPipe. Sentence: "It gives us the basis for several deductions" (Doyle, *The Hound of the Baskervilles*, 1901).

Head_Token_Id	Dep_Rel
2	nsubj
0	root
2	iobj
5	det
2	obj
8	case
8	amod
5	nmod

Table 3. "Shallow" annotation by UDPipe. Sentence: "There, however, stood only a single bowl" (Spyri, *Heidi*, 1880).

Token	Lemma	Upos	Feats
Da	Da	ADV	NA
stand	stehen	VERB	Mood = Ind | Number = Sing | Person = 3 | Tense = Past | VerbForm = Fin
aber	aber	ADV	NA
nur	nur	ADV	NA
ein	ein	DET	Case = Nom | Gender = Neut | Number = Sing | PronType = Art
einziges	einzig	ADJ	Degree = Pos | Gender = Neut | Number = Sing
Schüsselchen	Schüsselchen	NOUN	Gender = Neut | Number = Sing | Person = 3

Table 4. "Deep" annotation by UDPipe. Sentence: "There, however, stood only a single bowl" (Spyri, *Heidi*, 1880).

Head_Token_Id	Dep_Rel
2	advmod
0	root
2	advmod
5	advmod
7	det
7	amod
2	nsubj

For each token, the analysis gives the form as it appears in the text and its lemma. This information is not used in the method described here since our goal is to examine the discriminative power of morphosyntactic features. In addition, as noted above, general vocabulary may be largely dependent on genre or subject matter and may confound analysis. It is worth noting that the elimination of word forms and lemmas from consideration simplifies preprocessing and, to some degree, compensates for the time required to extract input features from the parsed text. Minimal clean-up of the .txt file is required; chapter titles and the like can be left in the document without affecting the results of the classification.

Leaving aside the form and lemma, the remaining columns in the UDPipe output shown in Tables 1–4 are essential to our method. The "upos" column contains the UD part-of-speech tags for each word. The "feats" column gives the morphological analysis. Morphology information has the form "TYPE = VALUE, with multiple features separated by a bar symbol (TYPE1 = VALUE | TYPE2 = VALUE). Consider, for example, the morphological data supplied for the word *us* in Table 1. The upos column assigns *us* to "pronoun" as its part of speech. The "feats" column then gives the following information: Case = Acc | Number = Plur | Person = 1 | PronType = Prs. This annotation can be read as follows. The grammatical case of *us* is accusative; its grammatical number is plural; it refers to the speaker, so it is considered grammatically a first-person word; lastly, *us* belongs to the pronoun subtype "personal".

A comparison of the "feats" column in Table 1 with that in Table 3 reflects an important typological difference among languages. Languages can vary significantly in their morphological complexity. For example, English nouns (*basis* and *deductions* in Table 1) are generally annotated only for grammatical number, while English adjectives (*several* in Table 1) show only grammatical degree (i.e., positive, comparative, and superlative). In contrast, German nouns (*Schüsselchen* "bowl" in Table 3) and adjectives (*einziges* "single") are considered to have grammatical gender (and case) as well as number. In addition to the natural differences between languages, complications can be introduced by the parser. For example, the UDPipe version used in this study ("german-hdt-ud-2.5-191206.udpipe" Wijffels 2019) does not assign a case to every instance of a noun or adjective. Instead, explicit annotation of case is generally restricted to words in which different cases are indicated morphologically: for example, occurrences of *Kindes* ("child's") are marked as genitive of the noun *Kind* ("child"); and occurrences of *bösem* are marked as dative of the adjective *böse* ("bad").

Parts of speech and morphology constitute what we may call "shallow" syntactic features. These features reflect some syntactical structures, but do not represent them directly. In contrast, the "head_token_id" and "dep_rel" columns are a direct representation of syntactic organization. The head token is the item that is the immediate syntactic "parent" of a given token. The "dep_rel" reports the dependency-type label as specified in the UD annotation guidelines. The dependency relation specifies the type of grammatical structure between the parent and target. From these columns, we can calculate the grammatical

structure of an entire sentence, as visualized in a dependency tree such as the one shown below (Figure 1).

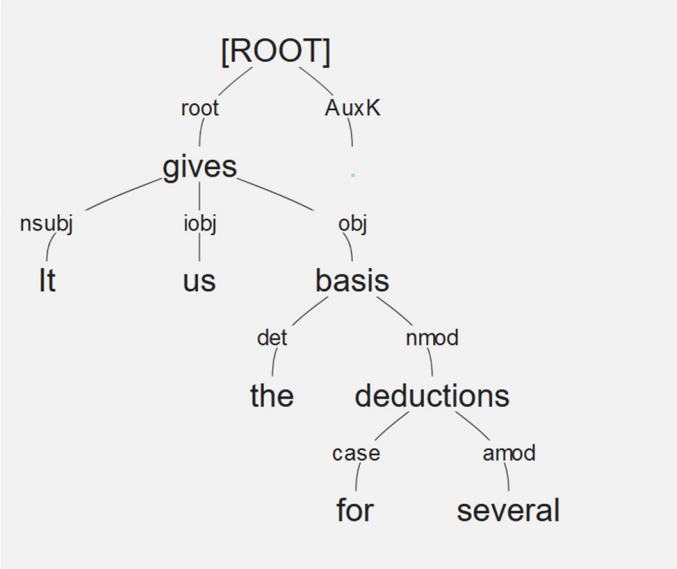

Figure 1. Universal Dependency tree for "It gives us the basis for several deductions".

The syntactic "path" from the sentence root to each "leaf" token is given by the combination of head id and dependency relationship. The syntactic function of each word is clearly and specifically defined by these two values. For example, the word *basis* is the *obj* of the word *gives*. *obj* is the UD label for what is traditionally called the "direct object" of a verb (a list of syntax labels along with examples can be found on the UD website: https://universaldependencies.org/en/dep/index.html, accessed 1 January 2024). The word *several* is labeled as *amod* of *deductions*. *amod* indicates an adjectival modifier. The word *deductions* itself is an *nmod* of *basis*. In UD annotation, *nmod* means "nominal modifier", a noun or noun phrase directly dependent on and specifying another noun (or noun phrase). For example, the prepositional phrase in "toys for children".

