Assessing the Reliability of Stress as a Feature of Authorship Attribution in Syllabic and Accentual Syllabic Verse

Petr Plecháč
Institute of Czech Literature,
Czech Academy of Sciences
plechac@ucl.cas.cz

David J. Birnbaum
University of Pittsburgh
djbpitt@pitt.edu

Abstract

This work builds on a recent study by one of the authors, which shows that statistics about versification may be used as a feature in the process of authorship attribution. One such statistic is what we have called the stress profile of a poem, a vector consisting of frequencies of stressed syllables at particular metrical positions.

Our initial hypothesis was that because syllabic versification (SV) regulates by definition the number of syllables in a line but not the distribution of stresses, it allows authors to individualize their rhythmical style much more than accentual syllabic versification (ASV), where the distribution of stresses is primarily determined by meter. For that reason, we expected the stress profile to be a more reliable indicator of authorship in Spanish SV than in Czech or German ASV. This hypothesis, however, was not supported by our analysis. For most of our samples, German ASV had lower accuracy than Spanish, which we had predicted, but, contrary to our expectations, the accuracy for Czech ASV and Spanish SV were more or less the same.

This result led us to hypothesize further that the traditional labels SV and ASV were misleading and we sought to measure the tonic entropy of our data. In this case, Spanish SV, as expected, was found to be the least tonically regular, while there was a significant difference between the two ASV systems: the values for Czech were even closer to Spanish than to the low-scoring German system. This explains why our initial grouping of Czech and German together into a single ASV category was insufficiently nuanced.

1 Introduction

In a recent article (Plecháč–Bobenhausen–Hammerich 2018) we proposed that statistics about versification may be used as a feature in the process of authorship attribution. A pilot experiment that we conducted with samples of poetry written in four different languages (Czech, German, Spanish, and English) has shown that rhythm and rhyme may be considered rather reliable indicators of authorship, in some cases outperforming even the features traditionally used for this purpose (e.g., frequencies of words, or frequencies of character *n*-grams).

In this article we further examine one of the feature sets employed in that previous paper, namely the so-called *stress profile*, a vector consisting of frequencies of stressed syllables at particular metrical positions (we refer to this as the metrical valence of the position; see Birnbaum 2018) across an entire poem or set of poems, which is widely used for formalizing rhythmical style. We proceed from a simple hypothesis: because syllabic versification regulates by definition the number of syllables in a line but not the distribution of stresses, it allows authors to individualize their rhythmical style much more than accentual syllabic (that is, *syllabotonic*) versification, where the distribution of stresses is primarily determined by meter. We thus expect the stress profile to be a more reliable indicator of authorship in Spanish syllabic versification than in Czech and German accentual syllabic versification.[1] In order to test this hypothesis we first set up attribution models with Czech, German, and Spanish samples of poetry using the stress profile as a feature set and a Support Vector Machine as a classification method. Next, we estimate the accuracy in the given languages by means of cross-validation of particular models. Finally, we compare the results to the degree of tonic regularity of the data calculated on the basis of entropy.

2 Method and data

We extract our samples from three corpora of poetry: *Corpus of Czech Verse* (Plecháč 2016; Plecháč–Kolár 2015), *Metricalizer* – corpus of German poetry (Bobenhausen–Hammerich 2015; Bobenhausen 2011), and *Corpus of Spanish Golden-Age Sonnets* (Navarro-Colorado et al. 2016; Navarro-Colorado 2015). TAB. 1 gives some basic information about these corpora.

As we intend to examine the relationship between the language of the sample set and the accuracy of the attribution, we need to control extraneous variables that might affect the result as much as possible. This in the first place requires samples to be represented by the same number of features, i.e., in the case of stress profile, to analyze lines of the same syllabic length. Since the Spanish corpus consists exclusively of 11-syllable lines (hendecasyllabs), we need to stick with this length. Fortunately, one

1 Let us emphasize that we are taking into account only strict accentual syllabic German poetry and not the rich tradition of German accentual verse (Senkungsfreiheit). Unlike in the previous article, we do not include English versification because of insufficient corresponding data.

