DHQ: Digital Humanities Quarterly
2022
Volume 16 Number 2

Universal Dependencies and Author Attribution of Short Texts with Syntax Alone

Robert Gorman <rgorman1_at_unl_dot_edu>, Department of Classics and Religious Studies, University of Nebraska-Lincoln

Abstract

Improving methods of stylometrics and classification so that they give good results with small texts is the focus of much research in the digital humanities and in the NLP community more generally. Recent work [Gorman 2020] has suggested that an approach using combinations of shallow and deep morpho-syntactic information can be quite successful. But because the data in that study were taken from hand annotated dependency treebanks, the wider applicability of such an approach remains in question. The present paper seeks to answer this question by using machine-generated morphological and syntactic annotations as the basis for a closed-set classification experiment. Texts were parsed according to the Universal Dependency schema using the “udpipe” package for R. Experiments were carried out on data from several languages covering a range of morphological complexity. To limit confounders, consideration of vocabulary was excluded. Results were quite promising, and, not surprisingly, a more complex morphology correlates with better accuracy (e.g., 100-token texts in Polish: 88% correct; 100-token texts in English: 74%).The method presented here has particular advantages for stylometrics as practiced in literary analysis and other fields in the humanities. The Universal Dependency annotation categories are generally similar to those used in traditional grammars. Thus, the variables which serve to distinguish the style of a given author are relatively easier to interpret and understand than, for example, are character n-grams or function words. This fact, combined with the availability of easy-to-use dependency parsers, opens up the study of a syntax-centered stylometrics to persons with a wide range of expertise. Even students at the early stages of their studies can identify and investigate the morpho-syntactic “signature” of a particular author. Therefore, the characterization of texts based on computational annotation of this type deserves a place in classification studies because of its combination of good results and good interpretability.

1. Introduction

Developing better methods of classifying short texts is becoming increasingly important in the field of computational stylometrics. Synoptic work (Eder 2015) suggests that many common approaches to at least one kind of text classification — Authorship Attribution — are unreliable when the text to be classified is less than a few thousand words. Maciej Eder’s study examines several classifiers (Burrows’ Delta, Support Vector Machine, and k-Nearest Neighbor) and finds that all suffer from a similar drop in effectiveness as the size of the target texts decreases. The same result is seen when considering the most usual independent variables (a.k.a. “features”) used as input to the various classifiers. These features include Most Frequent Words, character n-grams, and POS (part of speech) n-grams. Based on Eder’s results, one must conclude that success in classification of short texts needs improved algorithms and/or more informative input variables. 1

This study focuses on the second requirement and explores the value of morphological tagging and syntactic parsing to create a richly discriminative set of variables. Generally speaking, morphological and syntactic data are underused in Authorship Attribution as compared to lexical and character-based features. While a wide-ranging survey [Stamatatos 2009] was able to refer to a number of studies that used syntactic re-write rules or more complex morpho-syntactic features, an examination of references in a subsequent overview [Swain et al. 2017] indicates POS tags are almost the only such features in frequent use. However, a more recent experiment in Authorship Attribution relying entirely on 2

morpho-syntactic variables [Gorman 2020] suggests that such features can significantly enrich the information available for text classification.

The general applicability of R. Gorman's study is questionable for at least two reasons. First, it involves a morphologically complex language — ancient Greek. It is uncertain how its methods may be suited to a simpler morphological target such as English. Second, the study's approach is based on a dependency syntax corpus in which morphology and dependency relations have been hand annotated by a single language expert. The corpus is unusually large to be annotated in this way, and the resulting accuracy and consistency is a luxury available in few languages. 3

The present investigation will further explore the viability of a morpho-syntactic approach to text classification. It will test the effectiveness of such variables in several languages representing a wide spectrum of morphological complexity (English, German, Spanish, Finnish, and Polish). In addition, the requisite morphological and dependency annotation will be generated automatically, using a freely available program. As a proof of concept, classification will be carried out with only the resultant morpho-syntactic information. This restriction will simplify interpretation of results by reducing possible confounding elements within the proposed feature set. In addition, an important consideration in authorship classification is to eliminate as far as possible the effects of topic and genre and to avoid relying on any features that could be easily imitated or manipulated. Syntactic features are thought to meet this requirement better than lexical ones. Since identifying syntactic information requires extra processing with possibly noisy results, most investigations fall back on "function words" as a proxy for truly syntactic features. By directly using computationally generated syntactic analyses instead, this study will explore the benefits of such pre-processing in comparison to the costs of the noise it inevitably introduces. 4

The remainder of this article has the following structure. Part 2 discusses the formation of the various language corpora. It then details the creation of the input variables. These variables are somewhat complicated, in that they are based on combinations of morpho-syntactic elements and, at the same time, seek to preserve as much flexibility as possible in the incorporation of these elements. Part 3 explains the several steps of the classification experiment itself. Part 4 presents the results of the classification process and explores their implications. It lays particular emphasis on the interpretability of the input variables used in this study. Arguably, descriptions of authorial style based on traditional morpho-syntactic categories will prove more persuasive to those outside the ranks of scholars of computational stylometrics. 5

2. Corpora and Feature Extraction

This study involves corpora of texts in several languages: English, Finnish, German, Spanish, and Polish. An attempt was made to gather publicly available texts in a wider range of languages, but it was hindered by several factors. The design of the proposed experiment requires 20 single-author texts in a given language. All texts should be more than 20,000 tokens in length, and all texts in a language should be of the same type. Where possible, texts were to be drawn from a well-defined temporal range. These demands were most easily met by selecting works of fiction, in particular novels. The works were drawn from Project Gutenberg and the Computational Stylistics Group.[1] 6

It may be surprising that a study interested in classification of short texts would use novels rather than, for example, tweets or news articles. However, the focus here is on the informational value of morpho-syntactic variables in various languages and not on any particular kind of text, and in this case the reasons for choosing novels are fairly compelling. As noted, the literature suggests that classification of more than a few texts becomes ineffective when the targets are shorter than a few thousand words. Thus, our investigation starts with texts c. 2000 words. Since data are traditionally split into a training set comprising approximately 90% of the material, with the remainder in the test set, texts of 20,000 words are in order. While it might be possible to manufacture texts of this length by concatenating tweets or the like, it would not be easy to find, for a range of languages, large sets of short texts that meet a second requirement: that single authorship of each 20,000-word text group can be assumed. Third, the use of novels minimizes, at least as compared to concatenated collections, possible confounding effects of genre and topic. 7

The core of this investigation is variable extraction and preparation. The first step is to generate morpho-syntactic 8

annotations for each text. The processing is done with the “udpipe” package for the R Software Environment [Wijffels 2019]. The package includes functions to produce a universal dependency grammar analysis for texts in a wide variety of languages. Dependency grammar is an increasingly widely accepted approach to describing the structure of sentences. It offers an advantage to the present study in that, unlike phrase structure grammars, it deals well with non-projective grammatical sentence components — essentially, words closely related grammatically that are separated from each other by less closely related words. Non-projectivity occurs frequently in many languages, especially those which, unlike English, have a relatively complex system of morphology. The Universal Dependencies framework [Nivre 2015] has been developed by an international cooperative of scholars to further cross-linguistic language study. One of its most important contributions is to establish a common set of tags for morphological features and syntactic relationships. This standard is a great step forward in natural language processing across languages, since, *inter alia,* the parsers (i.e., programs to analyze syntax) sponsored by the Universal Dependencies (UD) project produce mutually compatible output: a single algorithm can pre-process texts from a range of languages.