It is important to recognize the special importance of the "head_token_id" annotation. Because its values specify the configuration of the dependency tree for each sentence, head token information can also be used to add structural/syntactic "depth" to the "shallow" morphological data. For example, examined against the background of the dependency tree, the German word *Schüsselchen* ("bowl") is no longer just a neuter noun, but a neuter noun that is dependent on a past tense verb, or a neuter noun that is dependent on the main verb, etc. Thus, the head token annotation allows us to consider the "syntactic sequence" of words, a hierarchically ordered analogue to the chronologically ordered sequence encoded in traditional n-grams.

The input features in our study are composed primarily of the three kinds of information discussed above: (1) morphological annotation; (2) syntactic information; and (3) morphosyntactic "n-grams" containing combinations of morphological and syntactic data from words that are hierarchically contiguous in the dependency tree of a sentence.

When constructing input features from morphosyntactic annotation, it is important to design the features in a way that preserves information while avoiding sparsity. We can achieve this goal by incorporating, in the input features for a single word, a series of combinations of individual morphosyntactic values. In many languages, a morphological analysis of a word may be relatively complex. For example, the UD annotation for the German word *stand* (Table 3) indicates that its morphology may be identified as a *past indicative third-person singular finite verb.* Naturally, as the number of more or less independent values in a given complex annotation increases, the frequency of that set of values will correspondingly decrease. Thus, while 12.5% of the words in Spyri's *Heidi* are

annotated with the part of speech *verb*, only 4.2% are marked with a combination of the *verb* annotation and the tense annotation *past*. If we take syntactic function into consideration, only 1.6% of words are a past tense verb whose relationship is annotated as *root* (i.e., the main verb of a sentence). Such a sharp tendency toward sparsity will rapidly compromise the effectiveness of morphosyntactic data. To avoid this effect, we distribute the full morphological and syntactic annotation for each word into a set of combinations made from its assigned grammatical values. In this way, for example, each verb is associated with an input feature giving its tense, another giving its mood, another its person, etc. Then, all binary combinations of types are generated (e.g., tense and mood, tense and person, or mood and person). The same is carried out for ternary combinations. The result is a framework for organizing input features that is satisfactorily informative while maintaining an acceptable level of sparsity.

In addition to encoding the morphosyntactic information for each word in the text, we take advantage of the opportunity afforded by the head token annotation to enrich the data, as mentioned above. For each word that is not annotated as the *root* of a sentence, we include input features constructed from the morphosyntactic annotation of that word's dependency "parent". For example, the input features for *deductions* in Table 1 would include combinations made from *basis* as well as *deductions* itself. These syntactically ordered n-grams bring a measure of structural depth to otherwise shallow "surface" morphological information.

In addition to the morphosyntactic categories discussed so far, we have also included a small additional group in the feature set. Natural language may be conceptualized as a hierarchical structure (as illustrated in dependency treebanks) projected onto a linear order, the chronological sequence of words in texts or speech. Word order, as well as word hierarchy, can represent crucial stylometric information. We capture some of this linear information by adding two values to the annotations provided by UDPipe: dependency distance (DD) and dependency direction (DDir).

DD is the distance in the linear order of a sentence between a given word and its parent word, measured by the number of words. More precisely, DD can be thought of as the absolute value of the difference between the linear index of a word (its position in the linear sequence) and the linear index of the word's parent. Thus, in our example sentence, "It gives us the basis for several deductions", the DD of the word *us* is 1: the index of *us* = 3, the index of parent word *gives* = 2; hence, $3 - 2 = 1$. As treebanks of many languages become widely available, research on DD is becoming more important. DD has been suggested as a proxy for sentence complexity [15–17] and as an explanation for aspects of word order. It is therefore reasonable to include DD among our input features on the assumption that it represents something important about the style of a text. A second addition to our set of categories is dependency direction (DDir). This category is quite simple. A value is assigned to each word (except for sentence roots) indicating whether it comes before or after its parent word in the linear order of the sentence. Word order has long been a staple of analyses of stylistics, so it naturally finds a place in a stylometric study (computational studies based on treebank data for DDir tend to be focused on typological questions rather than stylistics ones [18]).

The restriction of our input features to unary, binary, and ternary combinations of annotation categories is an attempt to balance the desire to include the widest range of possibly useful stylometric data with the need to avoid a sparse set of inputs. Nevertheless, additional culling of the input features is necessary. The limitation of combinations to more than three elements still allows for over 16,000. And each of these combinations is a type, not a variable. Each component of a given type may represent more than one value; the ternary combination gender–number–case may take one of 24 different value combinations in German (3 genders $\times$ 2 numbers $\times$ 4 cases). Thus, even our restricted set of combinations, when populated with the appropriate values, would be computationally unfeasible. We have addressed this problem with a naïve approach. Since we cannot know in advance which combinations may be most distinctive for authors and texts, we

have selected among them based on frequency alone. For each combination length, only those type–value pairs which occur in approximately 5% of the tokens in the corpus have been included as input variables for classification. The process of populating feature types with their values is computationally slow for combinations of more than two elements, so we have used a smaller sample corpus for each language. Thus, the 5% cut-off is an approximation. A separate set of variables has been identified in this way for each language. Because UDPipe produces different types of morphological annotation, and because syntactic annotation, although it largely consists of the same relationship labels, has different frequency distributions in various languages, the same selection procedure with the same 5% cut-off results in a different quantity of features for each language. Details are given in Table 5.

Table 5. Number of input features by number of type–value components in each feature.

	Unary	Binary	Ternary	Total
English	55	231	367	653
French	59	337	629	1025
German	63	325	673	1061
Polish	65	337	735	1137

The nature of these features may be difficult for the reader to visualize from a description alone. The examples given below in Section 5 should provide illustration.

3. Classification

The purpose of this study is to examine the effectiveness of morphosyntactic input features as stylometric markers of literary texts. In particular, we test the usefulness of the annotation produced by the UDPipe applications. As noted above, classification experiments are a standard means to evaluate the worth of different sets of input features. It is to be expected that the various steps implemented by UDPipe involve a greater or lesser degree of error. Since the information/noise ratio worsens for shorter input texts [19], the first round of classification tests will be performed using a range of shorter "texts" sampled from our corpora. The sample sizes are 2000, 1000, and 500 words. For the purpose of sampling, each text was treated as a "bag of words". Each token was—naively—treated as independent of all others; no further account was taken of the context of an individual token in sentence, paragraph or any other unit of composition.