	# of poems	# of lines	time span	website
CS	~ 75,000	~ 2,500,000	late 18th to early 20th century	http://versologie.cz
DE	~ 50,000	~ 1,700,000	16th to early 20th century	http://metricalizer.de
ES	~ 5,000	~ 70,000	16th to 17th century	https://github.com/bncolorado/CorpusSonetosSigloDeOro

TAB. 1: Basic information on corpora

of the corresponding accentual syllabic meters, feminine iambic pentameter, is very frequent in both remaining corpora, and we therefore have enough comparable data to work with.

Another important variable, as has been pointed out many times (e.g., Eder 2017), is the time span between analyzed samples. For example, it is self-evident that no matter what method is used, distinguishing Goethe (*1749) from Schiller (*1759) is a much more complicated task than distinguishing Cervantes (*1547) from García Lorca (*1898).

Given that, we extract the sample sets from each corpus in the following way:

(1) Each sample consists of 100 hendecasyllabic / feminine iambic pentameter lines written by one author. Multiple poems might be combined into a sample, and no poem contributes to more than one sample.

(2) Three sample sets are built from each corpus (es-A, es-B, es-C; de-A; de-B; de-C; cs-A, cs-B, cs-C), each set consisting of samples written by at least 5 authors born in a specified time period. We tried to keep all the periods to between 15 and 20 years long, although es-A and de-A had to be longer because of sparse data. Details about the sample sets are given in TAB. 2.[2]

Each sample was represented by a stress-profile vector consisting of 11 values corresponding to particular metrical positions. The principle of calculating the stress profile is illustrated in TAB. 3 with an example from Ludwig Tieck.

In order to consider the metrical valence of each position as a marker of equal importance, data were transformed to *z*-scores.[3]

2 We recognize that not all the relevant variables have been controlled this way. First, eras from which the data comes differ across the corpora, and one may, for instance, assume that romantic poets in general have put more effort into individualizing the rhythm of their poems than renaissance ones. Additionally, proximity in time is not the only factor here: we may assume that there was stronger mutual influence in the poetry of the Argensola brothers than, for instance, between the Austrian Betty Paoli and the German August von Platen.

3 The metrical valences of particular positions have a different scale. For example, the stressing of some positions varies across samples by tens of percents (this concerns mainly strong positions), stressing of others varies only in units of percents (this concerns mainly weak positions); cf. Sect. 3. Without any normalization of the data, SVM would treat positions with higher variance as more important. To avoid this, we transform the frequencies to *z*-scores: $z_S = (f_S - \mu) / \sigma$, where f_S is the original frequency of stressing of the position in sample S, μ stands for mean frequency at the given position across all samples

es-A (1515-1549)	**es-B (1550-1569)**	**es-C (1570-1589)**
de Acuña, Hernando (10)	de Arguijo, Juan (8)	de Molina, Tirso (7)
de Cervantes, Miguel (9)	de Argensola, Lupercio (7)	de Quevedo, Francisco (64)
de Cetina,Gutierre (31)	de Argensola,Bartolomé (19)	de Rojas, Pedro Soto (15)
de Herrera, Fernando (40)	de Góngora, Luis (14)	de Tassis y Peralta, Juan (25)
de la Torre, Francisco (9)	de Vega, Lope (168)	de Ulloa, Luis (13)
de-A (1770–1794)	**de-B (1795–1814)**	**de-C (1815–1834)**
Chamisso, Adelbert von (9)	Droste-Hülshoff, Annette von (7)	Geibel, Emanuel (16)
Eichendorff, Joseph von (7)	Grün, Anastasius (13)	Heyse, Paul (10)
Grillparzer, Franz (13)	Lenau, Nikolaus (15)	Keller, Gottfried (11)
Schlegel, Friedrich (9)	Platen, August von (7)	Lingg, Hermann von (7)
Tieck, Ludwig (9)	Paoli, Betty (7)	Otto, Louise (12)
cs-A (1825–1839)	**cs-B (1840–1854)**	**cs-C (1855–1869)**
Hálek, Vítězslav (18)	Čech, Svatopluk (58)	Kaminský, Bohdan (51)
Heyduk, Adolf (7)	Krásnohorská, Eliška (48)	Klášterský, Antonín (85)
Neruda, Jan (9)	Pokorný, Rudolf (13)	Kvapil, František (17)
Pfleger Moravský, Gustav (77)	Sládek, Josef Václav (29)	Mužík, Augustin Eugen (35)
Šolc, Václav (7)	Vrchlický, Jaroslav (417)	Škampa, Alois (24)

TAB. 2: Sample sets extracted from each corpus. The number enclosed in parentheses indicates the number of samples by the given author.

