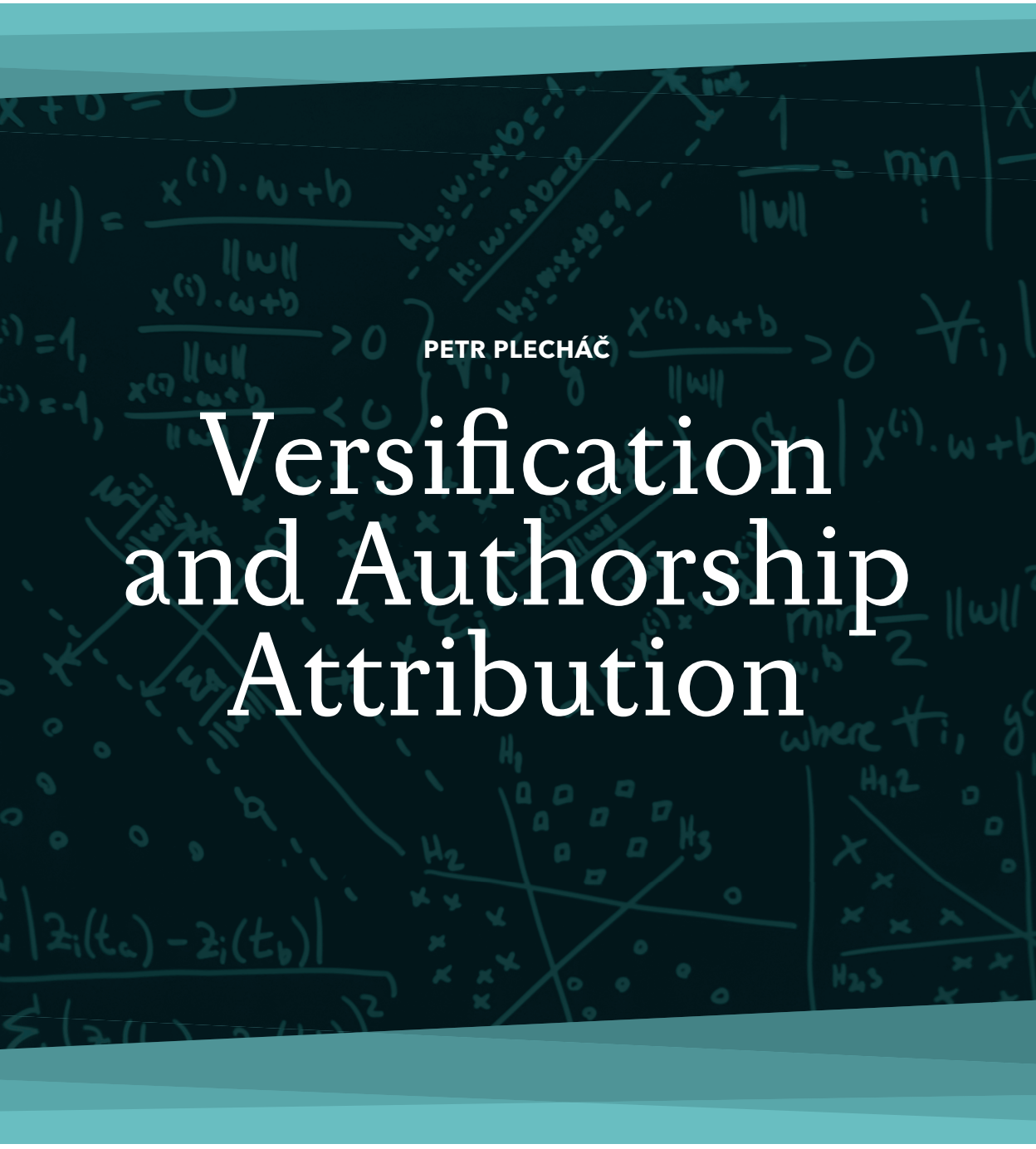
CZECH LITERATURE STUDIES



PETR PLECHÁČ

Versification and Authorship Attribution

INSTITUTE OF CZECH LITERATURE
KAROLINUM PRESS

3 Experiments

Petr Plecháč
Institute of Czech Literature, Czech Academy of Sciences
e-mail: plechac@ucl.cas.cz

https://doi.org/10.14712/9788024648903.4

3 Experiments

3.1 Data

I tested the performance of versification-based attribution on three corpora of poetic texts: *The Corpus of Czech Verse* (Plecháč 2016; Plecháč and Kolár 2015), *Metricalizer—the corpus of German Verse* (Bobenhausen and Hammerich 2015; Bobenhausen 2011) and the Spanish-language *Corpus de Sonetos del Siglo de Oro* (Navarro-Colorado, Ribes-Lafoz and Sánchez 2016; Navarro-Colorado 2015). For simplicity, these are denoted as CS, DE and ES respectively.

The general characteristics of these corpora are given in TAB. 3.1 and FIG. 3.1.

	# of authors	# of poems	# of lines	# of tokens
CS	613	80 229	2 727 632	14 923 528
DE	248	53 608	1 716 348	10 462 211
ES	52	5078	71 150	465 982

TAB. 3.1: Corpora size.

Attribution experiments clearly require the thorough tagging of all corpora. TAB. 3.2 shows that by default, only CS satisfied all of the required levels of annotation.

	CS	DE	ES
Tokenised	1	1	0
Lemmatised	1	0	0
Morphologically tagged	1	0	0
Phonetically transcribed	1	1	0
Metrically annotated	1	1	–
Stress annotated	1	1	1
Rhyme annotated	1	1	0

TAB. 3.2: Default tagging of corpora CS, DE, ES (1: tagged, 0: not tagged, –: not applicable).

https://doi.org/10.14712/9788024648903.4

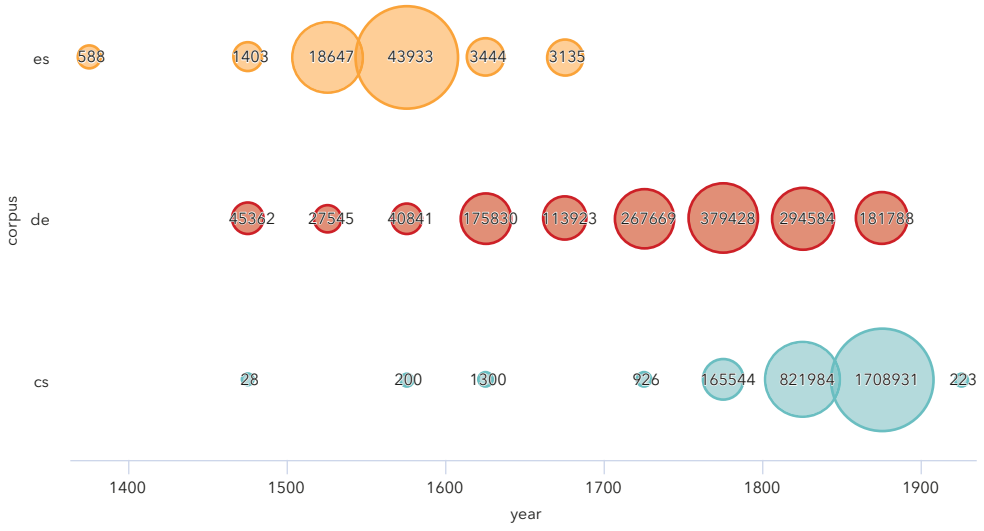


FIG. 3.1: Number of verse lines matched to the years of birth of their authors (50-year range). Circle size reflects the ratio of the given period to the total number of lines in the corpus.

It was therefore necessary to perform additional tagging. Tokenisation (ES), lemmatisation (DE, ES) and morphological tagging (DE, ES) were done with the stochastic tagger *TreeTagger* (Schmid 1994). Phonetic transcription (ES) took place via the popular TTS synthesizer *Espeak*. Rhyme recognition (ES) was done with the Python package *RhymeTagger* (Plecháč 2018).

3.1.1 Tagging Accuracy

The key issue for any automatic tagging system is its accuracy. For most of the tools used, published empirical accuracy estimations were available:

Morphological tagging, lemmatisation, tokenisation

— Spoustová et al. (2007) and Skoumalová (2011) each used a manually annotated **Czech** corpus to evaluate the morphological tagging of the combined stochastic rule-based tagger by which the CS corpus had originally been tagged. Both evaluations reported a value of **0.95** (share of correctly labelled tokens).