Recent years have seen the rapid development of many sophisticated classification algorithms. Deep learning approaches are appearing frequently in stylometric studies (for example, [20,21]). However, in spite of the accuracy achieved by some of these approaches, they are often uninterpretable; it is unclear exactly how the algorithm arrived at a particular classification, or even just what elements of a text were considered [22]. This is not a satisfactory outcome for stylometrics in a literary or historiographical context. In such fields, understanding and explaining the style of a text or author is often the principal goal.

In an effort to combine good accuracy and a high level of interpretability, we have chosen logistic regression as the approach for this study. Logistic regression has long been used extensively in many fields and is well understood. It is a straightforward approach to identify the contribution of each input feature to the predictions produced by this method. An additional advantage is that logistic regression is able to function well in the presence of many co-linear variables. Morphosyntactic data are by nature highly inter-dependent, and this may present a problem for some approaches. In this study, regression was implemented through the LiblineaR package for the R Project for Statistical Computing [23,24]. This package offers a range of linear methods; we selected the L-2 regularization option for logistic regression.

The first experiment was designed to discover if morphosyntactic features could distinguish among the individual novels in the corpora. For each input sample text size

in each language, 80% of the data were used for training the classifier and the remaining 20% were set aside for testing. Inclusion of a segment in the training or testing set was random. For example, to test 2000 word samples, each 20,000-word text in a corpus was split randomly into ten samples, eight of which were used for training and two were set aside for testing. Each training step of the classifier was therefore based on 360 samples (8 per novel for 45 novels). The procedure for other sample sizes was analogous. To validate the results of the classification testing, we used Monte Carlo sub-sampling [25] applied at two levels. As a rule, the populating of the segments with randomly selected tokens was carried out ten times. For each of these partitionings to create text segments, 50 additional random partitionings into a training set and a test set were made.

4. Results

We would expect the stylometric "signature" of individual novels to be very strong. This expectation is based on the (over-)simplifying assumption that a single literary work has a unitary style, arising from a shared theme, time of composition, etc). It should therefore not be surprising that morphosyntactic attribution with separate classes for each novel is quite successful. The results are given in Table 6, which gives the mean accuracy rate (correct "guesses"/total "guesses") for the 500 iterations in the top row of each cell, with the accuracy range reported below the mean. There are 45 classes in the data set for each language. All classifications were multi-class (one-versus-rest approach).

Table 6. Results of classification by individual novel (45 classes).

	500-Word Samples	**1000-Word Samples**	**2000-Word Samples**
English	90.6% (90.1–91.3%)	97.1% (96.4–98.1%)	99.4% (98.6–99.9%)
French	93.8% (93.1–94.6%)	96.9% (94.4–98.9%)	98.9% (97.7–100%)
German	96.3% (95.8–96.6%)	99.1% (98.8–99.3%)	99.8% (99.2–100%)
Polish	98.3% (96.8–98.9%)	99.5% (98.3–100%)	100% (100–100%)

Clearly, the works in each corpus are sharply distinguishable at the morphosyntactic level. Unfortunately, there is little published research to which these results may usefully be compared. Generally, recent stylometric research has a quasi-forensic tendency, focused on the ability to "prove" authorship of particular texts. In such cases, there is no reason to examine the discriminability of the individual works of an author. In contrast, our interest is in the *descriptive* value of stylometric measures as applied to works as well as authors. Our assumption in this study is that input features that both discriminate texts clearly and are understandable in terms of traditional stylistics may serve as the basis of valuable stylometric descriptions. Our results indicate that discriminability is high even with the relatively small 500-word samples; this success can be taken as an indication that a good deal of stylistic information is in fact conveyed by the features that we have proposed. We will examine some of the most important of these distinguishing features in the next section. It is worth mentioning here that the same procedure (albeit with different input features for each corpus) works quite well for each language tested. In fact, it is apparent from the 500-word samples that the morphosyntactic signal is somewhat stronger the more morphologically complex the language is. This complexity is reflected in the number of features as reported in Table 5: a sharper distinction seems to exist between works in Polish, which has 1137 total input features, compared to between works in English (653 total features).

5. Interpretability

In this section, we explore how the morphosyntactic input features presented here can be interpreted in a relatively straightforward manner using traditional grammatical terms. This advantage is not associated with more popular inputs such as character n-grams, which can achieve high accuracy in a classification test but do not lend themselves to clear interpretation.

In order to better illustrate the interpretability of our feature set, we will carry out our discussion against the background of an important open problem in stylometrics. This problem concerns the relationship between a stable authorial "signature" and the variability that all authors can be expected to display among their individual works. We have seen above (Table 6) that each novel in our four corpora has its own strong stylometric "signal" that allows it to be uniquely identified. Thus, all 2000-word samples in our classification were assigned to the correct work with an aggregated mean accuracy greater than 98%.

Matters are different if, instead of isolating the morphosyntactic characteristics that distinguish particular novels from each other, we try to abstract from the particular works the more general "style" of each author. Table 7 presents the results of such an experiment. Once again, data were randomly partitioned into samples of 2000, 1000, and 500 words. This time, however, classification followed the standard "leave-one-out" method. For each training iteration of the logistic regression classifier, all samples from one novel were withheld from the training data. The classes for attribution were the 15 authors in each corpus; the target class was modeled on the basis of the two novels by the relevant author remaining in the training set. In Table 7, the mean accuracy rate (correct "guesses"/total "guesses") for the 450 classification attempts is given in the top row of each cell, and the accuracy range is reported below the mean (the data for each novel were partitioned ten times into 2000-word samples; from each partitioning, 45 leave-one-out models were trained and the held-out set of samples was classified). There are 15 classes in the data set.

Table 7. Results of leave-one-out classification by author (15 classes).