	DISTRIBUTION OF STRESSED SYLLABLES (0: unstressed; 1: stressed)										
1. Viel Wunder in der Dichtkunst Garten blühen.	0	1	0	0	0	1	0	1	0	1	0
2. Es drohet als verschlingend Ungeheuer	0	1	0	0	0	1	0	1	0	1	0
3. Allem, was lebt, das hunger-grimme Feuer,	1	0	0	1	0	1	0	1	0	1	0
…	…	…	…	…	…	…	…	…	…	…	…
100. Der Sonnenschein, des blauen Himmels Helle;	0	1	0	1	0	1	0	1	0	1	0
SUM	13	79	3	73	3	73	6	74	2	94	2
RELATIVE FREQUENCY (STRESS PROFILE)	0.13	0.79	0.03	0.73	0.03	0.73	0.06	0.74	0.02	0.94	0.02

TAB. 3: Example of stress profile calculation (Ludwig Tieck‘s sample)

In each language we randomly selected 7 100-line samples by each author, so that each sample set was reduced to 35 samples. We then performed a leave-one-out cross-validation of Support Vector Machine driven authorship attribution on:

(1) each sample set (A; B; C)

(2) each consecutive pair of sample sets of one language merged together (A∪B; B∪C)

(3) all the samples of one language merged together (A∪B∪C)

(4) random selections of 4 authors from each consecutive pair of sample sets of one language merged together (A∪B(4); B∪C(4))

and σ stands for its standard deviation. This results in a distribution with mean value = 0 and standard deviation = 1.

(5) random selections of 5 authors from each consecutive pair of sample sets of one language merged together (A∪B(5); B∪C(5))

(6) random selections of 6 authors from each consecutive pair of sample sets of one language merged together (A∪B(6); B∪C(6))

(7) random selections of 4 authors from all the samples of one language merged together (A∪B∪C(4))

(8) random selections of 5 authors from all the samples of one language merged together (A∪B∪C(5))

(9) random selections of 6 authors from all the samples of one language merged together (A∪B∪C(6))

In order to obtain more representative results the entire procedure was repeated 500 times, with new random selections of both authors and their samples in each iteration.

3 Results

The experiments described above produced 15 sets of 500 accuracy estimations for each language. FIG. 1 summarizes the results. We also report the value of a random baseline calculated as

$$\text{random baseline} = \sum_{i=1}^{N} \left(\frac{a_i}{X}\right)^2$$

where N is the number of authors in a sample set, X is the number of samples and a_i is the number of samples written by author i.

While all the mean accuracy estimations are above random baseline values, there are significant differences between particular languages. Only the results for single sample sets (A; B; C) are consistent with our expectations based on the above-mentioned hypothesis: we attained the best scores for syllabic Spanish, and low scores for accentual syllabic German, with accentual syllabic Czech standing in the middle. In all other results, however, the situation is quite different: in agreement with our expectations, accentual syllabic German has lower values than Spanish, but, contrary to our expectations, the results for accentual syllabic Czech and syllabic Spanish are more or less the same.

The direct association between the type of versification and the accuracy of stress-profile-based classification thus has not been confirmed. On the other hand, we must keep in mind that the traditional labels “syllabic” and “accentual syllabic” are extremely simplifying. One needs to consider typology of versification as a continuous, rather than a discrete, system. For example, it has been shown that there is a certain degree of tonic (accentual) regularity in Spanish hendecasyllabs (e.g. Piera 1980) and that German versification exhibits a higher degree of tonic regularity than is the case with Czech (e.g. Levý 2011). In order to explore the consequences of those differences, in the following section we aim to measure the degree of tonic regularity of our data.