Raw text (.txt files) provided to the udpipe program gives an output in which each token is lemmatized, tagged for morphology, and parsed for universal dependency relationship. Formatted as a matrix, its most salient results look like this:

token_id	token	lemma	upos
1	Er	er	PRON
2	lachte	lachen	VERB
3	vor	vor	ADP
4	Vergnügen	Vergnügen	NOUN
5	,	,	PUNCT
6	sich	er|es|sie	PRON
7	über	über	ADP
8	den	der	DET
9	Katechismus	Katechismus	NOUN
10	mokieren	mokieren	VERB
11	zu	zu	PART
12	können	können	AUX

Table 1. udpipe Annotation.

token_id	feats	head_token_id	dep_rel
1	Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs	2	nsubj
2	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root
3	Case=Nom|Gender=Neut|Number=Plur	4	case
4	Case=Nom|Gender=Neut|Number=Sing	2	obl
5	NA	10	punct
6	Case=Acc|Number=Sing|Person=3|PronType=Prs|Reflex=Yes	10	obj
7	Foreign=Yes	9	case
8	Case=Acc|Definite=Def|Gender=Masc|Number=Sing|PronType=Art	9	det
9	Case=Acc|Gender=Masc|Number=Sing	10	obl
10	VerbForm=Inf	4	xcomp
11	Case=Nom|Definite=Ind|Gender=Neut|Number=Sing|PronType=Ind	10	mark
12	VerbForm=Inf	10	aux

Table 2. udpipe Annotation (continued)

The text is part of a sentence from *Buddenbrooks* by Thomas Mann: "*Er lachte vor Vergnügen, sich über den Katechismus mokieren zu können …*" ("He laughed with pleasure at being able to make fun of the catechism"). 10

For each token the analysis gives the form as it appears in the text and its lemma. This information is not used in our approach to classification and will be ignored here. The remaining columns shown, however, are integral. 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=VALUE1|TYPE2=VALUE2). For example, token #2, *lachte* ("laughed"), is identified as Indicative in the category Mood (Mood=Ind), Singular in the category Number (Number=Sing), etc. There is no limit to the number of features that may be assigned to a single token. 11

Part of speech and morphology constitute what we may call "shallow" syntactic features. Information of this type may allow us to infer some syntactical structures, but they do not represent them directly. In contrast, the "head_token_id" and "dep_rel" columns do constitute such a direct representation. The head token is the item that is the immediate syntactic "parent" of a given token. The dependency relation specifies the type of grammatical structure obtaining between parent and target. From these columns we may generate a description of the syntactic structure of the entire sentence in a representation of its "deep" syntax. The structure revealed by these data points is perhaps most clearly illustrated with the corresponding dependency tree: 12

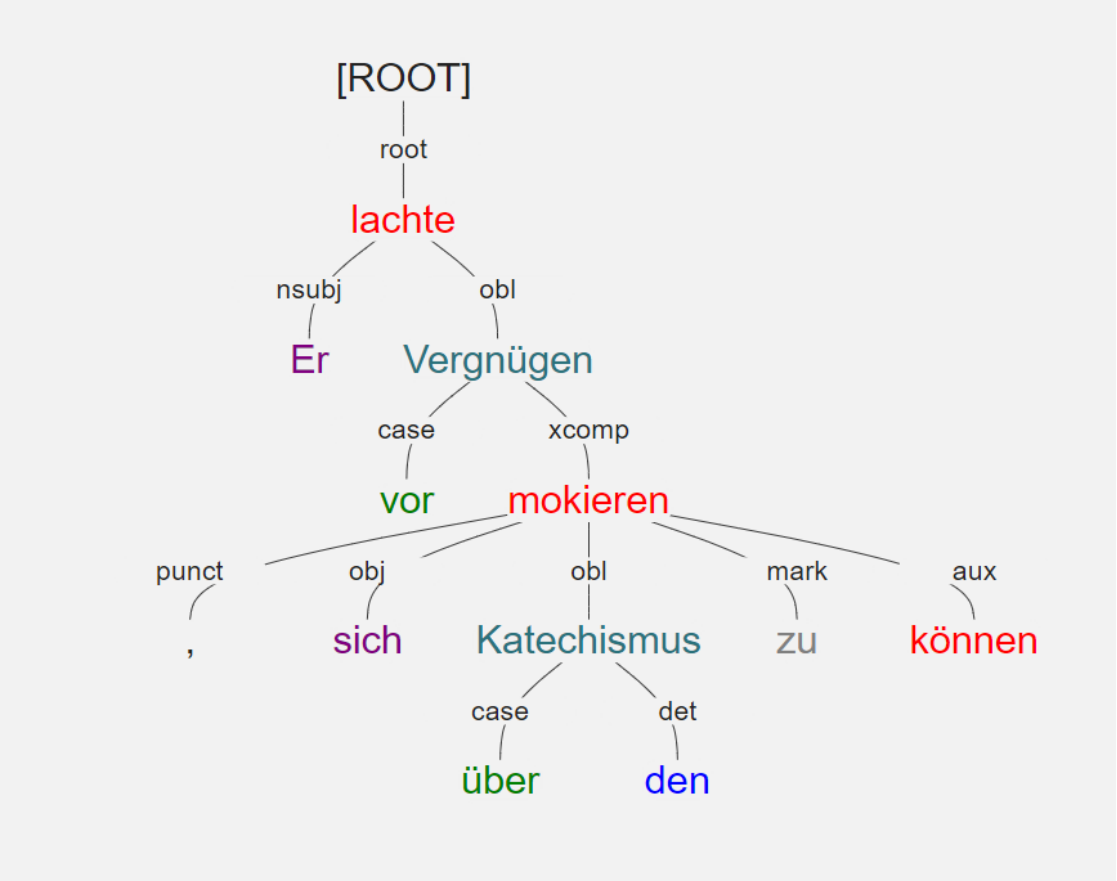

Figure 1. Dependency Tree of Example Sentence

As is apparent, the syntactic "path" from the sentence root to the each "leaf" token is given by the combination of head id and dependency relationship. 13

Both shallow and deep syntactic information are used to create the input variables for this investigation. Before looking at those variables in detail, it is worth noticing that pre-processing with the udpipe program introduces a certain amount of noise through mistaken analyses. To focus on the "feats" output, token 4, the noun *Vergnügen*, is incorrectly assigned NOMINATIVE case (should be DATIVE). In addition, the three indeclinable tokens in this text — i.e., words with a fixed and static form whose morphology is usually not further analyzed — are treated very strangely indeed: the preposition *vor* and the particle *zu* are assigned gender, number, and case (*zu* also gets classified for definiteness and pronoun type). The frequent preposition *über* is treated as a foreign word, perhaps as a dependent of the ecclesiastical (and ultimately ancient Greek) word *Katechismus*. Whatever the reason, the parser has introduced 4 errors in 12 tokens. Of course, the example text was chosen because it is short yet relatively complex; it may not be at all representative of the general error rate of the udpipe program. Nonetheless, it would be reasonable to worry that the noise introduced by erroneous morpho-syntactic annotations would undercut their value for classification input variables. However, as we will see below, bad tagging and parsing seems to make little or no practical difference in the classification results. We may evidently assume that these input errors are distributed randomly across the classified texts and that they have little effect as compared to the information carried by the input variables. 14

To turn now to the details of the features (i.e., input variables) themselves, there are two main principles guiding their creation: first, the set should include both shallow and deep syntactic information; second, the "degrees of freedom" present in the morpho-syntactic "system" of the target language should be, to the greatest extent possible, preserved and represented in the features.[2] 15