— Horsmann, Erbs and Zechs (2015) evaluated the morphological tagging provided by *TreeTagger* using **German** data from the *Tüba-DZ* corpus. Giesbrecht and Evert (2009) made a similar assessment using the *TIGER* corpus of newspaper texts. The author of *TreeTagger* published his own evaluation albeit based on a rather small body of texts (Schmid 1994). All of these studies reported values of approximately **0.97** (share of correctly labelled tokens).
— Göhring (2009) evaluated the morphological tagging provided by *TreeTagger* using a set of 200 manually tagged **Spanish** sentences. Instead of the portion of correctly labelled tokens, precision and recall values were reported for each tag in the tagset. Both these values achieved a micro-average of **0.94**.
— As for lemmatisation and tokenisation, these were assumed to be at least as accurate as the morphological tagging to which they are closely related in both taggers.

Metre and stress annotation

— Based on manually annotated samples, the accuracy of metrical recognition in the **CS** corpus was estimated at **0.95** (Plecháč 2016).
— Navarro-Colorado (2017) extracted a random sample of 100 sonnets from the **ES** corpus, and this was manually annotated by three subjects. The inter-annotator agreement was found to be 0.96. There was a **0.95** level of agreement of automated stress annotation between at least two of the annotators.
— For the **DE** corpus, no accuracy estimation of metre/stress annotation had been published.

Rhyme annotation

— The accuracy of *RhymeTagger* was estimated using manually annotated data in **Czech**, **English** and **French** (Plecháč 2018). Precision (P) and recall (R) were as follows: EN: P = 0.96; R = 0.88; FR: P = 0.94; R = 0.87; CS: P = 0.94; R = 0.96.

These values suggest that there was cause for optimism about the quality of the data annotation. On the other hand, the methods of evaluation differed across the corpora. Moreover, for any linguistic annotation (tokenisation, lemmatisation, morphological tagging), accuracy when tagging verse is almost certainly lower than reported owing to (1) the peculiarities of poetic speech (neologisms, word order inversions, etc.) and (2) the composition of works in older forms of the respective languages than those the tools were designed for and tested on.

For these reasons, I performed my own small-scale evaluation. I asked native speakers of each language, all of whom were professional linguists, to inspect random

samples from the corpora.[13] Since the annotations captured different linguistic levels (from individual sounds to rhymes that often spanned multiple lines), three kinds of samples were extracted from each corpus:

(1) Sample for the evaluation of tokenisation, lemmatisation, morphological tagging, phonetic transcription and stress annotation (CS: 52 lines / 287 tokens / 511 syllables, DE: 55 lines / 244 tokens / 377 syllables, ES: 98 lines / 627 tokens / 1078 syllables); the CS and DE samples consisted of at least the first eight lines[14] of randomly selected poems; the ES samples consisted of seven randomly selected sonnets;
(2) Sample for the evaluation of metre (CS: 120 lines, DE: 114 lines): each sample was made up of at least the first eight lines of randomly selected poems and
(3) sample for the evaluation of rhyme (CS: 86 rhymes, DE: 97 rhymes, ES: 183 rhymes); the CS and DE samples consisted of the initial lines of randomly selected poems that were extracted so that no rhyming lines were split; the ES sample consisted of 20 randomly selected sonnets.

TAB. 3.3 and TAB. 3.4 show the portion of tags that were evaluated as being correct. For rhyme annotation, I report both *precision* (the share of tags that corresponded with actual rhymes) and *recall* (the share of actual rhymes recognised). Since morphological tagging was used solely for rhyming words (cf. Section 2.2), I report not only overall accuracy but also the accuracy for line endings alone. As all of the values exceeded 0.9, all levels of annotation accuracy were found to be sufficient for my needs.

3.1.2 Subcorpora

I extracted eight subcorpora from CS, DE and ES (CS1, CS2, CS3, DE1, DE2, DE3, ES1, ES2). In each case, the authors in the subcorpus had been born in a preselected time span. These eras were chosen based on two factors: (1) the need to provide sufficient data (see below) and (2) the desire to approximate common literary periodisations where possible (e.g. CS1 comprised authors of the Czech National Revival; CS2

13 Generous assistance was provided by Michal Kosák (Institute of Czech Literature, Czech Academy of Sciences), Michael Wögerbauer (Institute of Czech Literature, Czech Academy of Sciences), Helena Bermúdez-Sabel (Université de Lausanne) and Clara Isabel Martínez Cantón (Universidad Nacional de Educación a Distancia, Madrid).

14 The logic behind the choice of opening lines was that this would provide evaluators with sufficient context in which to judge the results of disambiguation. This was important both from the standpoint of metre (e.g. possible metrical shifts within a poem might lead an evaluator to misclassify it) and rhyme (if one rhyming line fell outside a sample, then this too might result in misclassification).

	Tokenisation	Lemmatisation	Morphological tagging		Phonetic transcription
			Overall	Line endings	
CS	1	0.9692	0.9577	0.9302	1
DE	1	0.9385	0.9590	0.9836	1
ES	1	0.9426	0.9011	0.9984	0.9936

TAB. 3.3: Accuracy estimations for tokenisation, lemmatisation, morphological tagging and phonetic transcription.

	Rhyme annotation		Stress annotation	Metrical annotation
	Precision	Recall		
CS	0.9882	0.9767	1	1
DE	1	0.9794	0.9602	1
ES	0.9800	1	0.9944	–

TAB. 3.4: Accuracy estimations for annotations of rhyme, stress and metre.

comprised the "Lumír" generation; DE3 was mainly composed of German Romantic authors).

One metre or group of closely related metres was selected for each subcorpus. The breakdown was as follows: CS1: trochaic tetrameters with both strong and weak endings (T4); CS2–3: iambic pentameter with weak endings (I5w); DE1–3: accentual verse (F); and ES1–2: hendecasyllabic verse (11σ).[15]

Each author was represented by at least 10 samples written in the relevant metre(s). Each sample consisted of 100 lines and at least 40 rhyming pairs. Multiple poems could be combined in a sample, and no poem contributed to more than one sample.

Details of the subcorpora can be seen in TAB. 3.5.

3.2 Versification-Based Attribution

In the first battery of experiments, I tested the performance of attribution based solely on versification features.

To begin, I reduced each subcorpus to 50 samples as follows: (1) five authors were randomly selected (this did not apply to CS3 and ES1, which both comprised only five authors) and (2) 10 samples were randomly selected for each author. Each sample

15 This was the only metre in the ES corpus.

Subcorpus	Metre(s)	Era of birth	# of authors	Authors (# of samples)
CS1	T4	1760–1820	9	Čelakovský, František Ladislav (12); Havelka, Matěj (13); Hněvkovský, Šebestián (11); Kulda, Beneš Metod (27); Nejedlý, Vojtěch (17); Picek, Václav Jaromír (21); Pohan, Václav Alexander (10); Tablic, Bohuslav (16); Vinařický, Karel Alois (15)
CS2	I5w	1840–1855	7	Čech, Svatopluk (13); Kvapil, František (11); Mokrý, Otokar (15); Nečas, Jan Evangelista (10); Sládek, Josef Václav (16); H. Uden (17); Vrchlický, Jaroslav (281)
CS3	I5w	1860–1870	5	Klášterský, Antonín (64); Kvapil, Jaroslav (19); Leubner, František (10); Machar, Josef Svatopluk (22); Sova, Antonín (15)
DE1	F	1650–1699	6	Brockes, Barthold Heinrich (51); Drollinger, Carl Friedrich (11); Gottsched, Johann Christoph (29); Kuhlmann, Quirinus (30); Neukirch, Benjamin (21); Tersteegen, Gerhard (25)
DE2	F	1730–1754	5	Goethe, Johann Wolfgang (46); Jacobi, Johann Georg (12); Müller, Friedrich (15); Pfeffel, Gottlieb Konrad (28); Wieland, Christoph Martin (23)
DE3	F	1760–1794	7	Bernhardi, Sophie (12); Eichendorff, Joseph von (32); Grillparzer, Franz (52); Müller, Wilhelm (16); Schenkendorf, Max von (10); Schulze, Ernst (19); Tieck, Ludwig (28)
ES1	11σ	1500–1560	5	de Acunya, Hernando (10); de Borja, Francisco (17); de Cetina, Gutierre (31); de Góngora, Luis (14); de Herrera, Fernando (39)
ES2	11σ	1561–1599	6	Argensola, Bartolome (19); de Quevedo, Francisco (63); de Rojas, Pedro Soto (15); de Tassis y Peralta, Juan (25); de Ulloa y Pereira, Luis (13); de Vega, Lope (167)

TAB. 3.5: Subcorpora details (T4: trochaic tetrameter with both strong and weak endings; I5w: iambic pentameter with weak endings; F: accentual verse; 11σ: hendecasyllabic verse).

was then represented as a vector defined by the following versification features (as described in detail in Chapter 2):

(1) frequencies of rhythmic 2-, 3- and 4-grams for syllabic and accentual syllabic verse (CS, ES); frequencies of the 100 most common rhythmic types for accentual verse (DE);
(2) frequencies of morphological, phonetic and rhythmic rhyme characteristics; and
(3) frequencies of sounds.