	500-Word Samples	1000-Word Samples	2000-Word Samples
English	51.0% (49.6–53.5%)	56.9% (55.6–58.2%)	62.9% (60.2–64.8%)
French	54.0% (51.8–55.3%)	55.1% (54.0–56.1%)	57.1% (55.1–59.3%)
German	61.2% (59.9–63.2%)	63.0% (60.3–65.1%)	62.9% (59.7–64.0%)
Polish	44.8% (43.4–45.8%)	42.8% (42.4–44.7%)	41.5% (40.2–43.7%)

The sharp decrease in classification accuracy is striking. Presumably, an explanation is to be found in the greatly increased difficulty of the problem. The results of the most closely comparable previous studies point to the same conclusion. Maciej Eder has published three important studies on authorship attribution [19,26,27] in which the corpora are similar to our own. The accuracy of Eder's experiments is consistent with our results. For example, Eder (2010) classifies samples of various sizes drawn from 63 English novels; for samples of around 1000 words, accuracy falls between 40% and 50%. A more precise comparison is unfortunately not possible. All three of Eder's works present their results in graphs rather than tables. Thus, only rough estimates for the accuracy of a given sample size are possible. Most of Eder's data are based on the most frequent words. For a corpus of 66 German novels, samples ranging from 500 to 2000 words seem to yield accuracy scores from 30% to 60%. Evidently, the low accuracy of our authorship attribution tests (as compared to novel-by-novel classification) is not anomalous. Furthermore, it does not seem likely that the combination of input features and classifier that was quite good at identifying individual novels would become uninformative about the authorship of those same works. The field

of stylometry, at least as it pertains to the realm of literary writing, relies on the assumption that each author displays a number of linguistic "peculiarities" which remain stable for some significant length of time. Of course, this more or less stable authorial "signature" only exists as it is manifested in their individual writings, and these writings naturally vary in a host of ways. While we assume that the morphosyntactic dimension of authorial style is less affected by the "local" variation among texts than other aspects of style may be, true independence is of course impossible. It is essentially inconceivable that the author's personal linguistic "signature" could be completely separable from and unaffected by certain "external" factors—a novel's plot, setting or characters, for example—when it is only in the treatment of these factors that style comes into existence.

A comparison of Tables 6 and 7 show that, to speak loosely, the authorial signal as reflected by the morphosyntactic input features is only about half as strong as that produced by the combination of author and the "local" characteristics of the individual novel. This difference in results is the context against which we will examine the interpretability of our feature set. In particular, we will choose a single text, *Oliver Twist*, an 1839 work by Charles Dickens, on which to focus our discussion. We will briefly examine input features that allow this novel to be distinguished from the other 44 works in the corpus. Then, we will do the same for features that group *Oliver Twist* together with the other two Dickens works in the corpus while at the same time distinguishing "Dickens" as a class separate from the other 14 authors represented.

There are several simple ways to identify input features that are highly discriminative with a "one-layer" classifier such as logistic regression. For example, we could select those features to which the classifier assigned the most extreme weights (positive or negative). Similarly, we could guide selection by looking at the product of the model weight and the frequency of the feature, since it is on this basis that the algorithm assigns the probability for any class. However, to avoid complicating this discussion, we will limit our focus to the frequency of occurrence of the features. In particular, we examine the standardized value of input frequencies and choose those with the largest z-scores. Because morphosyntactic values are naturally interdependent, each input feature selected represents a group of correlated grammatical phenomena. For example, in many languages, only verbs are considered to have tense. Therefore, if a word is annotated as "tense = past", the part-of-speech annotation is redundant. Table 8 presents the selection of features "preferred" by *Oliver Twist*, as compared to the remainder of the corpus.

Table 8. Selection of input features "preferred" in *Oliver Twist*.

#	Feature	Frequency, *Oliver*	Frequency (Mean of Corpus)	Frequency Rank	Z-Score
1A	Number is singular, parent precedes	0.176	0.148	1	2.98
2A	Parent's own parent follows	0.124	0.104	1	2.59
3A	Parent is singular, parent's DD = 2	0.071	0.061	3	2.04
4A	Article, parent is singular noun	0.079	0.064	5	1.72
5A	Parent's dependency label is "object"	0.065	0.058	3	1.43

A few examples will help to illustrate the phenomena underlying these values. The first feature is grammatically transparent. This sentence from *Oliver Twist* has two examples: "They talked of **hope** and **comfort**". The two bold-faced nouns are annotated with feature #1; obviously singular, they are preceded by their dependency parent, *talked*. Although this dependency—a noun upon a verb—is the most frequent structure annotated with feature

#1 (a rate of 0.403 in *Oliver Twist*), a noun dependent on a preceding noun is also common (0.195), as in the following: "... an unwonted allowance of **beer**...". Here, the singular *beer* is dependent on *allowance.* One should also be aware that on rare occasions (0.018), the word annotated with feature #1 is itself a verb instead of the more usual noun (0.871) or pronoun (0.107): "I never knew how bad she **was**...". In this sequence, *was* is considered the head word of the indirect question clause *how bad she was;* UD grammar considers that the clause is dependent on the verb *knew*, which precedes it.

When interpreting a feature such as #1, which reflects more than a single grammatical value, it is a good idea to establish the relative contribution made by the components to the combined frequency. For example, both elements of feature #1 are more frequent in *Oliver Twist* than in the remainder of the corpus. Words annotated with *singular*: *OT* = 0.3171 and corpus = 0.308; words annotated with *parent precedes*: *OT* = 0.3394 and corpus = 0.3158. At the same time, the frequency of the combination of the two components (*OT* 0.17) is much higher than would be expected based on the parts (0.107). Thus, in a study of the style of the Dickens work, both feature #1 and its parts would be worthy of further analysis.

Feature #2 reflects a deep level of sentence structure since a word's annotation is based on its dependency parent and "grandparent". For example, in the sentence "If he **could** have known that... perhaps he would have cried the louder", *known*, the head word of the conditional clause, is the parent of the annotated word *could; cried*, the main verb of the sentence, is the parent of *known.* Since *cried* follows *known*, feature #2 is appropriately applied to *could*.

Feature #3, based simply on the dependency distance of the parent of the annotated word, should need no illustration. It is worth noting that, as for feature #1, both components of feature #3 are elements "preferred" by *Oliver Twist.* The frequency of words annotated with "parent is singular" is *OT* = 0.401 and corpus = 0.376; for "DD of parent is 2", *OT* = 0.134 and corpus = 0.129. Based on the individual frequencies, the expected rate of occurrence for the combination is 0.0537 for *Oliver Twist*. The actual rate is almost one-third larger.