The first criterion is met by including, alongside the morphological information for each word, its dependency relation and, for each word that is not a sentence root, the morphology and dependency relation for the parent word of the target 16

word. To illustrate from the German sentence given above, the variables for the word *Katechismus* would include the following:

- Self: POS = noun, case = accusative, gender = masculine, number = singular, dependency = oblique
- Parent: POS = verb, verb form = infinitive, dependency = open clausal complement

17 Including values for parent as well as target word creates a representation of the hierarchical structure that most linguistic theories assume for syntax. Of course, the dimensions of morpho-syntactic information produced by the udpipe tagger-parser vary sharply by language. The following table (Table 4) indicates the number of “basic” elements for each language.

	target token	target + parent
English	16	32
Finnish	23	46
German	15	30
Polish	33	66
Spanish	18	36

Table 3. Number of “Simplex” Variables

18 The second requirement mentioned above is meant to preserve the complexity of the morpho-syntax of the input texts. As we have noted, it is rare for an attribution attempt to take any morphological feature into account with the common exception of part of speech. In those studies that do include morphological details, the various values that make up the relevant information about a word seem to be coded in a single block. For example, Paulo Varela et al. (2018, 2016) incorporate in their variables a vector for “flexion” (number, gender, tense, etc.), but apparently combine the features of a word into a single value: thus the entry for *é* (Portuguese “is”) is reported as “PR 3S IND VFIN,” i.e, present indicative 3rd-person singular verb.[3]

19 Combining the morphological categories of a word into a single value obscures some of the stylistic information that the word may contain. For example, if we categorize a German noun into a consolidated unit of gender, number, and case (e.g., Masc Sing Nom), the computer can assign each noun in the corpus to one of the 24 resulting categories (3 genders x 2 numbers x 4 cases) and calculate a frequency distribution over them, but it cannot access the frequencies of, say, masculine nouns or of combinations of two categories such as feminine dative, etc.[4] There is no strong reason to presume that any frequency ratio among the ternary variables is more discriminative for the texts in a corpus than, e.g., the ratio between Masc Nom and Fem Nom. We may usefully think of the morpho-syntactic elements that make up combinations as “characters” in “syntactic n-grams.”[5] Just as the addition to word-based inputs of character n-grams with different values of *n* may improve classification, in the same way using a range of combination lengths for morpho-syntactic elements may be valuable. Thus, it is best *in principle,* to include variables made of all “arities” of morpho-syntactic categories, from one to the total number of such categories (i.e., $\binom{n}{1}\dots\binom{n}{n}$).[6]

20 Of course, such an extensive variable set is not feasible *in practice.* For example, the udpipe program for German returns 13 categories of morphological features. To these 13 must be added the part of speech and the dependency relationship, and since the characteristics of the parent word are to be combined with those of each target word, variable types will be combinations constructed from 30 categories. If all sizes of combination were included (i.e., $\binom{30}{1}\dots\binom{30}{30}$), the number of types would be essentially 2^{30}—over a billion separate combinations! Nor is the combinatorial explosion the only problem. Each of the 2^{30}combinations is a type, not a variable. We have seen that the ternary combination Gender-Number-Case may take one of 24 different values in German: MASC SING NOM, etc. Thus, the number of type-value pairs reaches truly astronomical values. Clearly, we must combine the generation of input features with a rigorous culling and reduction process.

The first step deals with the combinatorial dimension. Selecting grammatically sensible combinations such as Gender-Number-Case (and its unigram and bigram components) would be laborious and would require knowledge of the various languages to be classified. Thus, a naïve approach was preferred: an arbitrary maximum length was chosen. Given a set of 30 elements for combination, the length must necessarily be short. $\sum_{k=1}^{4}\binom{30}{k}$ already gives 31,930 unique combinations. When we take into account that each combination will have multiple values, trouble with data sparsity and computational limitations can be expected. Therefore, the maximum was set at three elements per combination. This limit results in 4,525 combinations. 21

The next step populates each combination (or variable type) with its various values, as drawn from the basic morpho-syntactic elements produced by udpipe. This process is computationally slow for combinations of more than two elements, so we have used a smaller sample corpus for each language; it is composed of 500 tokens (punctuation excluded) from each text. When populated, the various combinations of types exhibit a large number of unique values. To take German as illustrative: the 30 unary types produce 161 values, the 435 binary types produce 5,579 values, and the 4,060 ternary combinations yield 56,349 values. As one would expect with linguistic data, the frequency distributions of these values are quite skewed. For example, 53,209 of the ternary values occur in 1% or fewer words and 19,116 occur only once in the sample corpus of 10,000 tokens. To avoid severe sparsity, another culling of variables is clearly in order. 22

Once again, since we do not know without prior investigation which combinations may be most distinctive for authors and texts, we have had recourse to a naïve approach. For each combination length, only those type-value pairs that occur in 5% of the tokens in the sample corpus have been included as input variables for classification. A separate set of variables has been identified in this way for each language. The size of each sub-set of variables is given in the following table (Table 3). 23

	unary	binary	ternary	total
English	45	61	85	191
Finnish	48	226	467	932
German	58	216	318	592
Polish	63	331	726	1,120
Spanish	46	150	189	385

Table 4. Number of Variable Subsets with at Least 5% Frequency

Because the variables may be difficult to conceptualize, the following tables give an illustration, including, for selected languages, the most frequent type-value pair as well as the least frequent pair to make the 5% minimum. Examples will be given in each category for target word characteristics (marked “t:”), parent word characteristics (marked “p:”), and combinations of target and parent (Tables 4, 5, and 6). 24

	type	value	frequency
English	t: Num	Sing	29.80%
	p: POS	VERB	44.90%
	t: Rel	root	5.15%
	p: Rel	ccomp	5.56%
Polish	t: Num	Sing	48.20%
	p: Num	Sing	65.20%
	t: Rel	cc	5.20%
	p: Case	Loc	5.40%

Table 5. Examples of Simplex Variables

	type	value	frequency
English	t: POS & t: Num	NOUN/Sing	13.80%
	p: POS & p: Num	NOUN/Sing	28%
	t: POS & t: VerbForm	AUX/Fin	5.50%
	p: Num & p: Rel	Sing/nmod	5.20%
	t: Num & p: POS	Sing/VERB	16.30%
	t: POS & p: Rel	VERB/root	5%
Polish	t: Num & t: Gender	Masc/Sing	21.20%
	p: POS & p: Voice	VERB/Act	46.70%
	t: POS & t: Animacy	Verb/Hum	5%
	p: Gender & p: Aspect	Fem/Perf	5.20%
	t: Num & p: Num	Sing/Sing	33.70%
	t: Case & p: VerbForm	Acc/Fin	5%

Table 6. Examples of Binary Variables

	type	value	frequency
English	t: POS & t: Num & t: PronType	PRON/Sing/Prs	8.90%
	p: POS & p: VerbForm & p: Mood	VERB/Fin/Ind	19.90%
	t: Rel & t: Definite & t: PronType	det/Def/Art	5.50%
	p: Rel & p: VerbForm & p: Tense	root/Fin/Past	7.70%
	t: Num & p: POS & p: Tense	Sing/Verb/Past	9.10%
	t: Rel & p: POS & p: Mood	nsubj/Verb/Ind	5.10%
Polish	t: Mood & t: VerbForm & t: Voice	Ind/Fin/Act	14%
	p: POS & p: VerbForm & p: Mood	VERB/Fin/Ind	41.30%
	t: POS & t: Animacy & t: Voice	VERB/Hum/Act	5%
	p: Rel & p: Num & p: Aspect	conj/Sing/Perf	5%
	t: Num & p: POS & p: VerbForm	Sing/VERB/Fin	20.60%
	t: Num & p: POS & p: Rel	Sing/VERB/conj	5%

