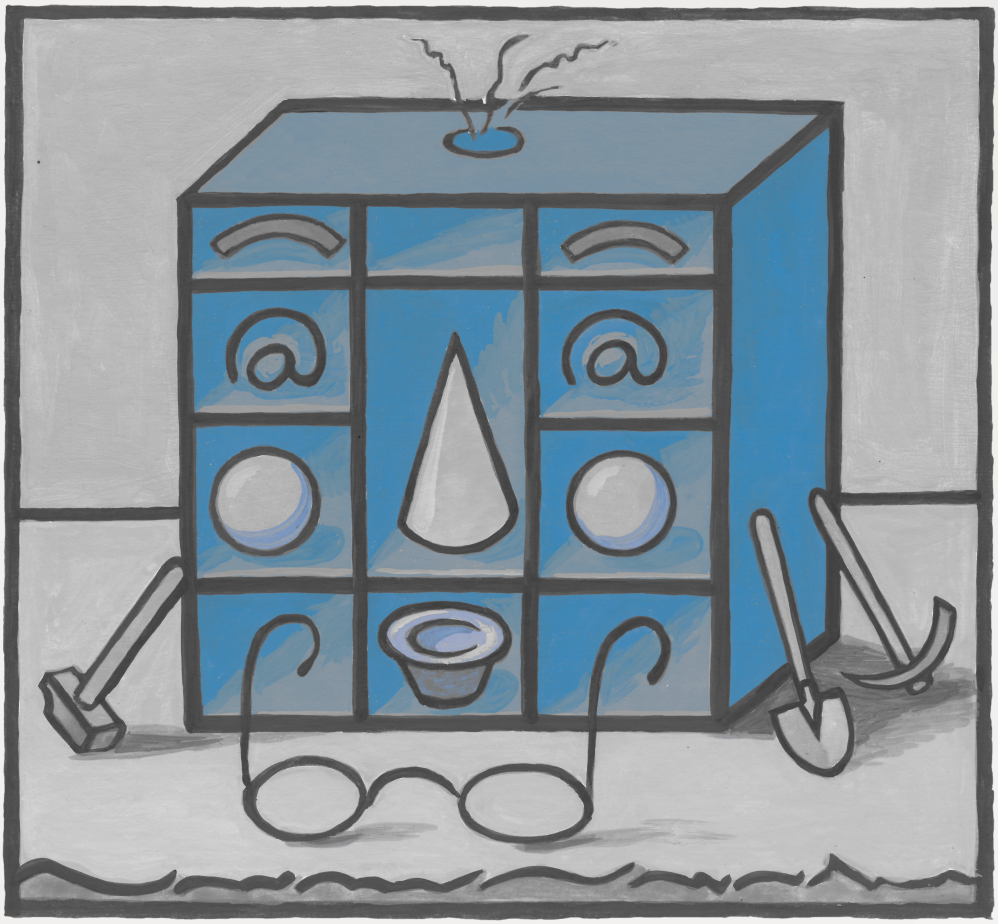
TACKLING THE TOOLKIT

Plotting Poetry through Computational Literary Studies



edited by
Petr Plecháč
Robert Kolár
Anne-Sophie Bories
Jakub Říha

Institute of Czech Literature
of the Czech Academy of Sciences

TACKLING THE TOOLKIT

Plotting Poetry through Computational Literary Studies

Editors:

Petr Plecháč (Institute of Czech Literature of the Czech Academy of Sciences)
Robert Kolár (Institute of Czech Literature of the Czech Academy of Sciences)
Anne-Sophie Bories (University of Basel)
Jakub Říha (Institute of Czech Literature of the Czech Academy of Sciences)

Each paper in this volume was assigned two reviewers and has undergone a double-blind peer review. We are grateful to scholars from the following institutions for their kind help with the review process: *Comenius University in Bratislava (SK); Charles University (CZ); Dartmouth College (US); Estonian Literary Museum (EE); Institute of Computational Linguistics "Antonio Zampolli" (IT); Institute of Linguistics, Russian Academy of Sciences (RU); Institute of Polish Language, Polish Academy of Sciences (PL); Moscow State University (RU); Mount Allison University (CA); National University of Distance Education (ES); Novetta (US); Tufts University (US); University of Adelaide (AU); University of Alicante (ES); University of Antwerp (BE); University of Massachusetts Lowell (US); University of Pittsburgh (US); University of Tartu (EE).*

Language supervision by Debra Shulkes.