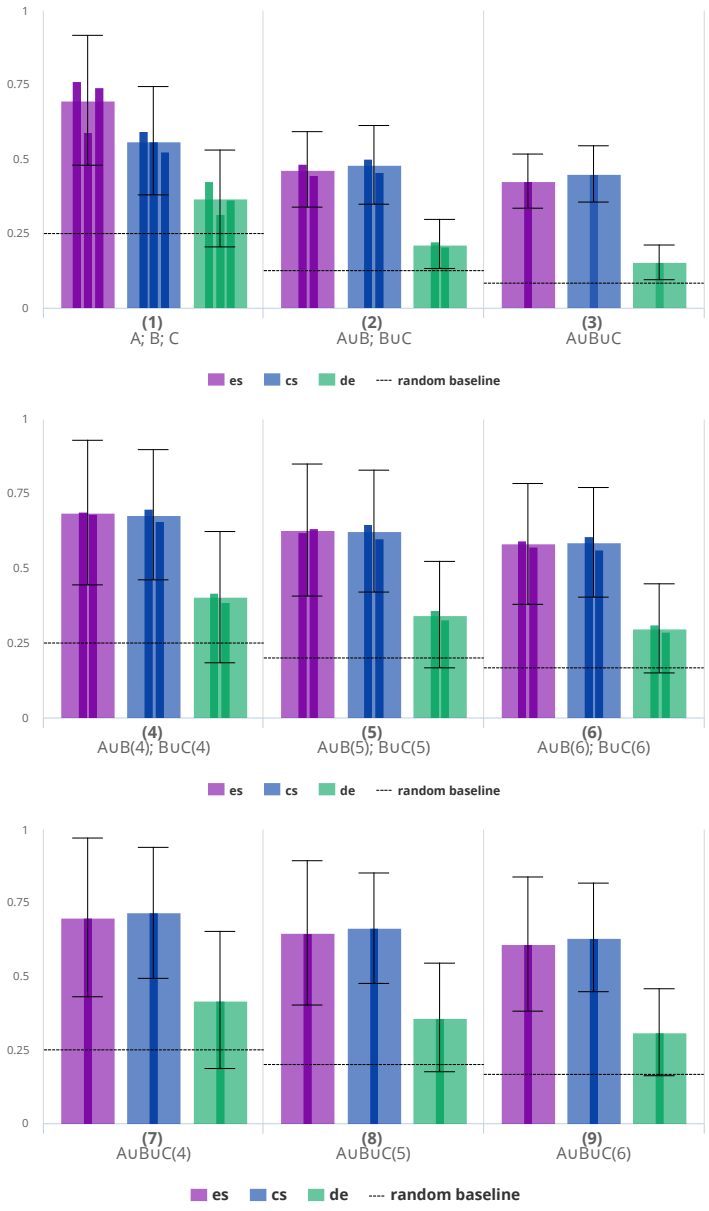

FIG. 1: Mean accuracy estimations. Thin columns correspond to the mean values for particular sample sets (A; B; C), merged pairs of sample sets (AUB; BUC), all samples merged together (AUBUC), random selections from merged pairs of sample sets ($AUB_{(4\text{-}6)}$; $BUC_{(4\text{-}6)}$), and random selections of all samples merged together ($AUBUC_{(4\text{-}6)}$). Thick columns correspond to the mean values of thin ones; whiskers denotes the 95% confidence interval of all the accuracy estimations under the given thick column.

4 Tonic regularity

In order to estimate the degree of tonic regularity we modify the procedure proposed in Dobritsyn 2016 and measure the entropy of what may be called *rhythmic types*. As a rhythmic type we denote the bit string representing the distribution of stressed syllables in a particular line (e.g., the example in TAB. 3 provides three different rhythmic types: 01000101010 (line 1 and 2), 01000101010 (line 3), and 01010101010 (line 100)). The entropy of rhythmic types in particular samples is calculated as:

$$S = -\sum_i \frac{n_i}{N} \log \frac{n_i}{N}$$

where n_i is the number of occurrences of rhythmic type i and N is the number of lines.