Table 7. Examples of Ternary Variables

It is important to emphasize that it is the *pairing* of the (possibly compound) types *and* their (possibly compound) values which constitute the input variables for this investigation. In Polish, for example, a type combining the Number of the target word, with the part of speech and dependency relation of the target word's parent can have many possible values. But only those combinations of values that occur no less often than the 5% frequency of Singular-Verb-conjunct have been taken into account. All less frequent values for this type have been excluded from the variables and ignored. [7] 25

This generation and selection of variables produce, for each text in a language corpus, a matrix in which each row represents a token of the text and each column represents a type-value pair as described above. The total number of variable columns differs for each language (see Table 4). For each row, each cell is assigned a value of one if the token has the morpho-syntactic characteristic for that column; otherwise, zero is assigned. The design of the variables means any given token can have positive values in numerous columns. The German corpus, for example, has a mean of 48.8 positive columns per token (of a total 592 variables). 26

3. Classification

27 The purpose of this study is to test whether automated Universal Dependency annotation, as supplied by the udpipe program, can be useful for authorship attribution and related problems. It seeks to understand whether such an input contains enough information to overcome the noise introduced by a level of mistakes, which can be relatively high as compared to the results of expert "hand annotation." Since it is known that the information/noise ratio worsens for shorter input texts [Eder 2017], this investigation focuses on the degree to which classification performance degrades as the size of the input decreases. This process requires texts of various lengths, with all other factors held constant, insofar as this may be possible. This requirement is met by creating a series of smaller "texts" for each author by sampling the full text.

28 At each stage of classification, each text in each language corpus was divided into smaller texts, and these smaller units were classified according to author. [Gorman 2020] has shown that — at least for expert-annotated ancient Greek prose — classification on the basis of combinations of morpho-syntactic characteristics is practically error-free for input texts larger than 500 tokens (> 99%). Exploratory tests for our study confirm this high level of accuracy for text sizes of between 2000 and 600 tokens. Thus, this paper will present only the data for the smallest texts in the investigation; it begins with texts of 500 tokens, and then decreases to 400, 300, 200, 100, and finally 50 tokens.

29 As is traditional in textual attribution studies, each text was treated as a "bag of words" for the purpose of division into samples. 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. Thus, segments may contain tokens from many parts of the original text without regard for their original order.[8]

30 Once segments of the appropriate length have been generated, the next step is to aggregate the values for all variables in each segment. For example, a matrix with 20,000 rows representing as many separate tokens, each with a 1 or a 0 in its variable columns, is replaced by a new matrix with 40 rows, each representing 500 tokens. Each variable column now contains the sum of the relevant column for the underlying tokens. Columns values are then normalized so that sums are replaced by relative frequencies. These "segment matrices" become the input for the classifying algorithm.

31 To present a difficult classification problem [Luyckx and Daelemans 2008] [Luyckx and Daelemans 2011], each language corpus contains one work apiece by 20 different authors. Given the number of suitable texts available in the public domain, this number represent an attempt to balance between including more languages with fewer authors for each and including fewer languages with more authors in each.

32 For the classification algorithm itself, logistic regression was chosen. This method has an ability to handle a large number of observations and variables. It is also able to function well in the presence of many co-linear variables. Specifically, the *LiblineaR* package for the *R Project for Statistical Computing* was chosen [Fan 2008] [Helleputte 2017]. This package offers a range of linear methods; we selected the L-2 regularization option for logistic regression.[9]

33 For each input text size in each language, 90% of the data was used for training the classifier and the remaining 10% set aside for testing. Inclusion of a segment in the training or testing set was random. However, the amount of text by individual authors in some language corpora varies a great deal, and this situation may affect results, with the classifier, for example, strongly preferring the most frequent author in the training set. Thus, the random assignment into training and test sets was guided by associating a selection probability with each segment so that each of the *n* authors in a corpus was represented by approximately $1/n$ of the segments in the test set. This balance prevents undue bias in the classifier. To validate the results of the classification testing, we used Monte Carlo subsampling [Simon 2007] applied at two levels. As a rule, the populating of the segments with randomly selected tokens was carried out ten times.[10] For each of these partitionings to create text segments, 100 additional random divisions into a training set and a test set were made. As noted below, because the classification process becomes much slower as text size falls, fewer iterations of both kinds of subsampling were used for the smallest segments. The results of the tests at each text size were averaged. These steps minimize the effects of the internal make-up of particular segments or of their inclusion or exclusion from the training/testing groups.

4. Results and Discussion

For each language corpus, the results of classification according to the size of the input text segments are given in Table 8. Since the goal of this study is to evaluate for authorship attribution the sufficiency of machine-generated Universal Dependency data, no consideration is given to various specialized measurements of classification success. Instead, the simplest measure of accuracy is used: percentage of successful attributions. The mean accuracy over all attempts is given in the top row of each cell; the range is given below the mean. 34

	500	400	300	200	100	50
Polish	99.35 (95.8-100)	99.05 (95.3-100)	98.18 (94.5-100)	96.28 (92.7-99.3)	88.32 (84.3-92.0)	74.19 (71.0-78.5)
Finnish	99.52 (99.3-99.7)	99.03 (98.7-99.3)	98.15 (97.8-98.4)	95.56 (95.1-96.0)	86.44 (85.8-86.9)	70.61 (70.2-71.1)
Spanish	98.87 (98.7-99.0)	98.25 (98.0-98.4)	96.75 (96.4-97.0)	93.18 (92.8-93.5)	81.2 (80.6-81.6)	61.7 (61.6-62.0)
German	96.7 (93.2-99.2)	95.4 (91.6-98.4)	92.85 (87.9-95.5)	87.99 (84.8-91.1)	76.95 (71.7-83.0)	56.07 (54.5-58.2)
English	98.26 (96.4-99.8)	97.25 (94.0-99.4)	94.75 (92.5-97.5)	89.7 (87.3-92.6)	74.45 (71.6-77.4)	53.8 (51.6-55.9)

Table 8. Classification Accuracy

These data make clear that morpho-syntactic variables derived from UD automated parsing suffice to identify author/work in a closed set of authors with a single work for each. For all tested languages, accuracy is greater than 90% with text segments larger than 300 tokens. Each language corpus contains 20 classes for attribution; 90% accuracy is roughly 18 times random chance since the “no-information rate” for each iteration varies slightly from 5%.[11] Such success offers promise for the feasibility of classification when vocabulary- and/or character-based variables are not appropriate. 35

A few more specific observations about these results are in order. In general, accuracy tracks well with the size of the variable set for each language (see Table 4). Languages with larger sets tend to have better accuracy. We may reasonably presume that this tendency is due in part to the correlation between the number of variables and the sparsity of the input matrices. Ordinarily, we expect sparsity — the number of cells of a matrix containing zeros — to increase as the number of variables becomes greater. And more sparsity in data for natural language processing usually means less accuracy. However, recall that in this study a single token may have positive values for many variables and that any variable which has positive values for fewer than 5% of tokens has been excluded. The combination of these two factors means that as the number of total variables increases, so too does the number of variables per token (Table 9). When the token matrices are aggregated to form representations of “texts” of various lengths, the variable columns are summed. The presence of positive values in a good number of columns in the original matrix makes it unsurprising when relatively few columns in the aggregated matrix sum to 0. We may illustrate this with the example of Spanish. The average token in the Spanish matrix has 34.7 columns with a positive value. Forming, e.g., a 50-token segment by the aggregation of the matrix rows means that, on average, 1,735 (34.7 × 50) positive values will be distributed among the 385 variable columns in the resulting matrix. It stands to reason that, all else being equal, we may expect relatively few aggregated columns containing zeros. In fact, with the variables used in this study, sparsity is almost non-existent for texts larger than 100 tokens; the sparsity rates for 50-token texts are themselves unexpectedly small given the number of variables (Table 9). This low sparsity may go far to explaining the good accuracy of the classification experiment. 36