Cover image *Tackling the Toolkit* by Michal Tomek.

This publication was created with the support of Research Development Program RVO 68378068 and published with support from the Czech Academy of Sciences.

Institute of Czech Literature
of the Czech Academy of Sciences

Czech Academy
of Sciences

ISBN 978-80-7658-032-9
ISBN 978-80-7658-033-6 (PDF)

https://doi.org/10.51305/ICL.CZ.9788076580336

Contents

Introduction

The field of digital literary studies continues to grow as an ever-widening range of tools enables new readings and expands our textual explorations. This work is not a refutation but rather a continuation of traditional literary research; new findings are tested against established theories while computational approaches are used to challenge intuitive hypotheses with more systematic data. Familiar hermeneutical questions yield new answers when approached from unfamiliar angles, and data-driven observations help us develop new perspectives on texts. For the most part, this methodological turn to computational tools has been well received by poetry scholars, who have long been preoccupied with quantitative material. The counting and sorting of metres, syllables, rhymes, patterns and the like have been enhanced by the use of digital corpora, natural language processing, databases and computation, and we are now turning to machine learning to automatize processes. Inside the vast sphere of the digital humanities, the community of poetry scholars is an active subgroup whose shared interests span many language areas and time periods as well as hermeneutic questions and techniques.

Founded after a conference in Basel, Switzerland in 2017, the group *Plotting Poetry* is a dedicated platform for computationally-minded poetry scholars. The name of the group is based on a—perfectly serious—pun by the French poet Guillaume Apollinaire. In his 1917 reflections on poetry and modernity, Apollinaire used the phrase "*machiner la poésie*", which can be understood as a call for poets to mechanise poetry although "*machiner*" really means to plot as one might plot a coup. Playing on this pun in its English translation, our group name embraces the double meaning of plot: *Plotting Poetry* brings together researchers who try to plot literary phenomena on their graphs and are plotting to overcome their own limitations. The latter plot is staged by combining computational methods with scholarly thoroughness and traditional questions with new approaches.

When the *Plotting Poetry* collective decided to hold a conference in Prague in the autumn of 2020, we knew what our goals were and how we wanted to achieve them. We hoped to strengthen our platform for poetry scholars from different regions and to ensure that these researchers at different career stages could meet, connect, share good practices and start new collaborations. We agreed on the timeliness of a discussion about methods and tools. And so, after

considering how to foster exchanges and encourage the reporting of failures as well as successes, we carefully planned our event. Then, when the Covid-19 pandemic came and persisted, we found that—like everyone else over the last two years—we needed to revise and reshape plans.

After an initial effort to reschedule our conference, we decided to forego the pleasure of meeting in person. However, holding an online meeting felt so far removed from the convivial spirit of past *Plotting Poetry* events that we opted to come together instead in this volume of selected articles. The resulting work reports on a broad sweep of current research in the computational study of poetry. Its contributors have developed innovative methods and pursued interdisciplinary inquiries along unexpected paths. We hope the reader will find here the dynamic mix of ideas and backgrounds that *Plotting Poetry* aims to foster.

In *Tackling the Toolkit*, we have chosen to focus on the methodological innovations, challenges, obstacles and even shortcomings associated with applying quantitative methods to poetry specifically and poetics more broadly. Using tools including natural language processing, web ontologies, similarity detection devices and machine learning, our contributors explore not only metres, stanzas, stresses and rhythms but also genres, subgenres, lexical material and cognitive processes. Whether they are testing old theories and laws, making complex concepts machine-readable or developing new lines of textual analysis, their works challenge standard descriptions of norms and variations.

