http://doi.org/10.15446/lthc.v25n2.108502

Rhythm-Based Authorship Recognition in Syllabic and Accentual-Syllabic Verse

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

David J. Birnbaum
University of Pittsburgh, Pittsburgh, United States
djbpitt@gmail.com

This contribution explores the extent to which rhythm-based features of poetic texts can contribute meaningfully to authorship recognition. We show that, although a binary categorization of languages as syllabic *vs*. accentual-syllabic fails to fully explain the differences. However, once we formalize accentual regularity as a *continuum*, our analysis shows that authorship attribution results improve as we move from the most to the least accentually regular languages. This result supports our hypothesis that accentual regularity is an inhibiting factor in authorship attribution as long as accentual regularity is understood as a continuous property.

Keywords: poetry; authorship recognition; poetic rhythm; versification.

Cómo citar este artículo (MLA): Plecháč, Petr & David J. Birnbaum. "Rhythm-based authorship recognition in syllabic and accentual-syllabic verse". *Literatura: teoría, historia, crítica*, vol. 25, núm. 2, 2023, págs. 184-190.

Artículo original. Recibido: 15/01/23; aceptado: 12/04/2023. Publicado en línea: 01/07/2023.

Reconocimiento de autoría basado en el ritmo en verso silábico y silábico-acentual

Este trabajo explora en qué medida las características rítmicas de los textos poéticos pueden contribuir al reconocimiento de la autoría. Se demuestra, en primer lugar, que una categorización de las lenguas como silábicas o silábico-acentuales no explica completamente las diferencias. Sin embargo, nuestros análisis muestran que, si la regularidad acentual se formaliza como una continuidad, el reconocimiento de autoría mejora cuando pasamos de las lenguas más regularmente acentuadas a las menos. Los resultados apoyan la hipótesis según la cual la regularidad acentual es un factor inhibidor en el reconocimiento de autoría si se le considera como una característica continua.

Palabras clave: poesía; reconocimiento de autoría; ritmo poético; versificación.

Reconhecimento da autoria baseado no ritmo em verso silábico e acentual-silábico

Esta contribuição explora em que medida as características rítmicas dos textos poéticos podem contribuir de forma significativa para o reconhecimento da autoria. Mostramos que embora uma categorização binária das línguas como silábicas *versus* acentual-silábicas não consiga explicar completamente as diferenças, uma vez que formalizamos a regularidade acentual como um *continuum*, a nossa análise mostra que os resultados da atribuição de autoria melhoram à medida que passamos das línguas mais regulares acentualmente para as menos regulares. Este resultado apoia a nossa hipótese de que a regularidade acentual é um fator inibidor na atribuição da autoria, desde que a regularidade acentual seja entendida como uma propriedade contínua.

Palavras-chave: poesia; reconhecimento da autoria; ritmo poético; versificação.

Introduction

In recent years, the use of rhythm-based features in authorship recognition of poetic texts has been studied thoroughly in several languages, including Czech, German, Spanish (Plecháč; Plecháč & Birnbaum), Portuguese (Mittmann), Latin (Nagy), Old English (Neidorf et al.), and Russian (Šeļa; Orekhov). In this contribution we expand our analysis of the relationship between versification type and authorship recognition accuracy provided in Plecháč & Birnbaum by performing an experiment with poetic texts in six languages: Czech (cs), German (de), English (en), Spanish (es), Italian (it), and Russian (ru).

Method

From each *corpus* we extract one or more subcorpora containing 11-syllable lines (except for en, where we use 10-syllable iambic pentameter) by five authors born in a specific time span (see Table 1). Ten 100-line samples are drawn at random for each author and each sample is represented by bitstrings that encode the stressed and unstressed syllables in particular lines (*e.g.*, *The curfew tolls the knell of parting day* ~ 0101010101). Leave-one-out cross-validation is performed to evaluate Support Vector Machine models (linear kernel) for bitstring-based authorship recognition. The entire procedure is repeated 10 times, resulting in 10 accuracy estimations for each sub*corpus*.

Table 1. List of subcorpora[1]