FIG. 2 gives the entropy of rhythmic types in sample sets A, B, and C in particular languages, as well as the same value for all the lines of a given meter in each entire corpus (columns ES, CS, DE).

Values across the sample sets as well as the entire corpus of one language vary in very small intervals, which suggests that we're actually capturing an important and constant feature of the versification systems in question. Syllabic Spanish, as expected, was found to be the least tonically regular. Interestingly enough, though, there is a significant difference between the two accentual syllabic versification systems: values for Czech are even closer to Spanish than to the low-scoring German. We may roughly identify which metrical positions contribute the most to the regularity/irregularity by plotting the stress profiles of our data on a chart (FIG. 3).

Once again, values across sample sets and corpora are more or less the same in each language. There is an iambic tendency in all of the data, which is most evident in German: the metrical valence of all the odd positions (weak) except for the first one is very close to zero, for even positions (strong) it is about 0.75, and the metrical valence of the penultimate almost 1. Spanish exhibits a noticeably weaker iambic tendency in the first four positions and the eighth position, but otherwise is comparable to German even in the highly regular ending; the metrical valence is 0 for the penultimate and 1 for the last position. Czech data exhibit the highest degree of irregularity at the very beginning and at the end of the lines: there are many more Czech lines with a stressed syllable in the first position than in German (even slightly more than in Spanish),[4] and, unlike in the

4 Czech has a fixed initial stress and is thus usually mentioned as having a natural propensity toward trochaic onset. On the other hand, there is generally a high degree of correspondence between verse-line boundaries and clause boundaries and many Czech connectors (conjunctions, pronouns, and others) are often realized as unstressed monosyllabic words. Trochees thus enforce rather simple syntax, while iambs (when trochaic onsets are not preferred) usually tend toward complex syntactic structures and complex interclause relations (cf. Červenka 2006: 81–111). Initial stress should nonetheless be taken into consideration as a factor in assessing the difference between the metrical valence of the first position in Czech and the German iamb, as there are sentence-initial words that naturally tend to bear stress on the first syllable.

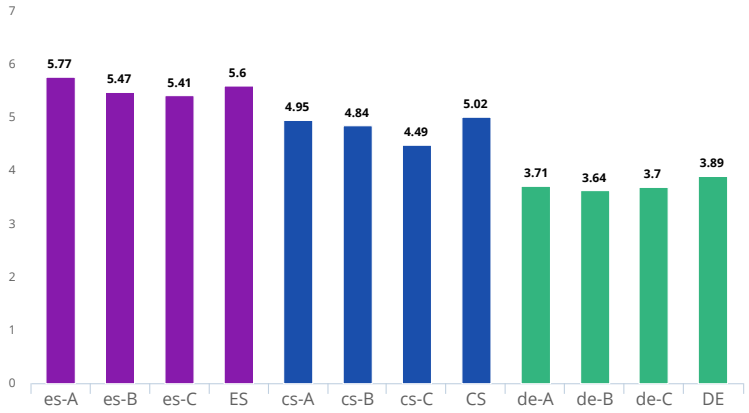

FIG. 2: Entropy of rhythmic types in particular sample sets and in the entire corpora (ES, CS, DE)