Anne-Sophie Bories

https://doi.org/10.51305/ICL.CZ.9788076580336.01
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

The Menzerath-Altmann Law in Czech Poems by K. J. Erben

Radek Čech
University of Ostrava, Czechia
cechradek@gmail.com
0000-0002-4412-4588

Ján Mačutek
Mathematical Institute,
Slovak Academy of Sciences / Constantine
the Philosopher University in Nitra, Slovakia
jmacutek@yahoo.com
0000-0003-1712-4395

Abstract

The aim of the paper is to test the validity of the Menzerath-Altmann law for Czech poems from K. J. Erben's ballad collection *Kytice z pověstí národních* (A Bouquet of Folk Legends). We focus particularly on the relationship between word length and syllable length. The Menzerath-Altmann law predicts that the mean syllable length will be longer in shorter words. The parameters of the mathematical model of this law for poems are compared with those for prose texts.

1 Introduction

According to Menzerath (1954), longer words in the German vocabulary tend to contain shorter syllables. A similar tendency, now known as the Menzerath-Altmann law (MAL), has been proven valid for multiple languages and several language units. In general, as Altmann observes, "[t]he longer a language construct the shorter its components (constituents)" (1980, p. 1). More specifically, when words are the constructs and morphemes their constituents, the mean length of morphemes decreases as the word length increases (Gerlach 1982). Similarly, when clauses are the constructs and syntactic phrases are the constituents, longer clauses consist of shorter syntactic phrases (Mačutek, Čech, et al. 2017). Overviews of work on this topic can be found in Cramer (2005) and Mačutek, Chromý, et al. (2019).

The most widely used mathematical model of the MAL is the function

$$y(x) = ax^b, \tag{1}$$

where x is the size of the construct, $y(x)$ is the mean size of the constituents in constructs of size x, and a and b are parameters. Buk et al. (2007, pp. 43–44) and Mačutek, Chromý, et al. (2019, p. 67) each provide a generalization of this model.

To the best of our knowledge, research on the MAL's application to poetic texts has so far been limited to two cases. A Slovak poem exemplifies the relationship between word length and syllable length in a study by Wimmer et al. (2003, pp. 105–106).[1] In addition, Čech et al. (2011, pp. 54–55) investigate the relationship between the word count of verses and the mean word length. The poetic texts in this case do not rhyme or follow any meter. The verses are also short, often consisting of only one word. In both of the above studies, the relations between unit lengths can be modeled by the MAL.

In what follows, we focus on the relationship between word length (measured in the number of syllables) and syllable length (measured in speech sounds) in a sample of Czech poetry. Our aim is twofold. First, we seek to test the validity of the MAL using a larger volume of "classical" poems that use rhyme and meter. In addition, these poems constitute a relatively homogeneous corpus as works written by one author. Second, we seek to compare the parameters of MAL for these poems with those for Czech prose texts.

2 Language Material and Methodology

Two groups of texts were chosen for our analysis. The first is comprised of 13 poems from the ballad collection *Kytice z pověstí národních* (A Bouquet of Folk Legends) written by Czech poet Karel Jaromír Erben.[2] This collection was first published in 1853. The poems it contains are highly influenced by folk poetry. For comparison, we again used Czech texts, namely eight short stories from the collection *Povídky malostranské* (Tales of the Lesser Quarter, first published in 1877) written by Jan Neruda.[3] We refer to the texts by Erben as E1–E13 and to the ones by Neruda as N1–N8. Table 1 provides an overview of the analyzed texts and their basic characteristics (i.e. the number of tokens and types they contain and the type-token ratio (TTR), which is here the ratio between the number of types and the number of tokens).[4]

In relation to the MAL, words were the constructs while the size of this construct was determined by the number of its syllables. Syllables were the constituents of the words, and their sizes were determined by the number of their speech sounds. We tested the MAL for word types (as opposed to tokens). As such, each word form was taken into account only once.[5]

Each text was analyzed separately. All non-syllabic prepositions were joined to the following word according to the preferred approach in quantitative linguistics, as described by Antić et al. (2006). First, for all of the words in the text, we determined the word length (WL), as calculated by the number of syllables.