sub*corpus*	**time span**	**authors**
cs1	*1833–1838*	*V. Hálek, A. Heyduk, J. Neruda, G. Pfleger Moravský, V. Šolc*
cs2	*1841–1853*	*S. Čech, E. Krásnohorská, J. V. Sládek, J. Vrchlický, J. Zeyer*
cs3	*1854–1861*	*B. Kaminský, K. Kučera, F. Kvapil, E. A. Mužík, A. Škampa*
de1	*1772–1802*	*A. von Chamisso, F. Grillparzer, N. Lenau, F. Schlegel, L. Tieck*
de2	*1806–1830*	*E. Geibel, A. Grün, P. Heyse, G. Keller, L. Otto*
en1	*1840–1865*	*A. Bierce, T. Hardy, A. Lang, O. Wilde, W. B. Yeats*
es1	*1490–1591*	*H. de Acunya, F. de Borja, J. Boscan, G. de Cetina, F. de La Torre*
es2	*1534–1562*	*B. Argensola, M. de Cervantes y Saavedra, L. de Góngora, F. de Herrera, L. de Vega*
es3	*1580–1603*	*G. Bocangel y Unzueta, F. de Quevedo, P. Soto de Rojas, J. de Tassis y Peralta, L. de Ulloa y Pereira*
it1	*1775–1814*	*G. Giusti, A. Guadagnoli, G. Leopardi, C. Porta, G. Prati*
ru1	*1783–1821*	*Y. Lermontov, A. N. Majkov, A. S. Pushkin, M. A. K. Tolstoj, V. A. Zhukovskij*

[1] Made by the authors.

Results

Figure 1 shows the results. All subcorpora significantly outperform the random baseline (0.2), yet there are substantial differences across languages.[2] We hypothesize that one of the key factors may be the versification type: while in syllabic versification only the number of syllables in a line is constrained and the distribution of stress is left to the author's preferences, in accentual-syllabic versification both syllable count and stress placement are subject to constraints, which leaves considerably less space for authors to individualize their rhythm.[3]

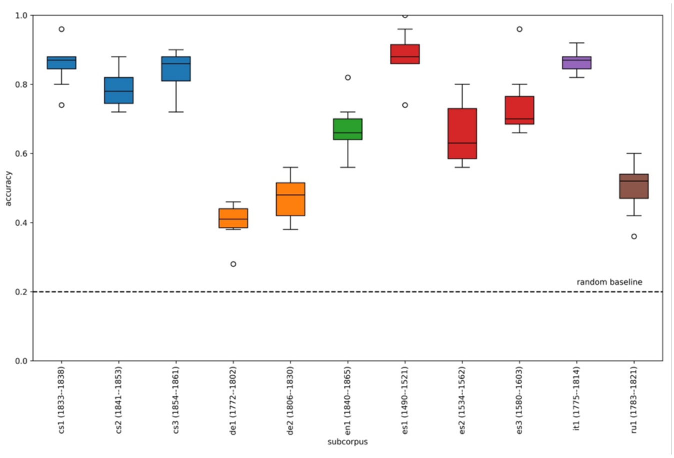

Figure 1. Rhythm-based authorship recognition accuracy; 30 random samplings per sub*corpus*; leave-one-out cross-validation; linear SVM. Boxplots are constructed in a following way: the box gives the interquartile range (IQR) of the 30 samplings; the horizontal line gives the median value; whiskers are plotted at the 1.5 IQR values; data beyond the whiskers are plotted as individual points.

[2] Within the same language subcorpora, there is only one significant difference, namely the high accuracy of ES1 as compared to ES2 (Mann-Whitney $U = 97.5$, $n_1 = n_2 = 10$, $P < 0.05$) and ES3 (Mann-Whitney $U = 89.5$, $n_1 = n_2 = 10$, $P < 0.05$). We cede the possible explanation to the experts in Spanish literature.

[3] Accentual-syllabic versification can be explored not only according to the distribution of stressed syllables, but also with respect to the position of word boundaries (see, for example, the discussion of the particular status of monosyllabic words in Russian iambic verse in Pilshchikov and Polilova). In this study we do not take word boundaries into account, regarding them as auxiliary to stress, and therefore more profitably explored after we have first understood what stress placement alone is able to contribute to authorship attribution.

Interpretation

At first glance, this hypothesis fails to explain the differences. Accentual-syllabic CS ranks among the best scoring languages and accentual-syllabic EN outperforms part of syllabic ES. We need, however, to keep in mind that the traditional categorical notion of versification types is misleading—it is, rather, a continuous scale (*cf.* Gasparov). To formalize the degree of accentual regularity, we measure the entropy of bitstrings in particular subcorpora:

$$S = -\sum_{i}^{n} f(\text{bitstring}_i) \log f\,(\text{bitstring}_i)$$

As Figure 2 shows, there is some sort of association between accentual regularity and the accuracy of authorship recognition as we proceed from strictly organized RU and DE; across semi-organized CS, EN, and ES; to loose IT, the accuracy tends to increase (linear regression, $R^2 = 0.5$).