Feature #4 is simple but is an example of the importance that very elementary syntactic structures can have in drawing stylometric distinctions. Essentially, this feature indicates the number of singular nouns that are modified by an article, either definite (*the)* or indefinite (*a/an*). Again, it is informative to analyze a compound feature according to its components. In this instance, *Oliver Twist* displays a preference for singular nouns (*OT*: 0.1534; corpus: 0.1404), but much of the distinctive force of feature #4 cannot be explained in terms of the frequencies of singular nouns. When feature #4 is controlled for the number of such words, the ratio of articles per singular noun is *OT* = 0.5182 and corpus = 0.4558. Clearly, a high frequency of articles is a stylistic characteristic of *Oliver Twist*.

Feature #5 is grammatically more complex. To interpret it, one must be familiar with two important aspects of dependency grammar: (1) verb valency and (2) functional syntax. According to dependency grammar in general, sentences are structures "built" according to the "requirements" of its component verbs. The primary requirement is the valency of a verb. Simply, "valency" in the appropriate sense is the number of dependents that are necessary to make a verb syntactically and semantically "correct". Such necessary components are called "arguments" of a given verb. For example, consideration of the sentence "Caesar died in Rome" shows that *die* here has a valency of one. If we subtract the dependencies, we produce "*died in Rome" and "Caesar died". Only the second is acceptable and indicates that *died* has a valency of one. A bivalent verb can be seen in "Brutus killed Caesar in Rome". *Kill* requires both *Brutus* and *Caesar*, but not *in Rome.* The argument of a monovalent verb is called the verb's *subject*; for bivalent verbs such as *kill*, one dependency is labeled as the verb's *subject*, the other as its *object*.

"Functional syntax" refers to the theory according to which dependencies are labeled primarily according to the role that they play with respect to their parent word. Less importance is given to the internal characteristics of the dependency. Consider the sentence "She put money in the bank". The verb *put* is trivalent since it requires a subject (*she*),

an object (*money*), and a third expression (*in the bank*) indicating a place/goal. This third expression in the case of trivalent verbs is also often called an object (or second object). The primacy of function is evident here in that the fact that *in the bank* is a prepositional phrase does not affect its dependency label. Its internal structure is irrelevant. Compare the sentence "She put money there". The dependency relationships in this version would be the same as in the first example. Although *there* in the second sentence is an adverb, it, like *in the bank*, is correctly annotated as object. The two expressions serve the same function with respect to *put*, and therefore receive the same relationship annotation.

The functional emphasis shown by dependency grammar reduces the number of dependency relation labels and, at the same time, groups a range of internally different phenomena in the same category. A few examples may help to clarify the phenomena that are reflected by feature #5. Under the label "object", UDPipe includes primarily nouns and adjectives. An example of a noun object is "Give it a **little** gruel. . .", where *little* is the annotated word and *gruel* its parent; *gruel* is the second argument (here a direct object) of *give.* One of Dickens's most famous sentences provides an example of an adjective in the *object* function: "'Please, sir', replied Oliver, 'I want **some** more'". Here, *more* is the direct object of *want*, and *some* (the annotated word) is a dependent modifier of the adjective *more*.

The word characterized by feature #5, in distinction to its *object* dependency parent, can also represent varying grammatical phenomena. To take only verbs, we find examples like "I shall take a early opportunity of **mentioning** it . . .". The annotated *mentioning* is a verb in its gerund form and is dependent on *opportunity*, the direct object of *take*. A different phenomenon is represented by "Bumble wiped from his forehead the perspiration which his walk had **engendered**. . .". The annotated *engendered* is the verb of the relative clause describing (and therefore dependent on) *perspiration*. UD grammar considers the verb the head of a relative clause, and therefore, *engendered* is the direct dependent of *perspiration*, which in turn is the direct object of the main verb *wiped.* Yet another difference is apparent in "The boy had no friends to **care** for. . .", where the annotated *care* is part of an explanatory infinitive structure which specifies the meaning of *friends*, the direct object of *had*.

This brief discussion of the dependency relationship *object* should make clear that the dependency labels are the most complicated annotation in the morphosyntactic data set created by UDPipe. However, since their complexity arises primarily from the grouping together of different grammatical "types" according to their grammatical "function", the interpretation of the relevant input features is time-consuming rather than conceptually difficult.

In addition to input features that are strongly "preferred" in *Oliver Twist*, there are others that are sharply "avoided". We will look only at three of the most important, as given in Table 9.

Table 9. Input features "avoided" in *Oliver Twist*.

#	Feature	Frequency, *Oliver*	Frequency (Mean of Corpus)	Frequency Rank	Z-Score
1B	Parent is an infinitive verb	0.076	0.10	43	−1.71
2B	Personal pronoun	0.085	0.114	43	−1.66
3B	Plural	0.041	0.051	42	−1.22

Feature #1B represents dependencies of the infinitive form of the verb. The English infinitive is morphologically the same as the dictionary lemma. It primarily occurs in one of two configurations. An infinitive can be "introduced" by the particle *to*, as in the following examples: "I have come out myself **to** take **him there**"; and ". . . the parish would like **him to** learn a right pleasant **trade** . . .". It is apparent that *to* plus the infinitive has a wide range of syntactic functions. In the first example, *to take* expresses the purpose for which the action of the main verb was undertaken. The infinitive phrase can be deleted from the sentence without making it ungrammatical. In contrast, the syntax (and semantics) of *like* in

the second example requires an object; *to learn* performs that necessary function and cannot be omitted without producing incorrect grammar. Note that in both example sentences, the infinitives have three words directly dependent on them (marked in bold). All of these words, then, would be correctly coded with feature 1B.

The second common configuration for the English infinitive is to be "introduced" by certain modal and auxiliary verbs. Examples occur in the following passages: "**Do I** understand that he **asked** for more. . .?"; ". . . [I]t **may** be as you **say**. . .". In the first example, the infinitive is *understand* which forms a verb phrase with the auxiliary *do.* In the second, the infinitive is *be*, a usage sometimes called complementary since the infinitive is necessary to complete the structure implied by a modal (here *may*). The bold-faced words in each example again indicate direct dependents of the infinitives, as required for feature 1B.