[1] It should be noted that this book is written in Slovak and therefore not easily accessible to a broader audience.

[2] For our analysis, we used the texts available at http://www.cist.cz/Poezie/kytice.htm.

[3] We used the digitized edition *Sebrané spisy Jana Nerudy – díl desátý – Povídky malostranské*. Pořádá Ignát Herrmann. Vydal F. Topič v Praze 1893.

[4] For several other approaches to the evaluation of the type-token ratio, see Wimmer (2005) and Mitchell (2015).

[5] We, thus, followed the original approach of Menzerath (1954), who analyzed German vocabulary, i.e. types.

text	title	# tokens	# types	TTR
E1	Dceřina kletba	217	121	0.56
E2	Holoubek	308	236	0.77
E3	Lilie	479	336	0.70
E4	Poklad	2239	1050	0.47
E5	Polednice	197	152	0.77
E6	Štědrý den	651	441	0.68
E7	Svatební košile	1422	713	0.50
E8	Věštkyně	1109	696	0.63
E9	Vodník	923	529	0.57
E10	Vrba	481	312	0.65
E11	Záhořovo lože	2587	1380	0.53
E12	Zlatý kolovrat	1367	648	0.47
E13	Kytice	112	94	0.84
N1	Doktor Kazisvět	1947	1007	0.52
N2	Hastrman	1718	943	0.55
N3	Jak si pan Vorel nakouřil pěnovku	1498	774	0.52
N4	O měkkém srdci paní Rusky	1630	868	0.53
N5	Pan Ryšánek a pan Schlegl	3062	1425	0.47
N6	Přivedla žebráka na mizinu	2207	1128	0.51
N7	Svatováclavská mše	3678	1728	0.47
N8	U Tří lilií	646	397	0.61

Table 1: Texts and their basic characteristics

For all cases except for extra-syllabic consonants (Crystal 2008, pp. 182–183), the number of syllables in the word equaled the number of sonority peaks in the word's sonority profile. A sonority scale consisting of three classes (vowels, sonorants, and obstruents) was applied to construct the sonority profiles. As a next step, word length was also determined based on the number of speech sounds.

Given that we work with mean syllable lengths (measured in speech sounds) in words of certain syllable lengths, the results become unstable if some word lengths do not occur frequently enough (e.g. there is only one five-syllable word in text E2). We therefore pooled some word length groups so that there were at least five words in each category.[6] We can illustrate this approach using text E2 from Table 2. This work has twelve four-syllable words and one five-syllable word. These two word lengths were combined into one category, which was then represented by the mean WL of all the words which it contains

[6] The limit we set (i.e. at least five words) was only our rule of thumb. In contrast, Mačutek and Rovenchak (2011) require 10 words as the minimum frequency per category and their analysis excludes construct lengths that do not satisfy this condition. We note that the results of fitting presented in this paper barely change if the latter approach is taken (the difference between the values of parameter b obtained by the two approaches is always within the limit of 0.02).

(i.e. $(12 \times 4 + 1 \times 5)/13 = 4.08$). The value 2.21 is the mean syllable length of all syllables occurring in words from this category.

3 Results

The MAL in the form of function (1) was fitted to the texts presented in Table 1. The goodness of fit of the model was evaluated in terms of the determination coefficient R^2. A fit is usually considered satisfactory if $R^2 \geq 0.9$; see Mačutek and Wimmer (2013).

It is obvious that if we substitute 1 for x in (1), we obtain $y(1) = a$. Parameter a can, thus, be (in this context) interpreted as the mean length (measured in speech sounds) of monosyllabic words; see Kelih (2010). Parameter b is the value which maximizes the determination coefficient. Fitting was performed using NLREG software (www.nlreg.com).