I opted for an SVM as a classifier using the *one-vs.-one* strategy for multiclass classification (cf. Section 1.4.2). Implementation took place through the *SVC* module of the *scikit-learn* library[16] with the following settings (cf. Section 1.4.1):

16 <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

— *kernel = "linear"* (no kernel transformation);
— *C=1* (default value of the penalising parameter; different settings had only a negligible impact on results).

For other parameters, default values were used.

Accuracy for each subcorpus was estimated using *leave-one-out* cross validation. As there was a fairly low number of samples per class, using standard *leave-one-out* validation might have biased the results since the actual author was only represented by nine samples in the training data while the other authors were each represented by 10 samples. To eliminate this risk, one randomly selected sample was dropped from the training data for every author besides the test sample author. This equalising approach was applied in all of the experiments described in this book, unless indicated otherwise.

To achieve more representative results, I repeated this entire process 30 times with a new random selection of both authors and samples in each iteration. The entire procedure is captured in the following code in Python:

```
'''
A dict contains authors' samples (represented by vectors):

samples = {
    'author1': [sample1, sample2, …],
    'author2': [sample1, sample2, …],
…
}
'''

import random
from sklearn.svm import SVC

classifier = SVC(kernel='linear', C=1)
n_authors = 5
n_samples = 10
n_iterations = 30

for iteration in range(n_iterations):
    selected_samples = {}
    correct_classifications = 0

    # Select 5 authors/10 samples at random
    for author in random.sample(samples.keys(), n_authors):
        selected_samples[author] = random.sample(samples[author], n_samples)

    # Cross-validation: iteratively select one sample as the test sample
    for test_author in selected_samples:
        for i,test_sample in enumerate(selected_samples[test_author]):

            # Add remaining samples of the test sample author to the training set
            X = selected_samples[test_author][:i] + selected_samples[test_author][i+1:]
            y = [test_author] * (n_samples - 1)
```

```
            # Add samples of other authors to the training set but always
            # drop one sample at random
            for a in [x for x in selected_samples if x != test_author]:
                X.extend(random.sample(selected_samples[a], n_samples - 1))
                y.extend([a] * (n_samples - 1))

            # Train the classifier and classify the test sample
            classifier.fit(X, y)
            predicted = classifier.predict([test_sample])
            if predicted[0] == test_author:
                correct_classifications += 1

    print('iteration #{0}: accuracy = {1}'.format(
        iteration + 1,
        correct_classifications / (n_samples * n_authors)
    ))
```

The results of cross-validation are given in FIG. 3.2.[17] Since each of the 300 values significantly exceeded the *random baseline* (for five authors represented by 10 samples, each RB = 0.2; cf. Section 1.4.4), I judged versification features to be a reliable indicator of a text's authorship.

These results, however, differed greatly across the subcorpora. Generally the models fell into two groups:

(1) *Highly accurate models* (CS1–3, ES1) whose medians ranged from 0.94 to 0.96 and lower quartiles ranged from 0.90 to 0.95 and

(2) *Accurate enough models* (DE1–3, ES2) whose medians ranged from 0.74 to 0.82 and lower quartiles ranged from 0.72 to 0.78.

There are many possible reasons for these differences, but they are almost impossible to trace since machine learning generally works like a "black box" (we have access to both the input and the output but what's going on inside is difficult to interpret). However, one plausible explanation may relate to the amount of data. For authors with a large number of samples—for example, Jaroslav Vrchlický (281 samples), Lope de Vega (168 samples), Francisco de Quevedo (64 samples), Johann Wolfgang Goethe (46 samples), Barthold Heinrich Brockes (51 samples) and Franz Grillparzer (52 samples)—recognition tended to be less accurate than it was for other authors in the same subcorpus (TAB. 3.6). If we assume that the larger an author's body of work (or more precisely, the longer their career), the greater its stylistic variation, this phenomenon is quite intuitive.

17 Unless stated otherwise, all boxplots in this book have the following format: The box shows the interquartile range ($Q_{0,25}$; $Q_{0,75}$); the midway line represents the median ($Q_{0,5}$); and its value is given in the label rounded to two decimal places. Whiskers represent the minimum and maximum of the distribution.

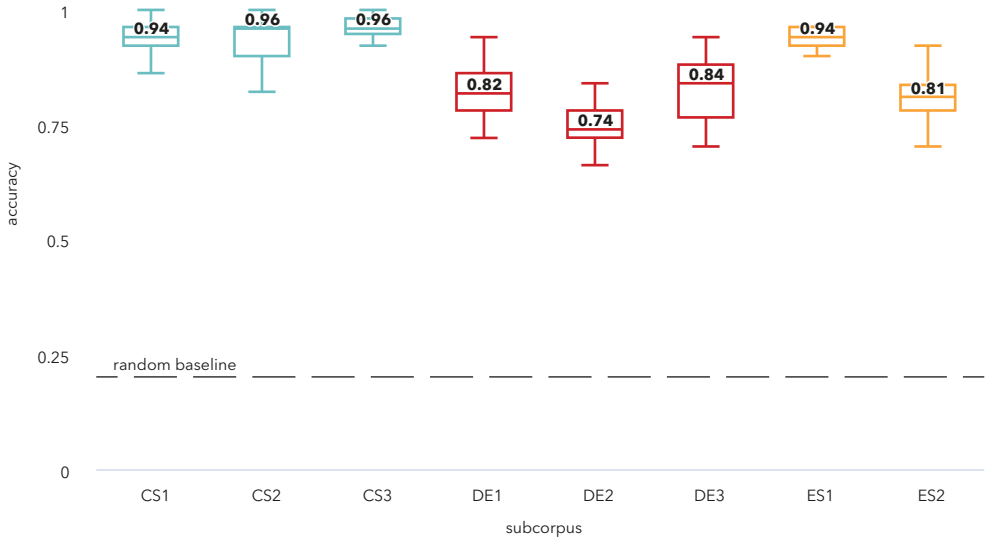


FIG. 3.2: Cross-validation results for versification-based models (30 iterations with random sampling).

The presence of two of the mentioned prolific authors in the ES2 subcorpus might help explain why its values were lower than those for ES1. In the case of the German subcorpora, we may also consider the impact of a specific type of versification (accentual verse) or a different method of rhythmic analysis (rhythmic types). Moreover a variety of cultural-historical factors may have been significant.[18] These factors are, however, beyond the scope of the present work.

3.2.1 Feature Importance

Aside from the performance differences across the subcorpora, it is also worth exploring the contribution of particular features. Failing to do this would leave open the possibility that some of the features were completely irrelevant. The option would remain that purely versification-based features yielded no information at all and the classification depended entirely on sound frequencies. In languages with a highly phonemic orthography like Czech or Spanish, this would basically mean that the

18 It may generally be assumed, for instance, that Romantic poets put more effort into individualising the rhythm of their poems than Baroque poets did.

		Čelakovský	Havelka	Hněv.	Kulda	Nejedlý	Picek	Pohan	Tablic	Vinařický
	Čelakovský	**0.91**						0.06		0.01
	Havelka	0.04	**0.95**					0.01		0.02
	Hněvkovský	0.04		**0.94**		0.04		0.01	0.06	
	Kulda				**0.99**			0.01	0.01	
CS1	**Nejedlý**			0.06		**0.95**		0.01	0.04	
	Picek	0.01	0.02				**1**	0.05		0.01
	Pohan	0.01	0.03					**0.83**		0.02
	Tablic				0.01	0.01			**0.89**	
	Vinařický							0.02		**0.94**
		Čech	**Kvapil**	**Mokrý**	**Nečas**	**Sládek**	**Uden**	**Vrchlický**		
	Čech	**1**	0.08			0.01	0.04	0.03		
	Kvapil		**0.9**			0.01		0.05		
	Mokrý			**1**						
CS2	**Nečas**				**1**	0.02				
	Sládek		0.01			**0.87**		0.04		
	Uden						**0.96**			
	Vrchlický		0.01			0.09		**0.86**		
		Klášterský	**Kvapil**	**Leub.**	**Machar**	**Sova**				
	Klášterský	**0.9**	0.02							
	Kvapil	0.07	**0.97**							
CS3	**Leubner**			**1**						
	Machar	0.02			**0.97**	0.03				
	Sova	0.01			0.03	**0.96**				
		de Acunya	**de Borja**	**de Cetina**	**de Góng.**	**de Herrera**				
	de Acunya	**0.94**		0.08						
	de Borja		**1**		0.01					
ES1	**de Cetina**	0.06		**0.92**		0.01				
	de Góngora				**0.81**					
	de Herrera				0.17	**0.99**				