From these examples and the brief discussion, it should be clear that English infinitives in their most frequent structures *cannot* appear without at least one direct dependent. According to the rules of UD, the particle *to* is considered a dependent of the infinitive it precedes. Likewise, auxiliaries and modals such as those in our second set of examples are annotated as immediate dependents of the infinitives with which they are associated. It will not be surprising, then, to learn that, on average, each infinitive has more direct dependents (*OT*: 2.516; corpus: 2.795) than finite verbs (*OT*: 1.752; corpus: 1.628). At the same time, these numbers indicate a sharp difference between *Oliver Twist* and the rest of the corpus with respect to the complexity of infinitive clauses. The increase in the average number of dependencies from finite verbs to infinitives is much smaller than we might expect, given that infinitives "automatically" come with at least one dependent word: dependency per word increases by 1.167 for the corpus, but just by 0.764 for *Oliver Twist.* Thus, measured by the number of dependencies, infinitive structures in *OT* are less complex than we would expect based on finite structures in the same novel.

The grammatical categories reflected in features 2B and 3B are self-evident and require no examples. We only note that the relative avoidance of personal pronouns (*I*, *you*, *she*, *he*, *it*, etc.) in *Oliver Twist* is no doubt associated with the same novel's relative preference for common and proper nouns: *OT* = 0.245 and corpus = 0.218. Relevant for the interpretation of feature 3B is the fact that the frequency of all words annotated with grammatical number—nouns, pronouns, verbs and a few determiners (*this*/*these*, etc.)—is lower for *Oliver Twist* than for the remainder of the corpus (*OT* = 0.358; corpus = 0.364). This difference, however, only partly explains *OT's* relative avoidance of feature 3B. In fact, the distribution of grammatical number within this subset of relevant parts of speech leans strongly toward the singular as compared to the rest of the corpus (*OT*: singular = 0.884 and plural = 0.115; corpus: singular = 0.857 and plural = 0.142).

As noted above, comparison of the classification results in Tables 6 and 7 reveals that for every corpus tested, the signal identifying each individual novel is much more discernable than the authorial signal. While truly understanding this phenomenon—the coexistence of "local" variability and "authorial" style—will no doubt require many years of intensive study, stylometrics at the morphosyntactic level can offer valuable data bearing on this issue.

A detailed discussion is beyond the limits of this paper, but a single straightforward example will serve as a useful illustration. The accuracy values given in Table 7 are averages that encompass a great deal of variation. The authorial signal for some writers in each corpus was very weak; other authors were comparatively quite easy to distinguish. To take the corpus of English language novels, the most distinguishable author, with the highest mean accuracy of classification, was E. M. Forster. Subsamples of Forster's novels were attributed to the author with an accuracy of about 85%. (Forster's works in the corpus are *Where Angels Fear to Tread* (1903), *A Room with a View* (1908) and *Howards End* (1910)). At the other extreme, accuracy for the works of Vernon Lee was generally less than 1%. (Vernon Lee was a *nom de plume* for the writer Violet Paget. The relevant works in the corpus are *The Countess of Albany* (1884), *Miss Brown* (1884) and *Penelope Brandling* (1903).) The authorial signal for Charles Dickens, whose *Oliver Twist* was the focus of our discussion of input

features, was squarely in the middle at about 50% (in addition to *Oliver Twist* (1839), the corpus also contains Dickens's *Bleak House* (1853) and *Great Expectations* (1861)).

As one might expect, the algorithm's success at discriminating authorial signals is to some degree correlated with stylometric consistency within the works of each author. In other words, authors whose works show greater variability in the values of the input features tend to be more difficult to correctly attribute an authorial signature to. Table 10 gives a distribution summary for the "intra-author" standard deviations of each morphosyntactic input feature.

Table 10. Summary of standard deviations of input features for selected authors.

	Min.	1st Quart.	Median	Mean	3rd Quart.	Max.
Forster	0.00005	0.0014	0.0027	0.0037	0.0050	0.0174
Dickens	0.00027	0.0040	0.0066	0.0072	0.0099	0.0248
Lee	0.00037	0.0088	0.0130	0.0149	0.0185	0.0733

Each value in Table 10 is based on the standard deviation of the three works of each author for each of the 653 input features used in the English corpus. It is evident that the works of Lee are much less consistent with each other than the works of authors with better classification results. Lee's weak signal is not unexpected, given that the mean standard deviation in her works is more than four times larger than Forster's, while Lee's median value approaches five times Forster's!

That inconsistency within a class is associated with a high noise-to-information ratio and with difficulty in discrimination, which will surprise few people familiar with classification experiments. On the other hand, extensive use of morphosyntactic annotation is rare in stylometric studies, and the data reported in Table 10 suggest that such information would indeed be useful in exploring stylometric variation for an individual author. In addition, because it preserves a relatively high amount of information even in short texts (see the 500-word samples in Tables 6 and 7), a morphosyntactic approach may even be effective in describing stylometric variability within a single work or a single chapter of a work.

We will conclude our investigation into morphosyntactic stylometry by returning to the work of Charles Dickens. Taken together, the works of Dickens in our corpus produce, as mentioned above, a moderate authorial signal. A few of the important input features through which the logistic regression model distinguishes "Dickens" from the other authors in the corpus are given in Table 11.

Table 11. Selected input features preferred or avoided by the class "Dickens".

#	Feature	Mean Frequency of Dickens's Works in Corpus	Mean Frequency of Remaining Corpus	Intra-Author Standard Deviation	Z-Score (Dickens's Works in Corpus)
1C	Parent precedes	0.336	0.314	0.0047	1.66
2C	Parent is a verb with DD > 2	0.136	0.107	0.0091	1.58
3C	Parent is a verb and head of an adverbial clause	0.069	0.055	0.0065	1.41
4C	Adjective	0.064	0.071	0.0056	−0.93
5C	Parent is sentence root	0.183	0.218	0.0082	−0.98

In Table 11, the intra-author standard deviation indicates that for each selected feature, the frequencies in the three works of Dickens in the corpus are quite close to each other. With regard to the z-scores, a distinction is noticeable between these and the scores in

Tables 6 and 7. The values in Table 11 indicate a selection of features with smaller differences between the class of interest (here "Dickens") and the corpus mean. This is not an accident of selection, since relatively large z-scores are less frequent for the mean of Dickens's three relevant works than for *Oliver Twist*. For example, for *OT*, 52 input features had a z-score with a magnitude greater than 1.5 (positive or negative). The corresponding count for the three-work mean is 9.