The results are presented in Table 2, Table 3, and Table 4. In these tables, MSL denotes the mean syllable length, f_{WL} is the frequency of word types of length WL, and a and b are parameters of the MAL; see (1).

As can be seen in Table 2, the fit is mostly very good. Among the texts presented in Table 2, text E7 is the only exception where the value of R^2 falls below 0.9. Since, however, this threshold is only a rule of thumb and the determination coefficient falls only slightly short of it, we may conclude that these texts abide by the MAL.

On the other hand, Erben's poem *Kytice* (text E13) clearly does not tend to reflect the MAL in the same way. This is evident in Table 3.

There are at least two possible explanations for the failure of this poem to follow the MAL. Admittedly, both these accounts are speculative as we do not have similar texts at our disposal which might verify or refute these hypotheses.

First, the poem *Kytice* is the shortest text in our sample, with just 112 word tokens and 94 types (see Table 1). It may be that this text is simply too short, and the "mechanism" behind the MAL is too "weak" to have any impact. We note that Čech (2015) suggests that when studying the structure of word frequencies, the "ideal" text length is somewhere between 200 and 6500 word tokens, and the same recommendation is made by Čech (2016, p. 57) for investigations of the thematic concentration of texts. This poem is the only text that is much shorter than 200 word tokens in our sample (the second shortest one, E5, contains 197 word tokens). In addition, its TTR achieves a very high value (see Table 1), which is typical of short texts.

Second, the poem's word length structure seems quite distinctive in the context of most of the other poetic texts analyzed in this paper. In particular, the ratio of the frequencies of three-syllable to two-syllable word types is roughly 0.79 (with 27 and 34 word types, respectively; see Table 3). The only other poem with this property is *Svatební košile* (E7) where the ratio is 0.68; otherwise, the ratio for works in the sample falls below 0.5. In both these exceptional cases, the iambic meter is often violated by a three-syllable word in an odd position of verse line. As such, these two poems have a less regular iambic character than other poems in the collection. It remains an open question what effect this

E1			E2			E3			E4		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.90	39	1	2.69	52	1	3.04	114	1	2.77	201
2	2.52	48	2	2.35	117	2	2.46	149	2	2.45	603
3	2.30	22	3	2.24	54	3	2.34	64	3	2.32	152
4	2.21	13	4.08	2.21	13	4	2.19	9	4.03	2.22	94
a	2.90		a	2.69		a	3.04		a	2.77	
b	-0.20		b	-0.16		b	-0.25		b	-0.16	
R^2	0.997		R^2	0.950		R^2	0.972		R^2	0.996	
E5			E6			E7			E8		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.87	53	1	2.86	87	1	3.14	182	1	3.05	126
2	2.51	63	2	2.42	236	2	2.45	311	2	2.47	360
3	2.32	31	3	2.32	100	3	2.30	213	3	2.28	163
4	2.15	5	4	2.16	24	4.14	2.32	7	4.02	2.26	47
a	2.87		a	2.86		a	3.14		a	3.05	
b	-0.20		b	-0.20		b	-0.26		b	-0.25	
R^2	0.997		R^2	0.981		R^2	0.883		R^2	0.948	
E9			E10			E11			E12		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.88	112	1	2.75	71	1	3.01	178	1	3.11	172
2	2.38	260	2	2.35	180	2	2.47	727	2	2.41	315
3	2.35	117	3	2.22	40	3	2.32	335	3	2.34	123
4	2.22	40	4.05	2.18	21	4	2.18	125	4	2.20	38
						5	2.16	5			
a	2.88		a	2.75		a	3.01		a	3.11	
b	-0.20		b	-0.19		b	-0.23		b	-0.27	
R^2	0.925		R^2	0.958		R^2	0.967		R^2	0.933	

Table 2: MAL in poems by Erben (texts E1–E12)

peculiarity of the word length structure has on syllable length. The impact may be somewhat neutralized in a relatively long poem like text E7, but text E13 stands out as an exception. Within the current sample, it has the lowest mean syllable length for one- and three-syllabic words and the highest mean syllable length for four-syllabic words.