		Argensola	de Quev.	de Rojas	de Tassis	de Ulloa	de Vega	
ES2	**Argensola**	**0.8**	0.04	0.02	0.04	0.02	0.07	
	de Quevedo	0.05	**0.72**		0.16	0.01	0.06	
	de Rojas		0.01	**0.95**			0.17	
	de Tassis y P.	0.09	0.18		**0.78**		0.02	
	de Ulloa y P.	0.04	0.04		0.01	**0.98**		
	de Vega	0.02	0.02	0.03			**0.68**	
		Brockes	**Droll.**	**Gott.**	**Kuhl.**	**Neu.**	**Terst.**	
DE1	**Brockes**	**0.76**	0.06	0.1	0.04	0.05	0.09	
	Drollinger	0.07	**0.84**	0.11	0.05	0.02		
	Gottsched	0.17	0.1	**0.77**	0.03	0.1	0.06	
	Kuhlmann				**0.88**			
	Neukirch			0.01		**0.83**		
	Tersteegen			0.01			**0.85**	
		Goethe	**Jacobi**	**Müller**	**Pfeffel**	**Wie.**		
DE2	**Goethe**	**0.53**		0.07	0.01	0.05		
	Jacobi	0.22	**0.83**	0.1	0.04	0.02		
	Müller	0.19	0.03	**0.76**	0.04	0.09		
	Pfeffel	0.03	0.13	0.01	**0.85**	0.08		
	Wieland	0.03		0.05	0.06	**0.76**		
		Bernhardi	**Eichen.**	**Grill.**	**Müller**	**Schen.**	**Schu.**	**Tieck**
DE3	**Bernhardi**	**0.97**						
	Eichendorff		**0.88**	0.1	0.1	0.03	0.05	0.03
	Grillparzer			**0.7**				
	Müller		0.01	0.03	**0.72**	0.02	0.04	0.03
	Schenkendorf		0.05	0.08	0.1	**0.89**	0.02	0.05
	Schulze	0.02	0.01	0.04	0.01		**0.82**	0.1
	Tieck	0.01	0.06	0.05	0.07	0.06	0.08	**0.79**

TAB. 3.6: Confusion matrices for versification-based models (relative counts). Rows show the author predicted by the model while columns show the actual author. Individual cells give the relative count of the relevant prediction.

classification was determined by a common stylometric indicator, that is, by character frequencies.

To explore how particular features contributed to the classification, I repeated the set of experiments described above. In lieu of cross-validation, this time all of the data were used to train the model for each of the 30 iterations with the *one-vs.-rest* strategy (each iteration, thus, constructed five hyperplanes). In this way, up to 30 hyperplanes were constructed for each author (the final number depended on how many times the author was randomly selected).

As discussed in Section 1.4.3 (formula 1.11), the separating hyperplane between two classes is defined by a normal vector $\mathbf{w}$ and a parameter b. Each iteration i in which author A occurred, thus, produced a normal vector $\mathbf{w}_{A,i} = (w_{A,i,1}, w_{A,i,2}, \ldots, w_{A,i,m})$, whose coordinates conveyed information about the importance of particular features. However, rather than the coordinates themselves, which might be either positive or negative, what mattered here was their absolute value. The importance of the j-th feature ($j \in [1,m]$) for the recognition of A in iteration i was, thus, assessed based on the value of $w_{A,i,j}$ squared. The overall importance of j to A across all N iterations was then assessed by means of a score calculated as follows:

$$s_{A,j} = \sum_{i=1}^{N} \frac{w_{A,i,j}^2}{N} \tag{3.1}$$

Finally, for each A, I collected the 30 features with the highest scores (i.e. the features that generally contributed most to author recognition).

As the total number of these features was in the hundreds, I regrouped them into the categories given in Chapter 2. TAB. 3.7 shows the distributions of the 30 highest-scoring features across these groups.

		r-2-gram	r-3-gram	r-4-gram	rh-pos	rh-snds	rh-stress	rh-word	snds-f
CS1	Čelakovský	0.07	0.13	0.13	0.1	**0.5**			0.07
	Havelka	0.1	0.1	0.1	0.03	**0.43**	0.07	0.07	0.1
	Hněvkovský	0.07	0.07	0.03	0.1	**0.57**	0.03		0.13
	Kulda	0.13	0.13	0.13	0.13	**0.33**			0.13
	Nejedlý	0.33	0.07	0.13	0.07	**0.33**	0.03	0.03	0.3
	Picek	0.17	0.2	**0.23**	0.1	0.2	0.03	0.03	0.03
	Pohan	0.07	0.17	0.23	0.07	**0.43**			0.03
	Tablic	0.07	0.1	0.13	0.03	**0.4**	0.13	0.1	0.03
	Vinařický	0.1	0.13	0.17	0.03	**0.23**	0.07	0.07	0.2
CS2	Čech			0.03	0.13	**0.63**	0.03	0.03	0.13
	Kvapil	0.07	0.13	0.23	0.07	**0.27**		0.03	0.2
	Mokrý	0.13	0.2	**0.3**		0.2	0.03	0.03	0.1
	Nečas	0.07	0.13	0.2	0.03	**0.33**	0.03	0.03	0.17

		r-2-gram	r-3-gram	r-4-gram	rh-pos	rh-snds	rh-stress	rh-word	snds-f
	Sládek	0.03	0.1	0.2	0.07	**0.43**	0.03	0.03	0.1
CS2	Uden	0.03	0.1	0.2	0.1	**0.37**		0.03	0.17
	Vrchlický	0	0.07	0.13	0.1	**0.47**	0.03	0.03	0.13
	Klášterský	0.2	0.2	**0.27**	0.03	0.17			0.13
	Kvapil	0.1	0.1	0.07	0.13	**0.4**		0.03	0.17
CS3	Leubner	0.13	0.13	**0.23**	0.1	**0.23**		0.07	0.1
	Machar	0.13	0.13	0.1	0.1	**0.17**	0.1	0.1	**0.17**
	Sova	0.07	0.17	**0.3**	0.07	**0.3**			0.07
	de Acunya	0.03	0.1	0.13	**0.33**	0.13			0.27
	de Borja	0.23	0.27	**0.3**		0.1			0.1
ES1	de Cetina	0.1	0.2	**0.3**	0.17	0.07			0.17
	de Góngora	0.07	0.03	0.1	**0.27**	**0.27**		0.03	0.23
	de Herrera	0.03	0.1	0.2	0.17	0.23			0.27
	Argensola			0.07	0.23	**0.4**		0.07	0.23
	de Quevedo	0.13	0.1	0.13	**0.27**	0.13		0.03	0.2
ES2	de Rojas	0.17	0.13	**0.27**	0.17	0.1			0.17
	de Tassis y P.	0.07	0.03	0.07	0.2	**0.3**		0.03	0.3
	de Ulloa y P.		0.1	**0.33**	0.17	0.1			0.17
	de Vega	0.13	0.13	0.07	**0.27**	0.2			0.2
	Brockes			0.23	**0.43**	0.23			0.03
	Drollinger			0.1	**0.37**	0.3	0.03	0.07	0.17
DE1	Gottsched			0.07	**0.5**	0.23		0.03	0.13
	Kuhlmann			**0.3**	0.2	0.27		0.07	0.2
	Neukirch			0.37	**0.5**	0.03		0.03	0.07
	Tersteegen			0.13	**0.5**	0.13	0.03	0.03	0.2
	Goethe			0.23	**0.5**	0.07	0.03		0.13
	Jacobi			**0.27**	**0.27**	**0.27**	0.03	0.03	0.17
DE2	Müller			0.3	**0.37**	0.17			0.13
	Pfeffel			0.2	**0.33**	0.17	0.17	0.03	0.13
	Wieland			**0.53**	0.33	0.03			0.03
	Bernhardi			0.3	**0.33**	0.27	0.03	0.07	0.07
	Eichendorff			0.23	**0.53**	0.13	0.03		0.03
	Grillparzer			**0.3**	**0.3**	0.2		0.03	0.13
DE3	Müller			**0.33**	**0.33**	0.2	0.03	0.07	0.07
	Schenkendorf			**0.47**	0.17	0.13		0.03	0.07
	Schulze			**0.37**	0.33	0.13		0.17	0.13
	Tieck			**0.33**	**0.33**	0.07	0.03	0.03	0.23

TAB. 3.7: Feature importance. (1–3) rhythmic *n*-grams/rhythmic types, (4) morphological characteristics of rhyme, (5) phonic composition of rhyme, (6) stress placement in rhyme, (7) word length in rhyme, (8) sound frequencies. The table shows the share of elements in these categories reflected in the 30 highest-scoring features for each author. The highest value in each row is highlighted in bold.