	Variable types total	Mean Variables Per token	Sparsity @50 tks.
Polish	1,120	102.9	5.8%
Finnish	932	69.7	3.8%
Spanish	385	34.7	3.4%
German	592	48.0	5.5%
English	191	16.7	5.0%

Table 9. Sparsity

A second aspect of the accuracy results reported in Table 3 needs closer elaboration. As is usual for author attribution studies that take text size into consideration, results in each language show a monotonic decline in accuracy as text size decreases. However, when viewed from the perspective of relative error rates, our morpho-syntactic variables seem quite robust against decreasing text size. 37

When considering classification accuracy of a range of text sizes, it is important to focus on the relationship between the change in accuracy and the change in text size. A 5% decrease in accuracy when text size drops from 2000 tokens to 1900 tokens is much more worrisome than the same decrease from 2000 to 1000 tokens. Table 10 gives the relevant information for this experiment. Specifically, it records the rates at which the error rate (1 – Accuracy) for each language increases as the text size is repeatedly halved: 400 to 200 tokens, 200 to 100, and 100 to 50. 38

	200/400	100/200	50/100
Polish	0.0372/.0095 = 3.92	0.1168/.0372 = 3.13	0.2581/0.1168 = 2.2
Finnish	0.044/0.0097 = 4.57	0.1356/0.044 = 3.05	0.2939/0.1356 = 2.16
Spanish	0.0682/0.0175 = 3.89	0.188/0.0682 = 2.75	0.383/0.188 = 2.03
German	0.1201/.046 = 2.61	0.2305/0.1201 = 1.91	0.4393/0.2305 = 1.90
English	0.103/0.0275 = 3.74	0.2555/0.103 = 2.48	0.462/0.2555 = 1.81

Table 10. Change in Error Rate by Text Size

The first row in each cell gives the error rate for a given text size divided by the error rate for the text that is double the size. For example, the Polish 200-token corpus had an average error rate of 3.72%, which is divided by the average error rate for the Polish 400-token corpus, 0.95%. The second line of the cell gives the corresponding result: while the text size decreases by a factor of 2, the error rate increases by a factor of 3.92. An overview of this table shows that for all languages the increase factor becomes significantly smaller as the input size itself decreases. While the absolute accuracy (or error) rate is still quite properly the metric of interest for literary and, most especially, forensic attribution methods, clearly presenting proportional relationships such as that of the rate of change in text size viz-à-viz the change in success rate should help us to better judge the effectiveness of different approaches to classification of small texts. 39

There are, of course, anomalies in the patterns observable in Table 8. For example, since the German matrix has both more total variables and more variables per token than does the Spanish, we might reasonably expect classification of German would outperform Spanish. This is not the case. Nor can an explanation be offered here.[12] We can only note that some of the differences between the success rates of various languages must be due not to characteristics of the morpho-syntactic variables, but to particularities of the language corpora. Although an effort was made, within the limits of the texts available in the public domain, to assemble corpora that were roughly similar among languages, important 40

aspects of the corpora must necessarily differ. It is inevitable, for example, that one corpus contains a higher proportion of texts that are more difficult to distinguish from each other than does another corpus. It is possible that this advantage, then, is so great that it may significantly affect the classification accuracy rates. Another complicating factor is introduced by the relative effectiveness of the udpipe annotation. Again, we must assume a possibly wide range between the most and the least accurate of the annotation algorithms. Accordingly, a relatively large amount of noise in the variables of an individual language may certainly be expected to suppress classification accuracy.

41 However, identifying such possible confounding elements and measuring their effects are not germane to the purpose of this study. Our experiments have been designed to test the hypothesis that machine-generated morpho-syntactic annotation can be used for accurate closed-set attribution, even for fairly small texts, in a range of languages. All comparisons among languages are merely illustrative and suggestive: greater morpho-syntactic complexity seems to correlate at least approximately with accuracy. Detailed investigation of this apparent relationship must be left to experts in the respective languages and literatures.

42 This investigation has set out to explore the discriminative value of a certain set of input features. It has advanced the hypothesis that, absent any vocabulary data, a mixture of a text's shallow and deep morpho-syntactic characteristics, when combined in a way that preserves as many "degrees of freedom" as possible, can produce strong results in closed-set classification. These results are significant, insomuch as most text classification methods rely primarily on vocabulary or character information, with a text's syntax represented only by function words and POS tags. However, the quantitative study of language and texts has recently seen the rise in prominence of approaches such as deep neural networks. These are often "end to end" learning systems; input consists of "raw" text with little pre-processing and no feature extraction. Results can be amazingly accurate. In the presence of such effective alternatives, the reader of this study may reasonably feel doubts about a method which requires syntactic parsing as well as significant effort given to feature engineering.

43 Yet, user-specified input features have their own advantages. The continuing efforts to find better methods of text classification is primarily driven, at least in the various disciplines of the humanities, by the desire to identify and describe the essential features of individual style. Investigations seek to confirm (or refute) the hypothesis that every user of language has a stylistic "fingerprint" or "signature" that allows written or spoken "texts" from that source to be distinguished from all others. In order for studies of this sort to be plausible, however accurate the results may be, we must be able to give a clear interpretation for the variables on which the classification depends.[13] Optimally, we must give this interpretation in terms of commonly accepted phenomena of language and communication. Generally speaking, hand-crafted input features meet this criterion. In contrast, end-to-end deep learning algorithms typically abstract their own features, and often it takes a great deal of effort to establish what phenomena the model is taking into account.[14]

44 The input features examined in this study are familiar in their basic elements and transparently interpretable. Anyone interested in the dimensions of authorial style presented here can quickly learn the outline of a language's morphology and the elementary structures of Universal Dependency grammar. As a result, classification based on UD parsing and morphology tagging puts even beginning students of style in a position to explore clear distinctions between texts and authors.

45 Since even a classification technique as fundamental as logistic regression complicates the matter by adding a layer of learned weights to the input features, the interpretability of these morpho-syntactic variables is well illustrated if looked at through the lens of a simple distance measure. Since its publication, Burrows's Delta has become one of the most widely used metrics for comparing text styles. In essence, Delta measures, for each feature of interest, how far the target text differs from the mean of the corpus, normalized by the standard deviation of each variable.[15] In spite of its simplicity, Delta can produce very good classification results [Eder 2015]. Conveniently, classification by this method is included in the "Stylo" package for R [Eder 2016]. Table 11 demonstrates the effectiveness of our morpho-syntactic variables by showing the results for the classification of 33 Polish novels using "Stylo."[16]

Text Size	Accuracy	Text Size	Accuracy	Text Size	Accuracy
2000	99.99 (96.8-100)	1500	99.98 (97.87-100)	900	99.30 (96.0-100)
1900	99.98 (97.3-100)	1400	99.84 (98.0-100)	800	98.82 (95.2-100)
1800	99.93 (97.5-100)	1300	99.90 98.1-100	700	98.17 (93.7-100)
1700	99.97 (95.0-100)	1200	99.68 93.0-100	600	97.1 (90.3-100)
1600	100 (100-100)	1100	99.3 (95.2-100)	500	95.90 (90.1-100)
		1000	99.67 (95.4-100)		