To turn to the details of the input features, feature 3C needs no further elaboration; any word that comes after its dependency parent in the linear order of the sentence is encoded with this feature. Feature 2C should likewise be self-explanatory at this point. The frequency of feature 4C is based on a simple count of parts of speech: a relative avoidance of adjectives is a shared characteristic of the three Dickens texts.

Features 3C and 5C contain types of dependency relationships and therefore require some explanation. In the UD annotation scheme, a sharp distinction is made between words that function as arguments to a verb (see above) and words that do not. Words that are not arguments are optional in the sense that they can be omitted without rendering the sentence ungrammatical (or nonsensical). *Adverbial* is the label for the most important class of "optional" words. If the word with this function is a verb, it is labeled as an *adverbial clause.* Two examples will point to the many possible ways that a verb can function as an adverbial for another clause: (1) "If **he could have** known that he was an **orphan**, . . . perhaps he would have cried the louder"; (2) "But he hadn't, **because nobody had** taught **him**". In sentence 1, *known* is the head verb of the conditional clause that is subordinate to the main verb *cried.* In sentence 2, *taught* is the head of a causal clause, itself dependent on *hadn't.* In both sentences, words annotated with "parent is head of an adverbial clause" are highlighted in bold.

The last of our exemplary "Dickensian" input features, feature 5C, indicates the frequency of sentence main clauses. Generally, the root of a sentence is a label given to the main verb, but a peculiarity of UD in this regard may be illustrated by the following example: "**Boys is wery** obstinit. . .". In equational structures such as this one, where the subject is "linked" with a predicate nominal (here *obstinit*) by a copula verb (*be* and similar verbs), UD grammar considers the predicate nominal, and not the verb, to be the head of the clause. Thus, the bold-faced words in the example are annotated as dependents of the adjective. As a result of this protocol, a not insignificant portion of sentence *roots* in the UD scheme are nouns, pronouns and adjectives.

The final step in our discussion of the "Dickensian" authorial signal is to give a very few examples of input features that weaken that signal. In particular, this is a selection of features for which the values are relatively diverse across the three Dickens novels in the corpus. Details are given in Table 12.

Table 12. Selected input features where frequency variability weakens the "Dickens" signal. *BH* = *Bleak House*, *GE* = *Great Expectations* and *OT* = *Oliver Twist*.

#	Feature	Frequency (Dickens)	Frequency (Remainder of Corpus)	Intra-Author s.d.	Frequency Rank
1D	Personal pronoun	*BH*: 0.117 *GE*: 0.134 *OT*: 0.085	0.113	0.020	*BH*: 20 *GE*: 7 *OT*: 43
2D	Parent is singular	*BH*: 0.384 *GE*: 0.355 *OT*: 0.401	0.376	0.0188	*BH*: 20 *GE*: 35 *OT*: 8
3D	Parent is past indicative verb	*BH*: 0.137 *GE*: 0.182 *OT*: 0.150	0.136	0.0188	*BH*: 24 *GE*: 3 *OT*: 13
4D	Parent is verb	*BH*: 0.414 *GE*: 0.449 *OT*: 0.409	0.403	0.0179	*BH*: 19 *GE*: 3 *OT*: 6

The morphosyntactic phenomena underlying these features should by now be clear and examples unnecessary. In fact, we have already looked closely at feature 1D, which appeared as feature 2B in Table 9. There, the avoidance of personal pronouns was identified as a distinguishing characteristic of *Oliver Twist*. Here, we see that this characteristic is not shared by the other two relevant works, in each of which personal pronouns are more frequent than in the corpus mean. Thus, unsurprisingly, the same morphosyntactic feature can be informative at one level of classification (e.g., novel by novel) and simultaneously increase noise at another (e.g., author by author).

The features in Table 12 disrupt the authorial signal for "Dickens" because of how much the frequencies vary among the three Dickens novels in the corpus. Comparison of the "author-internal" standard deviations in Tables 11 and 12 shows that the dispersal of the features in Table 12 is 1.9 to 4.2 times greater than in Table 11. Works with such a range of frequencies for any given input feature will hinder the detection of an authorial signal. At the same time, it is important to realize that a tight "grouping" of frequencies for a feature is not in itself enough to make that feature informative for classification. For example, there are many morphosyntactic input features in our set for which the "Dickensian" standard deviation is quite small but whose frequencies are very close to the corpus mean. Two such features are "part-of-speech is verb" and "part-of-speech is a preposition and parent is a noun". For the first of these, the mean frequency for the three Dickens novels is 0.134 (sd = 0.0043) and the corpus mean is 0.130. For the second, a feature that essentially reflects the number of prepositional phrases in the texts, the frequency for Dickens's works is 0.079 (sd = 0.0018), while the corpus mean is 0.078. Features with frequencies in this pattern are generally not valuable for logistic regression. From a stylometric or stylistic point of view, however, it may be just as interesting to know where Dickens's morphosyntactic characteristics adhere to the norm as where they depart sharply from it.