Among the Czech subcorpora, the phonic composition of rhymes tended to be the most prominent category. In contrast, for German works, morphological characteristics played this role, and for the Spanish subcorpora, the results were somewhere in between. Rhythmic characteristics also played an important part in all three corpora. Of the rhythmic *n*-grams (CS, ES), rhythmic tetragrams were most prominent. The significance of word length and stress placement in rhyme was fairly weak across all the subcorpora.

Concerning the stress placement in rhyme, all values were zero in both Spanish subcorpora. The explanation for this was quite simple: one constant of the Spanish hendecasyllable is that the final stress falls on the penultimate syllable:

¡Peñascos Altos, de la mar batidos,
rhythm: 0 1 0 1 0 0 0 1 0 1 0
de nubes coronadas las cabezas,
rhythm: 0 1 0 0 0 1 0 0 0 1 0
donde se rompen en diversas piezas
rhythm: 0 0 0 1 0 0 0 1 0 1 0
cristales espumosos resistidos
rhythm: 0 1 0 0 0 1 0 0 0 1 0
(Lope de Vega)

There was no exception to this rule across the ES corpus. The null variability of this stress placement on rhyming words, thus, led to its null applicability for classification. On the whole, however, none of the categories appeared dominant and none could be dismissed as irrelevant.

3.3 Comparison with Lexicon-Based Models

The goal of the second battery of experiments was to compare the performance of versification-based models with that of models based on standard stylometric features (again for simplicity, these are referred to—albeit imprecisely—as "lexicon-based" models). Through these same tests, I also assessed the performance of models combining versification-based and lexicon-based features.

3.3.1 Fine-Tuning

Before proceeding with these comparisons, it was necessary to choose a domain (words, lemmata or character *n*-grams) and the number of types of each feature to be analysed. To find the optimal solution, I first trained and cross-validated many different models and found the best-performing settings.

When fine-tuning, it is good practice to employ different datasets to the ones that will be used to measure accuracy. Since in this case, there was no need to limit the poems to any particular metre, plenty of data were available in CS and DE to build alternative subcorpora for validation (denoted here as CS' and DE'; see TAB. 3.8 for details). This unfortunately was not the case for ES where there was no option but to use ES1 and ES2 themselves for this purpose. The results for those subcorpora, thus, provide only a very general comparison.

Subcorpus	Era of birth	# of authors	# of samples
CS1'	1760–1820	32	986
CS2'	1840–1855	24	1190
CS3'	1860–1870	27	1476
DE1'	1650–1699	8	486
DE2'	1730–1754	10	598
DE3'	1760–1794	16	1295

TAB. 3.8: Validation of the subcorpora.

In training the models, I followed the design sketched above for five randomly selected authors and 10 randomly selected samples (cf. Section 3.2). Over 30 iterations, I then performed *leave-one-out* cross-validation using an SVM with the set of features below:

(1) frequencies of the *n* most common words,
(2) frequencies of the *n* most common lemmata,
(3) frequencies of the *n* most common character bigrams,
(4) frequencies of the *n* most common character trigrams and
(5) frequencies of the *n* most common character tetragrams,

where $n \in \{50, 100, 150, \ldots, 2000\}$.

The results (FIG. 3.3) confirmed a pattern observed in previous studies, namely that the relationship between the number of types analysed (*n*) and the attribution accuracy rose sharply, and then, after reaching a certain value, tended to stabilise (cf. Eder 2011; Rybicki-Eder 2011; Smith-Aldridge 2011). While the value appeared similar for

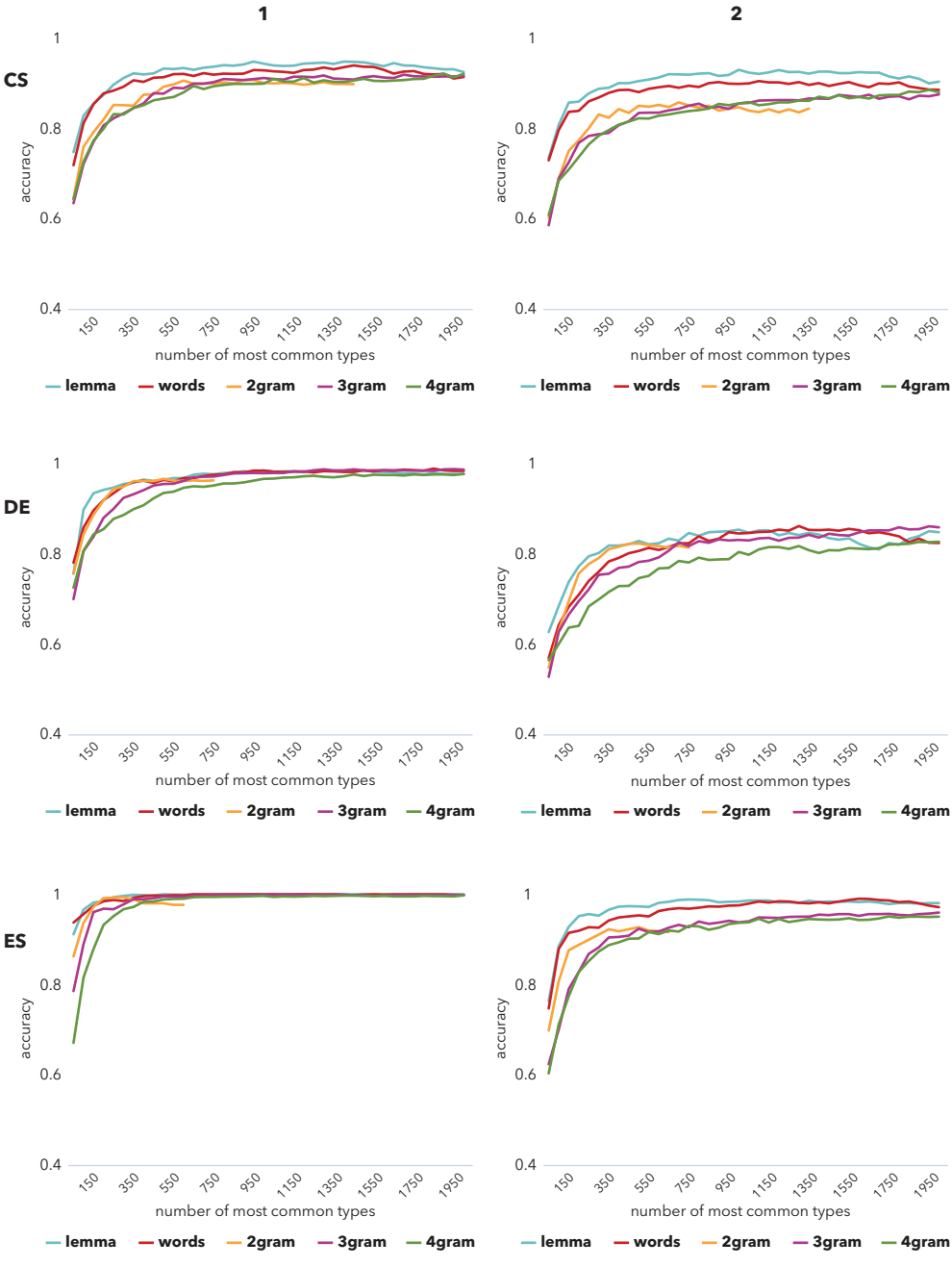
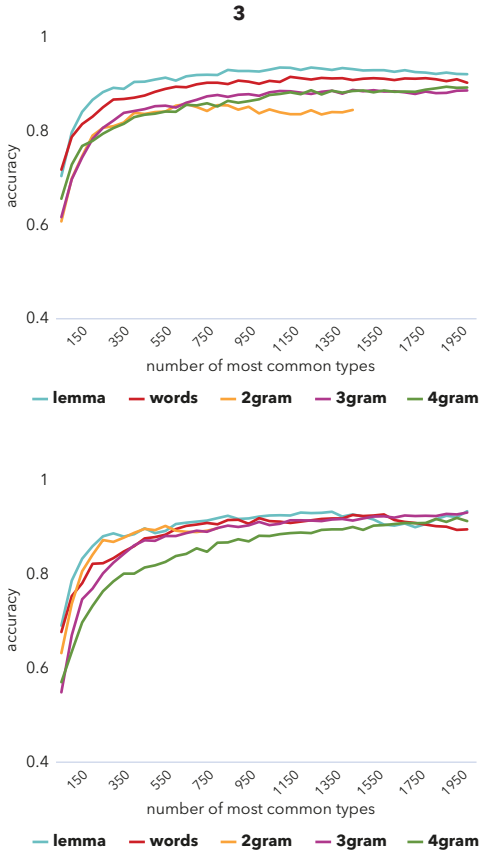


FIG. 3.3: Cross-validation results for lexicon-based models (50, 100, 150, …, 2000 most frequent character bigrams, character trigrams, character tetragrams, lemmata and words).



all the features within a single subcorpus, it differed vastly across the subcorpora. Although accuracy for ES1 peaked at $n \approx 200$, in the cases of CS3' and DE2', it continued to increase up to the highest values of n observed.