Table 11. Classification Accuracy (Burrows's Delta)

Clearly, the variables used in this study can accurately distinguish texts without the addition of weights and calculations that may obscure interpretation. A very few brief examples should suffice to make this point. 46

According to the application of Burrows's Delta to the morpho-syntactic variables presented here, the two stylistically most distinct works in our corpus of English fiction are James Fenimore Cooper's *The Last of the Mohicans* (1826) and Mark Twain's *The Adventures of Huckleberry Finn* (1885). The most distinctive feature in the set (where the two authors differ by almost four standard deviations) is Cooper's fondness for an oblique dependent of a verb that itself is modified by its own dependent. In UD grammar, the oblique of a verb is a nominal dependent of a verb that is not an argument of the verb. Here is an example from Cooper's first paragraph: "The hardy colonist, and the trained European who fought at his side …." Here, *side* in "at his side" is an oblique of *fought* and is modified by its dependent possessive *his.* If the annotations created by "udpipe" are to be trusted, Cooper used such structures to an unusual degree, while Twain avoided them. Comparing Cooper to the entire English corpus, we find that the author was also inordinately fond of modifying the object of a verb with a dependency.[17] The unusually high frequency of this construction remains even if we control for the number of objects in general, as well as when we consider the total frequency of all verbs. 47

To examine a morphologically more complex language, in our German corpus the work with the greatest mean distance to all the others is Arthur Achleitner's *Im grünen Tann* (1897?). One of the strongest *differentiae* is the author's marked preference for verbs in the Present Indicative Active 3rd-person Singular. Because the data set contains values for simple morphological elements as well as combination, we can quickly see that Achleitner's verbal tendency can be explained in terms of his very sharp relative avoidance of past tense verbs, alongside a more moderate relative deficit of plural verbs; we may assume that mood, voice, and person do not play a significant role in this stylistic peculiarity. With respect to nominal forms, Achleitner shows a relatively high number of nouns with a dependent definite article; this frequency is only partly explained by the relative number of nouns in general. At the same time, the frequency of pronouns is much lower than average. Accordingly, we might formulate, perhaps for our students, a "thumbnail" outline of Achleitner's stylistic quirks: he loves *liebt* and hates *liebte*; loves *der, die, das* and hates *er, sie, es.* It is a gross simplification to be sure, but it emphasizes the transparency of stylistic distinctions drawn on the basis of our feature set. 48

In conclusion, the evidence presented in this study has shown that morpho-syntactic features can form a basis for the successful classification of texts. Results are good across languages ranging from the morphologically complex (e.g., Polish) to the simple (e.g., English). And, while it is unlikely that the frequency of morpho-syntactic elements is unaffected by topic, genre, etc., we can reasonably suppose that it is less affected by such elements external to the author than are variables based primarily on aspects of vocabulary. Therefore, morpho-syntactic features seem essential to characterizing an author's stylistic signature. 49

A second goal of the present work has been to demonstrate that we do not need to rely on expensive expert-annotated 50

data to provide satisfactory syntactic information. Automated parsers such as udpipe are able to output Universal Dependency annotation that is sufficient to form the basis of effective variables representing both the “shallow” and “deep” syntax of a text. Because, with most variable sets used in the attribution literature, noise tends to overwhelm information when text size decreases sufficiently, this investigation has focused on short texts. Even with texts of 50 tokens, accuracy remains many times better than random chance (74%-53%). These results were achieved with the most naïve methods of feature selection and without any optimization of the parsing or classifying algorithms. We may thus suggest that further explorations of the value of UD parsing for classification promise to be productive.

Finally, we have argued that the morpho-syntactic features discussed here are advantageous, at least from the point of view of interpretability and pedagogical usefulness. Unlike n-grams, for example, morpho-syntactic variables represent traditional grammatical terms and categories. Unlike function words, such variables — at least when the preservation of “degrees of freedom” is emphasized in their construction — contain information based on every word in a text. As a result, stylistic traits formulated in terms of these variables are easily identified and illustrated in a way relatively more likely to be persuasive in the classroom or in the pages of a specialized literary journal. 51

In sum, this investigation of short text classification on the basis of machine-generated Universal Dependency annotation indicates that, at the very least, morpho-syntax should occupy a larger role in the future development of text attribution. Perhaps it should even take a central role. 52

Appendix: List of Works in Corpora

English, Finnish, German, and Spanish texts are taken from Project Gutenberg (https://www.gutenberg.org/) and are listed here by author and title. Polish texts are drawn from the web site of The Computational Stylistics Group (https://computationalstylistics.github.io/; texts at https://github.com/computationalstylistics/100_polish_novels). No bibliographic data are provided on this site. Thus, Polish texts are listed by the file name used by the repository. 53

English Corpus

Alcott, *Little Women* 54
Austen, *Pride and Prejudice*
Barrie, *Peter Pan*
Baum, *The Wonderful Wizard of Oz*
Bronte, *Wuthering Heights*
Cather, *My Antonia*
Christie, *The Mysterious Affair at Styles*
Conan Doyle, *The Hound of the Baskervilles*
Cooper . *The Last of the Mohicans*
Dickens, *A Christmas Carol*
Eliot, *Middlemarch*
Hardy, *Tess of the d'Urbervilles*
Joyce, *A Portrait of the Artist as a Young Man*
London, *White Fang*
Melville, *Moby Dick*
Montgomery, *Anne of Green Gables*
Shelly, *Frankenstein*
Sinclair, *The Jungle*
Twain, *Adventures of Huckleberry Finn*
Wharton, *Ethan Frome*

Finnish Corpus

Aho, *Hellmannin herra; Esimerkin vuoksi; Maailman murjoma* 55
Airola, *Keksijän voitto*

Alkio, *Eeva*
Anttila, *Hallimajan nuoret*
Canth, *Köyhää kansaa; Salakari*
Elster, *Päivän valaisemia pilven hattaroita*
Ervast, *Haaveilija*
Gummerus, *Peritäänkö vihakin?*
Haahti, *Valkeneva tie*
Hahnsson, *Huutolaiset*
Hannikainen, *Erakkojärveläiset*
Heman, *Kysymysmerkkejä*
Högman, *Merimiehen matkamuistelmia 1*
Ivalo, *Aikansa lapsipuoli*
Jääskeläinen, *Iloisia juttuja IV*
Jahnsson, *Hatanpään Heikki ja hänen morsiamensa*
Järnefelt, *Isänmaa*
Järventaus, *Kaukainen onni*
Järvi, *Harry*
Kataja, *Lain varjolla*

German Corpus

Achleitner, *Im grünen Tann*
Ahlefeld, *Die Bekanntschaft auf der Reise*
Aldersfeld-Ballestrem, *Die Falkner vom Falkenhof 1*
Bechstein, *Der Dunkelgraf*
Bernhard, *Die Glücklichen*
Bonsels, *Eros und die Evangelien*
Feuchtwanger, *Die häßliche Herzogin*
Fontane, *Effi Briest*
Gjellerup, *Der Pilger Kamanita*
Goethe, *Wilhelm Meisters Lehrjahre 1*
Grillparzer, *Das Kloster bei Sendomir*
Hauff, *Der Mann im Mond*
Hesse, *Unterm Rad*
Hoffmann, *Klein Zaches, genannt Zinnober*
Huch, *Der Fall Deruga*
Mann, H., *Der Untertan*
Mann, Th., *Buddenbrooks*
Meyrink, *Der Golem*
Spyri, *Heimatlos*
Zweig, *Die Liebe der Erika Ewald*