In Neruda's short stories, the relationship between word length and syllable length also follows the MAL, with R^2 values above the threshold of 0.9 for all eight texts (see Table 4).

The graph in Figure 1 depicts the MAL for texts E6 and N7.

Kytice (text E13)		
WL	MSL	f_{WL}
1	2.50	26
2	2.44	34
3	2.21	27
4	2.34	7
a	2.50	
b	-0.07	
R^2	0.571	

Table 3: MAL in the poem *Kytice* by Erben (text E13)

N1			N2			N3			N4		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.99	138	1	2.95	146	1	2.98	124	1	2.95	139
2	2.56	420	2	2.59	392	2	2.56	329	2	2.54	365
3	2.38	289	3	2.39	276	3	2.35	222	3	2.36	250
4	2.33	118	4	2.27	99	4	2.25	87	4	2.35	86
5	2.24	36	5	2.27	24	5.31	2.21	12	5	2.21	21
6	1.81	6	6.17	1.96	6				6	1.84	7
a	2.99		a	2.95		a	2.98		a	2.95	
b	-0.21		b	-0.19		b	-0.20		b	-0.21	
R^2	0.921		R^2	0.955		R^2	0.978		R^2	0.901	
N5			N6			N7			N8		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	3.08	189	1	3.04	168	1	3.17	212	1	2.94	86
2	2.58	606	2	2.54	496	2	2.57	701	2	2.50	179
3	2.40	407	3	2.37	333	3	2.40	544	3	2.36	93
4	2.28	170	4	2.30	101	4	2.32	211	4	2.16	32
5	2.28	39	5	2.17	24	5	2.18	47	5	2.20	7
6	1.90	8	6.17	1.86	6	6.24	1.90	13			
7	2.12	6									
a	3.08		a	3.04		a	3.17		a	2.94	
b	-0.22		b	-0.23		b	-0.25		b	-0.20	
R^2	0.926		R^2	0.955		R^2	0.968		R^2	0.968	

Table 4: MAL in short stories by Neruda (texts N1–N8)

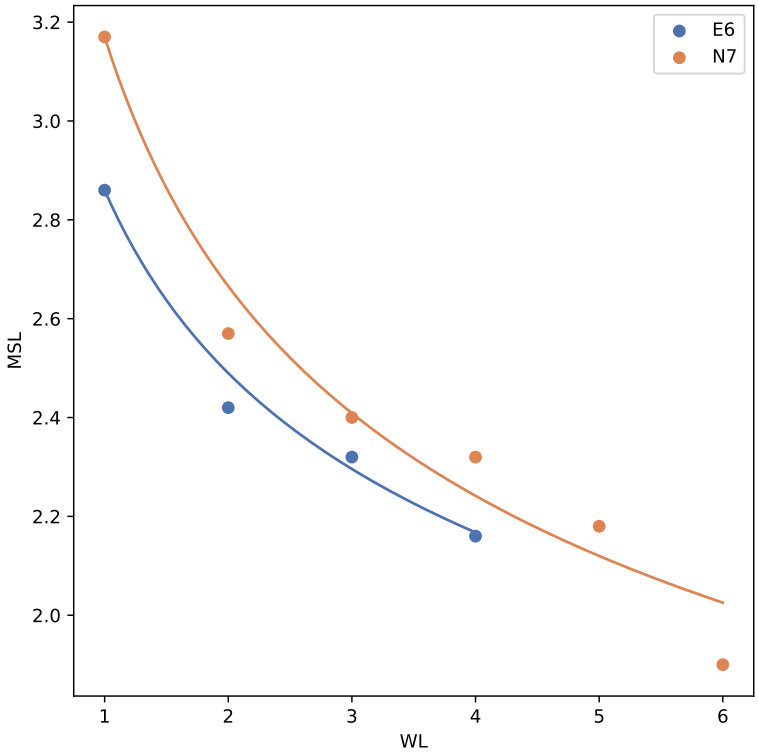


Figure 1: MAL fitted by function (1) for texts E6 and N7

4 Discussion and Conclusion

The results in Section 3 indicate that the MAL is a valid model for the relation between word and syllable lengths not only in prose texts, but also (with one exception, which can be at least partially explained) in poetic texts. Nevertheless, they also reveal several differences between the two groups of texts.[7]