FIG. 3.3 also shows that across all the subcorpora, lemmata outperformed words and all of the character n-grams. In the n-gram group, character trigrams proved more accurate than both character bigrams and character tetragrams in each subcorpus.

At first glance, it may seem, then, that the most reliable models were those based on the highest values of n. However, we should be aware of the risk of overfitting: when we take a higher number of common types into account, there is more chance that the classifier will not actually recognise the peculiarities of an author's style but only respond to specific themes. An example may be found in one of the experiments I performed with the samples from Sigismund Bouška (1867–1942) and František Cajthaml-Liberté (1868–1936) where $n = 2000$. A list of the 10 most important features

(lemmata) for each of these authors (TAB. 3.9) shows that the classification was based primarily on thematic differences (Catholic themes vs. working-class themes). Had this model been applied to poems on different themes, it would probably have failed to distinguish the authors.

	Bouška	**Cajthaml-Liberté**
1	svatý (holy)	práce (work)
2	boží (godly)	černý (black)
3	nebesa (heaven)	bída (poverty)
4	jaký (which)	lid (people)
5	Kristus (Christ)	chléb (bread)
6	mluvit (to speak)	zítra (tomorrow)
7	volat (to call)	dělník (workman)
8	nebeský (heavenly)	ležet (to lie)
9	otec (father)	ruch (tumult)
10	klín (lap)	již (already)

TAB. 3.9: Most important features (lemmata) for the classification of works by Sigismund Bouška and František Cajthaml-Liberté (n = 2000).

I set out to test this hypothesis with another experiment. The goal was to assess how accurately lyric poems were classified by classifiers trained with narrative poems and *vice versa*. (I assumed here that literary genre had a similar effect to theme.) For this purpose, I selected five authors from CS2' who had written narrative and lyric poems. These individuals were Svatopluk Čech, Eliška Krásnohorská, Rudolf Pokorný, Ladislav Quis and Jaroslav Vrchlický (see TAB. 3.10 for details).

Author (# of lyric samples / # of narrative samples)	**Lyric poems**	**Narrative poems**
Čech (23/20)	*Jitřní písně*; *Nové písně*	*Václav z Michalovic*; *Lešetínský kovář*; *Petrklíče*
Krásnohorská (37/25)	*Vlny v proudu*; *Letorosty*	*Vlaštovičky*; *Šumavský Robinson*; *Zvěsti a báje*
Pokorný (17/9)	*S procitlým jarem*; *Vlasti a svobodě*	*Mrtvá země*
Quis (11/10)	*Písničky*	*Hloupý Honza*; *Třešně*
Vrchlický (42/34)	*Dni a noci*; *Hořká jádra*; *È morta*	*Hilarion*; *Sfinx*; *Poutí k Eldoradu*

TAB. 3.10: Lyric samples and narrative samples selected from CS2'.

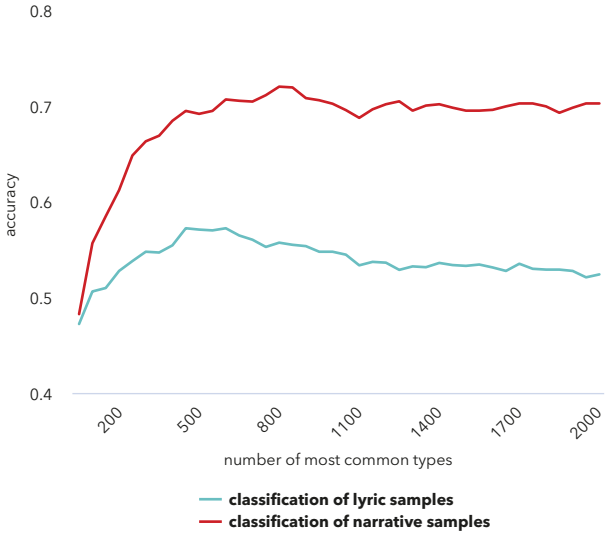


FIG. 3.4: Classification accuracy of lyric samples using models trained with narrative samples and *vice versa*.

Over 30 iterations, nine narrative samples were randomly selected for each of the five authors.[19] In each iteration, 40 different models were trained with $n \in \{50, 100, 150, \ldots, 2000\}$ and these were then used to classify nine randomly selected lyric samples by each of the authors. The entire process was repeated in order to train the models with the lyric samples and classify the narrative samples.

The results are given in FIG. 3.4. While the recognition of narrative samples generally followed the pattern seen in FIG. 3.3, the recognition of lyric samples peaked at $n = 450$ and then declined significantly. In other words, this was another case of overfitting to the training data.

On this basis, I chose the 500 most common lemmata as the optimal reference for versification-based models. At this level, accuracy had either already peaked or only limited improvements could be expected while the risk of overfitting could still be considered rather low. Notably, these 500-dimensional vectors have often been used for authorship attribution with poetic texts (e.g. Craig and Kinney 2009; Smith and Aldridge 2011). For the sake of comparison, I also included two lower levels used elsewhere including in two influential studies: $n = 150$ (Burrows 2002) and $n = 250$ (Koppel and Schler 2004).

19 Here the number of samples was made equal to that of the author with the least samples (Pokorný).

3.3.2 Results

To compare the versification-based and lemma-based models, I applied the procedure I had used with versification-based models alone. This entailed 30 iterations in which the subcorpora were each reduced to 50 samples (i.e. five authors with 10 samples each). During each iteration, I cross-validated the following:

(1) versification-based models (same feature set as in Section 3.2);
(2) lemma-based models ($n = 500$);
(3) combined models (concatenation of versification-based and lemma-based vectors).

The entire process was repeated with lemma-based models when $n = 150$ and $n = 250$.

The results are given in FIG. 3.5. They showed that:

(1) As expected (see Section 3.3.1), within lemma-based models, accuracy tended to grow as n increased.
(2) The accuracy of versification-based models was more or less stable across different samplings.
(3) In six cases (CS1 with $n = 150$, CS2 with $n \in \{150, 250\}$ and CS3 with $n \in \{150, 250, 500\}$), versification-based models outperformed lemma-based models while in the remainder, lemma-based models proved more accurate.
(4) Both versification-based and lemma-based models were outperformed by combinations of these models in the cases of CS1–3 and DE2–3; this occurred at each of the three examined levels of n (all of these differences were statistically significant at a conventional significance level $\alpha = 0.05$; see TAB. 3.11). For DE1 and ES1–ES2, however, combined models brought no improvement over lemma-based ones.

Along with the concatenation of feature spaces, I also considered how the lexicon-based model and versification-based model might work as a voting ensemble. In this scenario, there were three possible classification outputs:

(1) correct prediction (the output of both models is the same and it identifies the actual author);
(2) false prediction (the output of both models is the same and it does not identify the actual author); and
(3) ambiguous prediction (the output of one model differs from that of the other).

FIG. 3.6 shows the results of testing this approach with the same samples used in the last battery of experiments. Though this approach excluded some samples as

n	CS1	CS2	CS3	DE1	DE2	DE3	ES1	ES2
150	**$< 10^{-4}$**	**$< 10^{-4}$**	**$< 10^{-4}$**	0.3878	**$< 10^{-4}$**	**0.0013**	0.1	0.54
250	**$< 10^{-4}$**	**$< 10^{-4}$**	**$< 10^{-4}$**	0.1739	**0.0132**	**0.0347**	0.36	0.33
500	**$< 10^{-4}$**	**0.0001**	**$< 10^{-4}$**	0.3608	**0.0002**	**0.0077**	0.08	0.11

TAB. 3.11: *P*-values for the difference between lemma-based and combined models (Wilcoxon signed-rank test). Statistically significant values ($\alpha = 0.05$) appear in bold.

ambiguous, there was a significant improvement in accuracy among the samples classification of which was unequivocal (i.e. both models predicted the same author) when compared to the results of the standalone (i.e. lemma-based, versification-based and combined) models tested above (Wilcoxon signed-rank test; $\alpha = 0.05$) except in two instances. These were ES1 with $n \in \{250, 500\}$ (where there was no room to improve) and DE1 with $n = 500$.