Spanish Corpus

Blasco Ibáñez, *La Catedral*
Caro, *Amar es vencer*
Carrere, *La copa de Verlaine*
Conscience, *La niña robada*
Delgado, *Angelina*
Dourliac, *Liette*
Espina, *Agua de Nieve*
Fernández y González, *Los hermanos Plantagenet*

Gil y Carrasco, *El señor de Bembibre*
Halévy, *El Abate Constantin*
Larra, *Si yo fuera rico!*
Larreta, *La gloria de don Ramiro*
Leumann, *Adriana Zumarán*
Mancey, *Las Solteronas*
Ocantos, *Quilito*
Ortega y Frías, *La Gente Cursi*
Palacio, *La alegría del capitán Ribot*
Pardo Bazán, *Una Cristiana*
Pereda, *Al primer vuelo*

Polish Corpus (20 texts used for logistic regression)

balucki_burmistrz_1887.txt
beczkowska_gniezdzie_1899.txt
dabrowska_nocednie2_1932.txt
dmochowska_dwor_1903.txt
domanska_paziowie_1910.txt
godlewska_kwiat_1897.txt
iwaszkiewicz_czerwone_1934.txt
kaczkowski_olbrachtowi_1889.txt
kossak_oreza_1937.txt
krzemieniecka_fatum_1904.txt
kuncewiczowa_twarz_1928.txt
marrene_mezowie_1875.txt
mostowicz_hanki_1939.txt
nalkowska_romans_1923.txt
prus_faraon_1897.txt
rodziewicz_lato_1920.txt
sienkiewicz_quo_1896.txt
sygietynski_calvados_1884.txt
zapolska_tagiejew_1905.txt
zeromski_przedwiosnie_1924.txt

Polish Corpus (33 texts used for classification by Burrows's Delta)

balucki_burmistrz_1887.txt
beczkowska_bedzie_1897.txt
berent_diogenes_1937.txt
dabrowska_nocednie1_1931.txt
deotyma_panienka_1893.txt
dmochowska_dwor_1903.txt
domanska_historia_1913.txt
dygasinski_as_1896.txt
godlewska_kato_1897.txt
gojawiczynska_dziewczeta_1935.txt
iwaszkiewicz_czerwone_1934.txt
kaczkowski_grob_1857.txt
korzeniowski_emeryt_1851.txt
kossak_bog_1935.txt
kraszewski_kordecki_1850.txt

krzemieniecka_fatum_1904.txt
kuncewiczowa_cudzoziemka_1936.txt
makuszynski_basie_1937.txt
marrene_bozek_1871.txt
mniszek_gehenna_1914.txt
mostowicz_hanki_1939.txt
nalkowska_granica_1935.txt
orzeszkowa_gloria_1910.txt
prus_emancypantki_1894.txt
reymont_chlopi_1908.txt
rodziewicz_lato_1920.txt
samozwaniec_ustach_1922.txt
sienkiewicz_rodzina_1894.txt
swietochowski_twinko_1936.txt
sygietynski_wysadzony_1891.txt
zapolska_tagiejew_1905.txt
zarzycka_wiatr_1934.txt
zeromski_syzyfowe_1897.txt

Notes

[1] https://www.gutenberg.org/ and https://computationalstylistics.github.io/resources/. The list of authors and works is given in the appendix.

[2] Here the metaphor “degrees of freedom” is drawn not from the statistical technical term but rather from traditional physics. Thus, the phrase is meant to call to mind the number of independent parameters that a mechanical system can have, such as the yaw, pitch, roll, etc. of a ship or airplane. Analogously, as discussed below, a German noun has three independent morphological parameters, gender, number, and case, each with multiple possible values. *Ex hypothesi*, any combination of these parameters can carry important information about an author’s style, and therefore all “degrees of freedom” should be taken into consideration.

[3] Varela et al. (2018, 2016) do not explicitly explain how the internal structure of the values in the flexion vector are handled in their approach [Varela et al. 2018] [Varela et al. 2016]. However, the number of input variables in their analysis (179 reduced to 132) is not consistent with the preservation of all degrees of freedom in their five morpho-syntactic vectors.

[4] Certainly, an algorithm could be designed to recognize the internal structure of a composite variable and calculate the frequency of the components, but this procedure would be a more cumbersome route to the same result as including more granular variables from the beginning.

[5] [Sidorov et al. 2012] explore what we might call “word-level” syntactic n-grams. The *n* in their conception represents the number of hierarchically contiguous words included, where hierarchy is determined by dependency and is analogous to linear order in the more familiar form of n-gram. [Gorman and Gorman 2016] explore a similar procedure. In the present study, greater weight is given to intra-word characteristics alongside syntactic sequences between words. The analogy with character n-grams thus seems apt.

[6] From the perspective of information theory, [Shannon 1948, 12] has pointed out that entropy is sub-additive, in that the sum of the individual entropies of several variables is greater than the joint entropy of those variables, except in the case where the variables are statistically independent. While calculating the relative contribution to information of two variables is elementary, specifying the contributions of multiple variables is an open question [Willliams and Beer 2010]. It may therefore be more prudent to turn linguistic features into a larger number of more basic variables than a smaller number of consolidated inputs.

[7] Note that it follows that one cannot pay attention to the values alone: the value Fem/Perf may be included in the variable set when it gives the values for the Gender and Aspect of a given word’s dependency parent, but the same morphological features may be excluded when they are the values of the target word’s *own* Gender and Aspect.

[8] Because udpipe works sentence by sentence, all annotation must be completed before creating new smaller “texts” by subsampling since this random selection of course disrupts sentence structure. Thus, for strictly practical reasons, the sampling process is applied to the matrices of token plus annotation rather than to the raw text. This procedure seems to imply no theoretical implications beyond those entailed by the application of the bag-of-words method in general. My thanks to the anonymous reader for raising this question.

[9] Exploratory testing revealed that the L-2 option was the quickest and most accurate of those available. L-2 regression (also called “Ridge Regression”) is also preferable for the way it handles collinearity. L-2 regression distributes the weight that collinear variables will have in a model across those collinear variables. The alternative method, L-1 or “Lasso Regression,” by contrast, arbitrarily singles out one of the collinear variables and assigns it the full weight of those variables. L-2 appears more conducive for the interpretation of the importance of individual variables since L-1 may lead the unwary to dismiss the force of a large number of collinear inputs.

[10] Processing becomes slow for text sizes of 100 tokens and fewer, depending on the morphological complexity of the language. The number of times the texts were partitioned was reduced accordingly.

[11] We may take “random chance” here to be equivalent of the reported “no-information rate.” The no-information rate in a classification experiment is the occurrence rate of the most frequently appearing class in the data. For example, if texts by Thomas Mann made up 50% of our material, then a model that classified all test segments as “Mann” would achieve 50% accuracy simply “by chance.” As indicated above, to minimize the no-information rate, our approach ensured that the representation of each class in the training set was roughly equal.

[12] It should be recognized that the number of variables produced by UD parsing and the selection process outlined above is not a precise measure of a language’s morphological complexity. Some relevant studies identify the complexity of Spanish as greater than that of German [Bittner et al. 2003] [Marzi et al. 2019].

[13] [Gabay 2021, 360] argues persuasively that stylometry should go beyond analysis of quantitative observations to focus on the exploration of “stylistic features with an interpretative yield.”