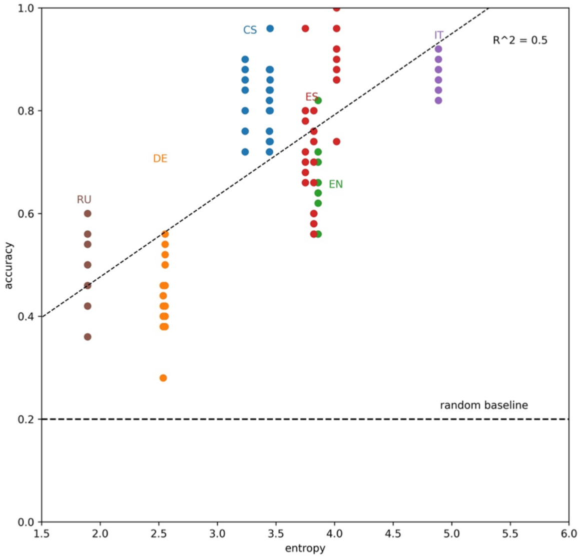

Figure 2. Relationship between rhythm-based authorship recognition accuracy and bitstrings entropy; linear regression ($R^2 = 0.5$)

Conclusions

The experiment described here yields two types of results. The first, which extends our earlier work in Plecháč & Birnbaum, is a demonstration that rhythmic organization (and, specifically, accentual regularity) can function as a meaningful feature for authorship recognition. The second is our observation that the categorical identification of verse traditions as either syllabic or accentual-syllabic masks the actually continuous nature of the feature, and we propose a formula that remedies that limitation by quantifying the accentual regularity of verse corpora as a continuous property.

Cited works

Gasparov, Mikhail, and Marina Tarlinskaja. "A Probability Model of Verse (English, Latin, French, Italian, Spanish, Portuguese)". *Style*, vol. 21, no. 3, 1987, pages 322-358.

Mittmann, Adiel, et al. "What rhythmic signature says about poetic corpora". *Quantitative Approaches to Versification*. Edited by Petr Plecháč et al. Prague, ICL, 2019, pages 153-–172.

Nagy, Benjamin. "Metre as a stylometric feature in Latin hexameter poetry". *Digital Scholarship in the Humanities*, vol. 36, no. 4, 2021, pages 999–1012. DOI: https://doi.org/10.1093/llc/fqaa043

Neidorf, Leonard, et al. "Large-scale quantitative profiling of the Old English verse tradition". *Nature Human Behaviour*, vol. 3, 2019, pages 560–567. DOI: https://doi.org/10.1038/s41562-019-0570-1

Orekhov, Boris. "Mikrodiakhroniia stikhovedcheskikh parametrov u russkikh poetov". *VAProsy Iazykoznaniia*. Edited by A. A. Kibrik et al. Moscow, Buki Vedi 2020. pages 161–164.

Pilshchikov, Igor, and Vera Polilova. "Quantitative Approaches to Versification: Conference report (June 24–26, 2019, Prague, Czech Republic)". *Studia Metrica et Poetica*, vol. 7, no. 1, 2020, pages 116–137. DOI: https://doi.org/10.12697/smp.2020.7.1.05

Plecháč, Petr. *Versification and Authorship Attribution*. Prague, Karolinum/ICL, 2021. DOI: https://doi.org/10.14712/9788024648903

Plecháč, Petr, and David J. Birnbaum. "Assessing the reliability of stress as a feature of authorship attribution in syllabic and accentual syllabic Verse".

Quantitative Approaches to Versification. Edited by Petr Plecháč et al. Prague, ICL, 2019, pages 201–210.

Šeļa, Artjoms, et al. "Fenomen Batenkova i problema verifikacii avtorstva". *Acta Slavica Estonica*, vol. 12, 2020, pages. 131–165.

Acknowledgments

The study was supported by the Czech Science Foundation, project GA23-07727S (European Poetry: Distant Reading).

Data and code are available at https://github.com/versotym/Versification-Type-Authorship.

Data sources:

CS: *Corpus* of Czech Verse (https://github.com/versotym/*corpus*czechverse)

DE: Metricalizer (https://metricalizer.de/en/)

EN: Gutenberg English Poetry *Corpus* (https://doi.org/10.3389/fdigh.2018.00005)

ES: *Corpus* of Spanish Golden-Age Sonnets (https://github.com/bncolorado/*Corpus*SonetosSigloDeOro)

IT: Biblioteca Italiana (http://www.bibliotecaitaliana.it/)

RU: Russian Poetry *Corpus* (https://ruscorpora.ru/en/page/*corpus*-poetic/)

About the authors

Petr Plecháč is head of the Versification Research Group at the Institute of Czech Literature, Czech Academy of Sciences. He received PhDs in Literary Theory and Mathematical Linguistics from Palacký University Olomouc and Charles University in Prague, respectively. His main areas of interest are the quantitative analysis of poetic texts and the problems of authorship recognition.

David J. Birnbaum is Professor Emeritus in the Department of Slavic Languages and Literatures from the University of Pittsburgh. He earned his PhD in Slavic Languages and Literatures, with a specialization in Slavic Linguistics, at Harvard University. His main areas of research are the use of computational methods in textual studies (especially digital editions) and the exploration of formal features of Russian verse.