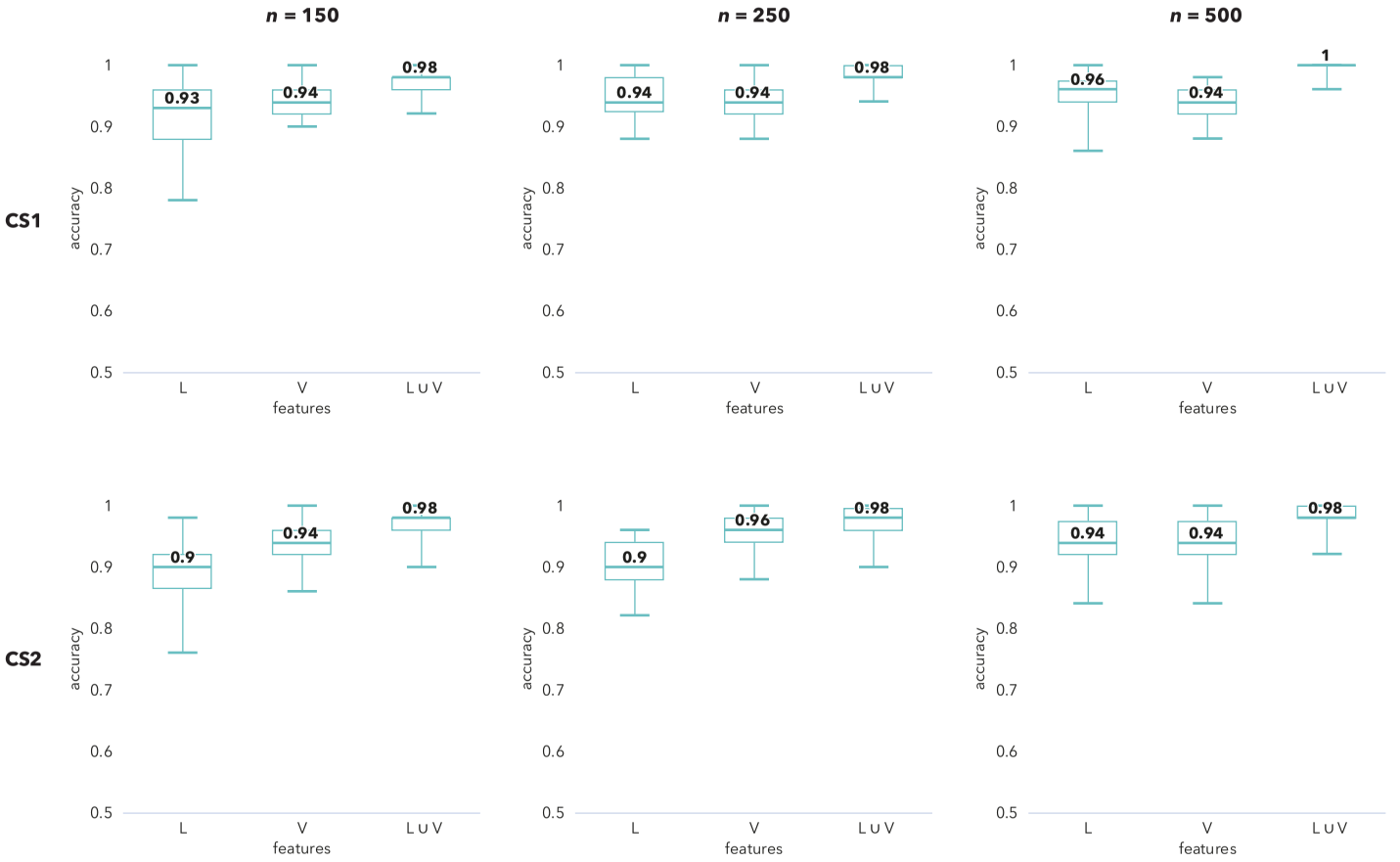
It may be objected that an approach which throws away a significant portion of samples (in the case of DE2, up to 50%) is, in fact, useless. This is a valid concern when both models are weighted equally, but it does not apply when a lemma-based model (the type that is usually more accurate) is treated as primary and the versification-based model only serves as supplementary evidence (i.e. in case of ambiguous prediction, we let the lemma-based model decide). In other words, if a lemma-based model predicts the same author as a versification-based model, the attribution is generally more reliable than one based on lemmata only.

3.4 Summary

The results presented in this chapter show that versification features are a reliable stylometric indicator. In particular, we can draw four conclusions:

(1) The accuracy of versification-based models is significantly higher than the random baseline.
(2) Versification-based models occasionally outperform lexicon-based models.
(3) Both versification-based models and lexicon-based models are usually outperformed by models combining both feature sets.
(4) If a lexicon-based model confirms the prediction of a versification-based model, the attribution is generally more reliable than one based on lexical features alone.

FIG. 3.5: Cross-validation of models based on the 150, 250 and 500 most common lemmata (L), versification features (V) and the concatenation of both feature spaces (L ∪ V); 30 iterations with random sampling.