[14] For example, one method of interpretation is to inactivate one at a time hidden nodes within a neural network model to see if the “ablation” affects the outcome of interest [Lakretz et al. 2020a] [Lakretz et al. 2020b].

[15] $\Delta_{Bur}(Text_1, Text_2) = \sum_{i=1}^{n} |z_i(Text_1) - z_i(Text_2)|$, where *z* is the “z-score”: (observation value – population mean) / population standard deviation. As a sum of absolute values, Burrows’s Delta is a species of Manhattan Distance [Evert et al. 2017].

[16] The top line of each “Accuracy” cell gives the mean of all classifications for that size; the second line gives the range. For text sizes of 2000-600 tokens, data were partitioned into text segments five time; each such partition was then classified 100 times, with each classification using a different random test and training set. Because of slowing processing speed, the 500-token texts were classified only 50 times for each partitioning of segments.

[17] In UD grammar, object does not refer only to a direct object of a transitive verb, but to the second argument of any verb. For example, in “she went to the store” *store* would be considered the object of *went* on the assumption that, with that verb, an expression of goal is usually mandatory.

Works Cited

Bittner et al. 2003 Bittner, D., Dressler, W. U., and Kilani-Schoch, M. (eds). *Development of Verb Inflection in First Language Acquisition: A Cross-Linguistic Perspective.* Mouton de Gruyter, Berlin (2003).

Eder 2015 Eder, M. “Does Size Matter? Authorship Attribution, Small Samples, Big Problem”, *Digital Scholarship in the Humanities*, 30.2 (2015): 167–82.

Eder 2016 Eder, M., Rybicki, J., and Kestemont, M. “Stylometry with R: a Package for Computational Text Analysis”, *R Journal*, 8.1 (2016): 107–21.

Eder 2017 Eder, M. “Short Samples in Authorship Attribution: A New Approach,” *Digital Humanities 2017: Conference Abstracts*. McGill University, Montreal (2017), pp. 221–24. https://dh2017.adho.org/abstracts/341/341.pdf.

Evert et al. 2017 Evert, S. Proisl, T., Fotis, J., Reger, I., Pielström, S., Schöch, Ch., and Vitt, Th. “Understanding and Explaining Delta Measures for Authorship Attribution, Digital Scholarship in the Humanities”, *Digital Scholarship in the Humanities* 32.suppl. 2 (2017): ii4–ii16. https://doi.org/10.1093/llc/fqx023.

Fan 2008 Fan, R.–E., Chang, K.–W., Hsieh, C.–J., Wang, X.–R., and Lin, C.–J. “Liblinear: a Library for Large Linear Classification.” *Journal of Machine Learning Research*, 9 (2008): 1871–74.

Gabay 2021 Gabay, S. “Beyond Idiolectometry? On Racine’s Stylometric Signature”, *Computational Humanities Research Conference*, November 17–19, 2021, Amsterdam, The Netherlands (2021): 359-76. http://ceur-ws.org/Vol-2989/long_paper39.pdf

Gorman 2020 Gorman, R. “Author Identification of Short Texts Using Dependency Treebanks without Vocabulary.” *Digital Scholarship in the Humanities*, 35.4 (2020): 812–25. https://doi.org/10.1093/llc/fqz070

Gorman and Gorman 2016 Gorman, R. and Gorman, V. “Approaching Questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry”, *Open Linguistics*, 2 (2016): 500–10.

Helleputte 2017 Helleputte, T., Gramme, P., and Paul, J. *LiblineaR: Linear Predictive Models Based on the Liblinear C/C++ Library.* R package version 2.10–8 (2017).

Lakretz et al. 2020a Lakretz, Y., Dehaene, S., and King, J-R. “What Limits Our Capacity to Process Nested Long-Range Dependencies in Sentence Comprehension?” *Entropy* 22.4 (2020a): 446. doi.org/10.3390/e22040446.

Lakretz et al. 2020b Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. “Exploring Processing of Nested Dependencies in Neural-Network Language Models and Humans.” Preprint: arXiv:2006.11098 [cs.CL] (2020b). https://www.researchgate.net/publication/342352527_Exploring_Processing_of_Nested_Dependencies_in_Neural-Network_Language_Models_and_Humans

Luyckx and Daelemans 2008 Luyckx, K. and Daelemans, W. “Authorship Attribution and Verification with Many Authors and Limited Data.” In Donia Scott and Hans Uszkoreit (eds), *Proceedings of the 22nd International Conference on Computational Linguistics*, Coling (2008), pp. 513–20. https://dl.acm.org/citation.cfm?id=1599146 .

Luyckx and Daelemans 2011 Luyckx, K. and Daelemans, W. “The Effect of Author Set Size and Data Size in Authorship Attribution”, *Literary and Linguistic Computing*, 26.1 (2011): 35–55.

Marzi et al. 2019 Marzi, C., Ferro, M., and Pirrelli, V. “A Processing-Oriented Investigation of Inflectional Complexity”, *Frontiers in Communication*, 4 (2019). doi: 10.3389/fcomm.2019.00048

Nivre 2015 Nivre, J. “Towards a Universal Grammar for Natural Language Processing.” In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing.* CICLing 2015. Lecture Notes in Computer Science, 9041 (2015). Springer, Cham. DOI: 10.1007/978-3-319-18111-0_1

Shannon 1948 Shannon, C. E. “A Mathematical Theory of Communication”, *The Bell System Technical Journal*, 27 (1948): 379-423, 623-56.

Sidorov et al. 2012 Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona–Hernández, L. “Syntactic Dependency-Based N-Grams as Classification Features”, *Lecture Notes on Computer Science*, 7630 (2012): 1–11.

Simon 2007 Simon, R. “Resampling Strategies for Model Assessment and selection.” In W. Dubitzky, M. Granzow, and D. Berrar. (eds.), *Fundamentals of Data Mining in Genomics and Proteomics*. Springer, Boston (2007), pp. 173–86.

Stamatatos 2009 Stamatatos, E. “A Survey of Modern Authorship Attribution Methods”, *Journal of the American Society for Information Science and Technology*, 60.3 (2009): 538-556. doi: 10.1002/asi.21001

Swain et al. 2017 Swain, S., Mishra, G., and Sindhu, C. “Recent Approaches on Authorship Attribution Techniques — An Overview”, *International Conference of Electronics, Communication and Aerospace Technology* (ICECA), Coimbatore (2017), pp. 557-66. doi: 10.1109/ICECA.2017.8203599.

Varela et al. 2016 Varela, P. J., Justino, E., Britto, A., and Bortolozzi, F. “A Computational Approach for Authorship Attribution of Literary Texts Using Sintatic Features,” 2016 *International Joint Conference on Neural Networks* (IJCNN), Vancouver, BC (2016), pp. 4835-42. doi: 10.1109/IJCNN.2016.7727835.

Varela et al. 2018 Varela, P. J., Albonico, M., Justino, E. J. R., and Bortolozzi, F. “A Computational Approach for Authorship Attribution on Multiple Languages,” 2018 *International Joint Conference on Neural Networks* (IJCNN), Rio de Janeiro (2018), pp. 1-8, doi: 10.1109/IJCNN.2018.8489704.

Wijffels 2019 Wijffels, J. “udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the ‘UDPipe’ ‘NLP’ Toolkit”, (2019). https://CRAN.R-project.org/package=udpipe

Willliams and Beer 2010 Williams, P. and Beer, R. “Nonnnegative Decomposition of Multivariate Information.” Preprint: arXiv:1004.2515 [cs.IT] (2010). https://arxiv.org/pdf/1004.2515.pdf.