6. Conclusions

This paper has presented arguments for the potential value of morphosyntactic annotation for stylometric analysis. It has demonstrated that UDPipe parsers currently available for many languages produce annotations whose inevitable errors do not seriously undermine the stylometric usefulness of this information, judged by accuracy in classification experiments. Based on the assumption that morphosyntactic characteristics are not closely dependent on the specific subject matter of the target texts, the input features described in this study are to a significant degree topic agnostic. Further, this work has explored the advantage offered by morphosyntax in terms of stylometric interpretability. The input features used here are, for the most part, made up of grammatical concepts likely to be familiar to anyone seriously investigating the style of a literary work or author. Admittedly, the concepts underlying dependency grammar may be new to many investigators, but the syntax of natural languages is itself a complex structure. Dependency grammar, assuming no "hidden" structures, reflects this complexity in a fairly straightforward way. In view of the demonstrated advantages of morphosyntactic information, it seems clear that it should have a larger role in stylometric scholarship.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Mosteller, F.; Wallace, D.L. Inference in an Authorship Problem. *J. Am. Stat. Assoc.* **1963**, *58*, 275–309. [CrossRef]

2. Koppel, M.; Schler, J.; Bonchek-Dokow, E. Measuring Differentiability: Unmasking Pseudonymous Authors. *J. Mach. Learn. Res.* **2007**, *8*, 1261–1276.
3. Coyotl-Morales, R.M.; Villaseñor-Pineda, L.; Montes-y-Gómez, M.; Rosso, P. Authorship Attribution Using Word Sequences. In *Progress in Pattern Recognition, Image Analysis and Applications*; Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 844–853.
4. Rochon, E.; Saffran, E.M.; Berndt, R.S.; Schwartz, M.F. Quantitative Analysis of Aphasic Sentence Production: Further Development and New Data. *Brain Lang.* **2000**, *72*, 193–218. [CrossRef]
5. Kestemont, M. Function Words in Authorship Attribution from Black Magic to Theory. In Proceedings of the 3rd Workshop on Computational Linguistics for Literature, Gothenburg, Sweden, 27 April 2014; pp. 59–66.
6. Koppel, M.; Schler, J.; Argamon, S. Computational Methods in Authorship Attribution. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 9–26. [CrossRef]
7. Ranaldi, L.; Pucci, G. Knowing Knowledge: Epistemological Study of Knowledge in Transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
8. Revesz, P.Z. A vowel harmony testing algorithm to aid in ancient script decipherment. In Proceedings of the 24th International Conf. on Circuits, Systems, Communications and Computers, Chania, Greece, 19–22 July 2020; IEEE Press: New York, NY, USA, 2020; pp. 35–38.
9. VanOrsdale, J.; Chauhan, J.; Potlapally, S.V.; Chanamolu, S.; Kasara, S.P.R.; Revesz, P.Z. Measuring vowel harmony within Hungarian, the Indus Valley Script language, Spanish and Turkish using ERGM. In Proceedings of the 26th International Database Engineered Application Symposium, Budapest, Hungary, 13 September 2022; pp. 171–174.
10. Burke, M. Stylistics: From classical rhetoric to cognitive neuroscience. In *The Routledge Handbook of Stylistics*; Burke, M., Ed.; Routledge handbooks in English Language Studies; Routledge, Taylor & Francis Group: London, UK; New York, NY, USA, 2014.
11. Eder, M.; Górski, R.L. Stylistic Fingerprints, POS-tags, and Inflected Languages: A Case Study in Polish. *J. Quant. Linguist.* **2022**, *30*, 86–103. [CrossRef]
12. Nivre, J.; De Marneffe, M.C.; Ginter, F.; Goldberg, Y.; Hajic, J.; Manning, C.D.; McDonald, R.; Petrov, S.; Pyysalo, S.; Silveira, N.; et al. Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 1659–1666.
13. De Marneffe, M.C.; Manning, C.D.; Nivre, J.; Zeman, D. Universal dependencies. *Comput. Linguist.* **2021**, *47*, 255–308. [CrossRef]
14. Wijffels, J. Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. 2019. Available online: https://CRAN.R-project.org/package=udpipe (accessed on 1 January 2024).
15. Lui, H. Dependency distance as a metric of language comprehension difficulty. *J. Cogn. Sci.* **2008**, *9*, 159–191.
16. Chen, R.; Deng, S.; Liu, H. Syntactic complexity of different test types: From the perspective of dependency distance both linearly and hierarchically. *J. Quant. Linguist.* **2021**, *29*, 510–540. [CrossRef]
17. Ferrer-i-Cancho, R.; Gómez-Rodríguez, C.; Esteban, J.L.; Alemany-Puig, L. Optimality of syntactic dependency systems. *Phys. Rev. E* **2022**, *105*, 014308. [CrossRef]
18. Chen, X.; Gerdes, K. Classifying languages by dependency structure: Typologies of delexicalized universal dependency treebanks. In Proceedings of the Fourth International Conference On Dependency Linguistics, Pisa, Italy, 18–20 September 2017; pp. 54–63.
19. Eder, M. Short Samples in Authorship Attribution: A New Approach. In Proceedings of the Digital Humanities 2017, Conference Abstracts, 8–11 August 2017; McGill University: Montreal, QC, Canada, 2017; pp. 221–224. Available online: https://dh2017.adho.org/abstracts/341/341.pdf (accessed on 1 January 2024).
20. Hay, J.; Doan, B.L.; Popineau, F.; Elhara, O.A. Representation Learning of Writing Style. In Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020), Online, 19 November 2020; Available online: https://aclanthology.org/2020.wnut-1.30 (accessed on 1 January 2024).
21. Wegmann, A.; Schraagen, M.; Nguyen, D. Same Author or Just Same Topic? Towards Content-Independent Style Representations. *arXiv* **2022**, arXiv:2204.04907.
22. Patel, A.; Rao, D.; Callison-Burch, C. Learning Interpretable Style Embeddings via Prompting LLMs. *arXiv* **2023**, arXiv:2305.12696.
23. Fan, R.E.; Chang, K.W.; Hsieh, C.J.; Wang, X.R.; Lin, C.J. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
24. Helleputte, T. LiblineaR: Linear Predictive Models Based On The Liblinear C/C++ Library. R Package Version 2.10-12. Available online: https://www.csie.ntu.edu.tw/~cjlin/liblinear/ (accessed on 1 January 2024).
25. Simon, R. Resampling strategies for model assessment and selection. In *Fundamentals of Data Mining in Genomics and Proteomics*; Dubitzky, W., Granzow, M., Berrar, D., Eds.; Springer: Boston, MA, USA, 2007; pp. 173–186.
26. Eder, M. Does size matter? Authorship attribution, short samples, big problem. In *Digital Humanities 2010: Conference Abstracts*; King's College London: London, UK, 2015; pp. 132–135.
27. Eder, M. Does size matter? Authorship attribution, small samples, big problem. *Digit. Scholarsh. Humanit.* **2015**, *30*, 167–182. [CrossRef]