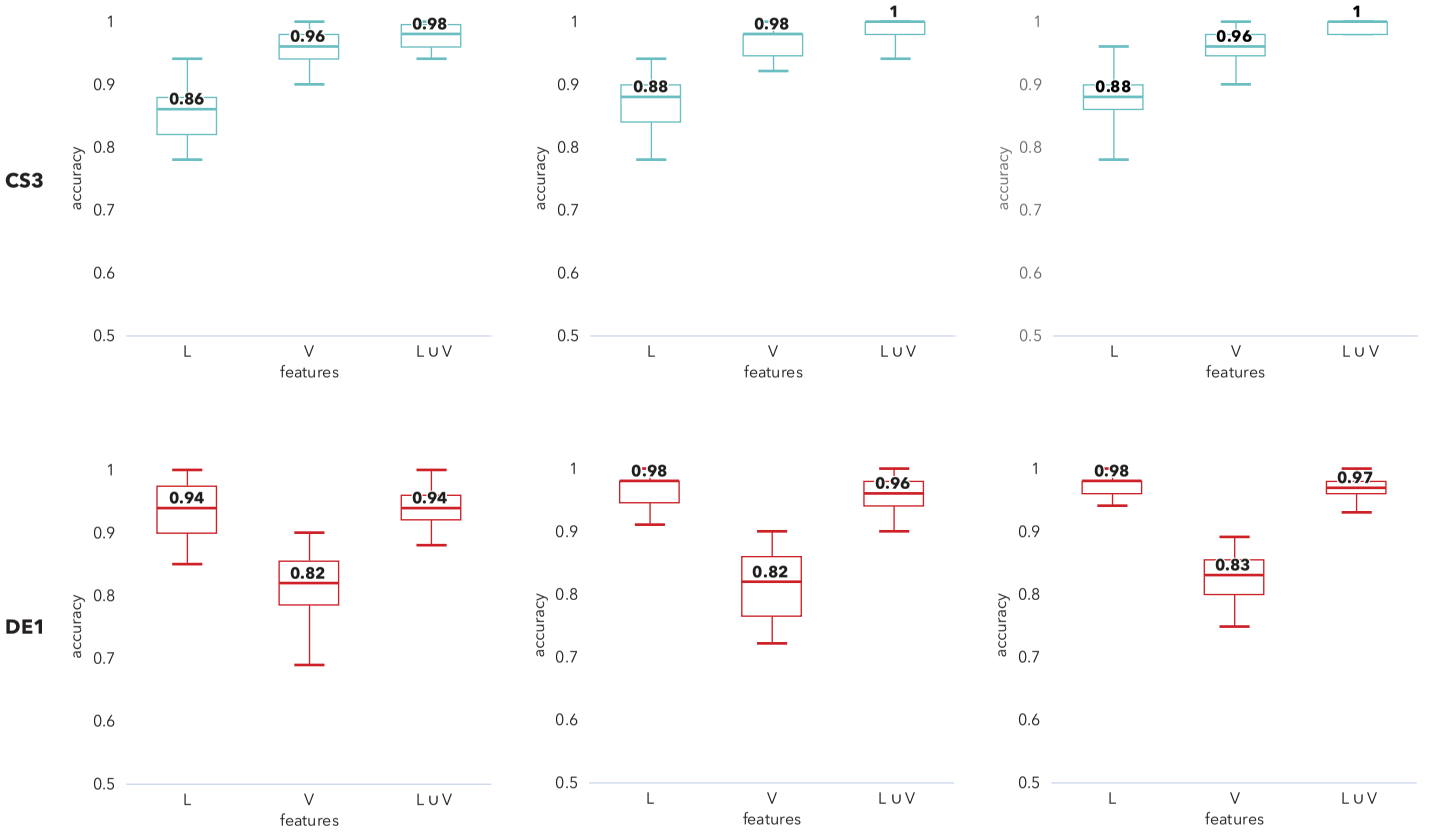
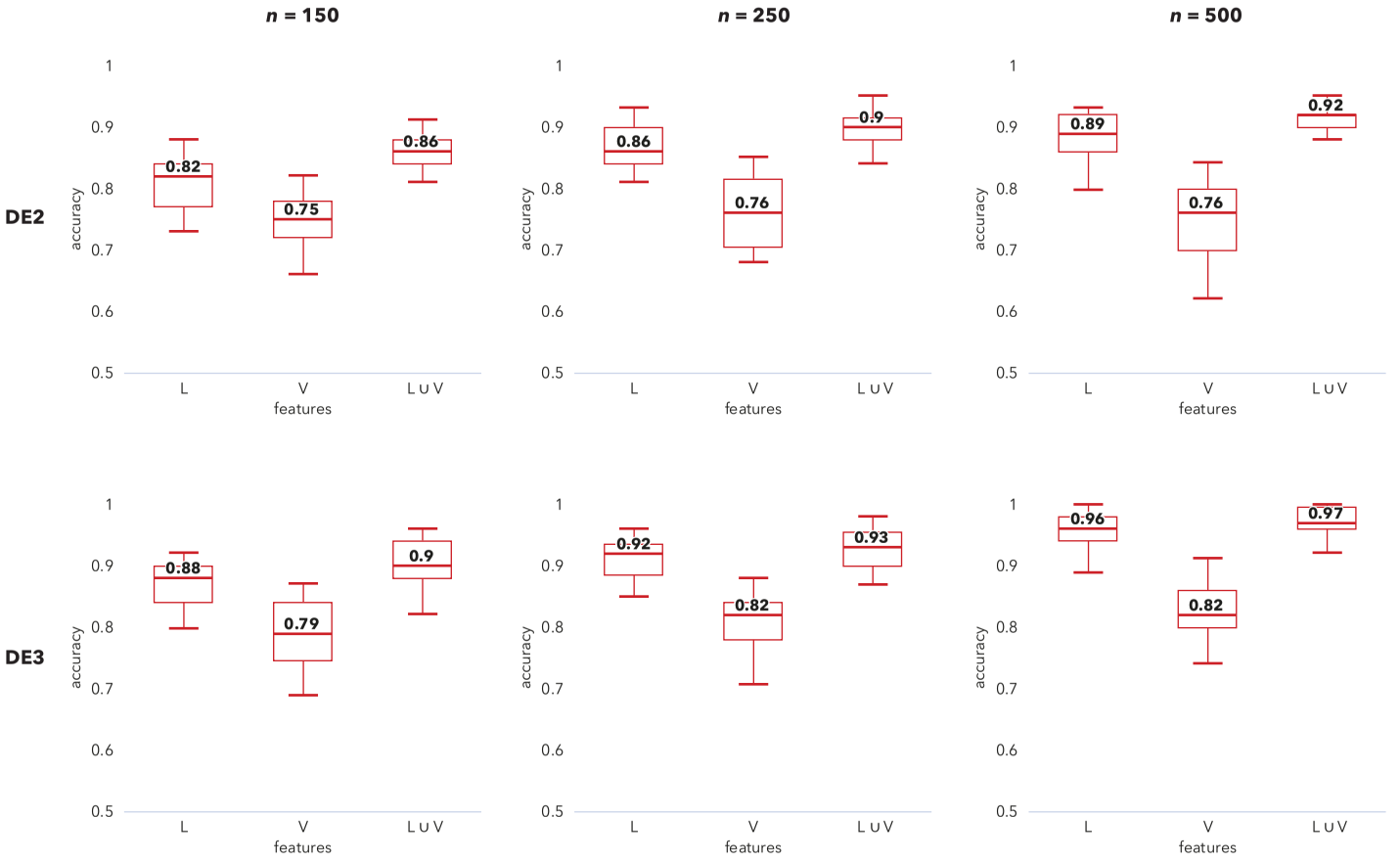

CS3
DE1
accuracy
features
L
V
L ∪ V
0.86
0.96
0.98
0.88
0.98
1
0.88
0.96
1
0.94
0.82
0.94
0.98
0.82
0.96
0.98
0.83
0.97

FIG. 3.5: Cross-validation of models based on the 150, 250 and 500 most common lemmata (L), versification features (V) and the concatenation of both feature spaces (L ∪ V); 30 iterations with random sampling.



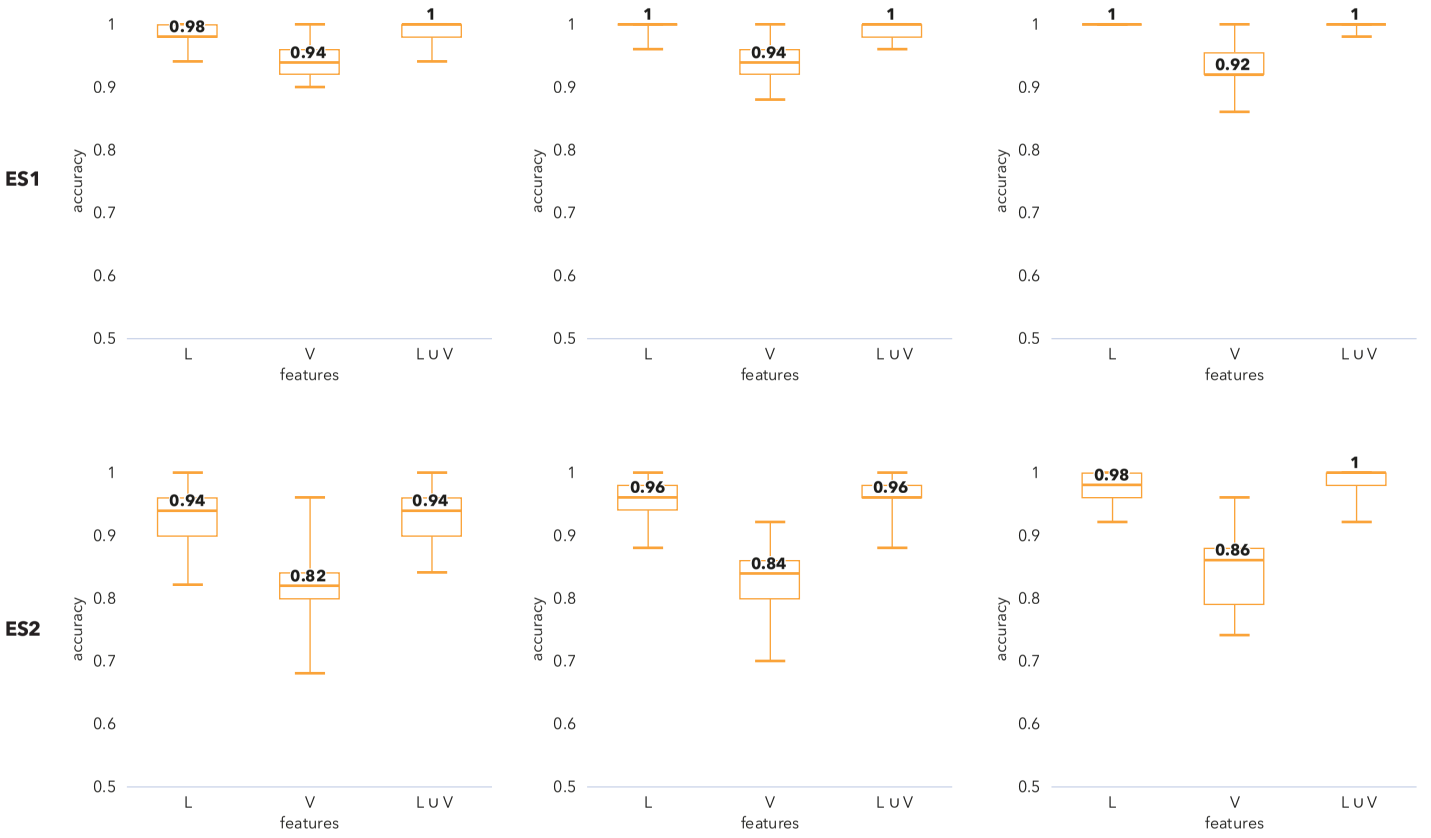

ES1
ES2
accuracy
features
L
V
L ∪ V
0.98
0.94
1
1
0.94
1
1
0.92
1
0.94
0.82
0.94
0.96
0.84
0.96
0.98
0.86
1

FIG. 3.6: Frequency of ambiguous predictions (lemma-based model predicts a different author than versification-based model) per iteration; frequency of correct predictions (both models predict the same author) within all unequivocal predictions; 30 iterations with random sampling.