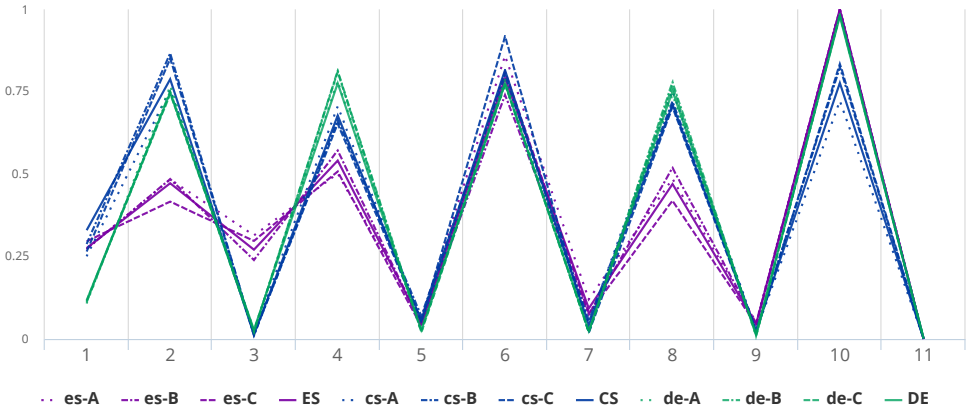

FIG. 3: Frequency of stressed syllables at particular metrical positions in particular sample sets and in the entire corpora (ES, CS, DE)

other two languages, the metrical valence of the penultimate does not exceed the metrical valences of other even positions. In Czech there is also a slightly higher variability in the fourth position than in German, and a slightly lower variability in the second position.[5]

5 The latter may come from the phonetic rules of the language itself. Unlike German and Spanish, it is not possible to have two adjacent unstressed syllables at the beginning of the line in Czech. The sum of the first and the second values in each Czech series thus inevitably has to be greater than 1.

5 Conclusion

We have proceeded from a simple hypothesis that Spanish syllabic poetry should perform better in stress-profile-based authorship attribution than Czech and German accentual syllabic poetry. This, however, was not confirmed with our data. While German versification showed the weakest performance in all the experiments conducted, Spanish and Czech yielded more or less the same scores in most of the experiments. The problem seems to be in the initial assumption: it has been shown that all the data exhibit tonic regularity to some extent, and that Czech is in this respect closer to less regular Spanish than to more regular German. We thus may conclude that the degree of tonic regularity of the versification affects the accuracy of stress-profile-based attribution to some extent; it is not a direct correlation, but, rather, a question of a certain limit where the meter ceases to offer authors enough space to individualize the rhythm of their poems. As a consequence, we may expect that in order to achieve acceptable reliability, versification-based attribution tasks dealing with German poems (or poems from other strongly regulated versification systems) would require more features than languages with less regular versification systems.

Acknowledgment

The authors were supported by the Czech Science Foundation, project GA17-01723S (Stylometric Analysis of Poetic Texts).

References

Birnbaum, D. J. (2018, November 5). Strong and weak metrical positions. Retrieved from <http://poetry.obdurodon.org/metrical-analysis.xhtml>.

Bobenhausen, K. (2011). The Metricalizer: Automated Metrical Markup for German Poetry. In C. Küper (Ed.), *Current Trends in Metrical Analysis* (119–131). Frankfurt am Main: Peter Lang.

Bobenhausen, K. – Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer². *Langages* 199, 67–87.

Červenka, M. (2006). *Kapitoly o českém verši* [Chapters on Czech Versification]. Praha: Karolinum.

Dobritsyn, A. (2016). Rhythmic entropy as a measure of rhythmic diversity (The example of the Russian iambic tetrameter). *Studia Metrica et Poetica* 3(1), 33–52.

Eder, M. (2017). Short samples in authorship attribution: A new approach. In R. L. Lewis et al. (Eds.), *Digital Humanities 2017: Conference Abstracts* (221–224). Montreal: McGill University.

Levý, J. (2011). *The Art of Translation*. Amsterdam: John Benjamins.

Navarro-Colorado, B. (2015). A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (105–113). Denver (Co): NAACL.

Navarro-Colorado, B. – Ribes-Lafoz, M. – Sánchez, N. (2016). Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (4360–4364). Portorož: ELRA.

Piera, C. J. (1980). *Spanish Verse and the Theory of Meter* (Ph.D. thesis). Los Angeles: University of California.

Plecháč, P. (2016). Czech verse processing system KVĚTA: Phonetic and metrical components. *Glottotheory* 7(2), 159–174.

Plecháč, P. – Bobenhausen, K. – Hammerich, B. (2018). Versification and authorship attribution: Pilot study on Czech, German, Spanish, and English poetry. *Studia Metrica et Poetica* 5(2), 29–54.

Plecháč, P. – Kolár, R. (2015). The Corpus of Czech Verse. *Studia Metrica et Poetica* 2(1), 107–118.