There is a statistically significant difference (a p-value below 0.1) between the values for parameter a, i.e. the mean length of monosyllabic words, in the poetry and prose works. The same is also true of the differences between the mean syllable lengths of two-, three-, and four-syllabic words in the texts by Erben on the one hand and those by Neruda on the other.[8] This suggests a tendency to use shorter words in poetry. At the same time, the difference between parameters b in poems by Erben and short stories by Neruda was not significant. This means that mean syllable length decreases with increasing word length at the same

[7] All data were first tested for normality using the Shapiro-Wilk test. Depending on whether or not the normality hypothesis was rejected, we then applied either the t-test or the Wilcoxon–Mann–Whitney test. The tests were performed in R (www.r-project.org).

[8] Differences in the mean syllable lengths of longer words were not tested because of the insufficient data in these poetic texts.

	Erben				Neruda			
	# types	TTR	a	b	# types	TTR	a	b
# tokens	0.99	-0.74	0.31	-0.19	> 0.99	-0.92	0.94	-0.82
# types		-0.66	0.32	-0.21		-0.92	0.92	-0.81
TTR			-0.44	0.33			-0.82	0.71
a				-0.97				-0.91

Table 5: Pearson correlation coefficients between basic text characteristics and MAL parameters

rate for both groups of texts. However, the curves representing model (1) for prose tend to be higher than the ones for poetry (as can be seen in Figure 1 for two texts).

Another difference relates to the Pearson correlation coefficients between some basic text characteristics (text length in tokens and in types; the TTR) and the parameters of the MAL. In Table 5, we can see that the values of both parameters correlate closely with text length (expressed in terms of both tokens and types) and the TTR in Neruda's short stories. However these correlations are much weaker for the poems by Erben. The correlation between parameters a and b is very strong in all of the texts. These correlations hint towards an interpretation of the MAL parameters for prose texts. The dependence of parameter a (i.e. the mean length of monosyllabic words) on text length conforms with the findings of Kelih (2012), who showed that the mean length of word types increases with increasing text length.

The far weaker correlations between MAL parameters and text length in poems may be explained by the tendency of the words in poems to have shorter syllables than those in prose texts. Shorter syllables have a higher proportion of vowels, which may relate to attempts to achieve euphonic effects.[9] If an author (poet) deliberately opts for shorter syllables,[10] then since we assume a synergetic model of language such as the one suggested by Köhler (2005) where all linguistic properties are interrelated, this choice will be reflected in other changes. One such change may be that text length has a far weaker influence, and other "language forces" (Altmann and Köhler 1996) come into play. Identifying these forces is one of our future challenges. In addition, the MAL in poetic texts may be analyzed at other levels, and phenomena that do not appear in prose (e.g. verses and stanzas) emerge as new candidates for reasonable units. Precisely which properties distinguish poems from other genres remains to be seen.

[9] The idea that the proportions of vowels and consonants differ between prose and poetry can be traced back to at least the second half of the 15th century. A 1830 commentary explicitly mentions that euphony depends partly on this proportionality, and it specifies the values of the proportions for several languages. See Grzybek (2013) for more details.

[10] The words in their poems will, thus, also be shorter than those in prose in terms of speech sounds.

Acknowledgments

J. Mačutek's work on this paper was supported by VEGA grant no. 2/0096/21.

References

Altmann, Gabriel (1980). "Prolegomena to Menzerath's law". In: *Glottometrika 2*. Ed. by Rüdiger Grotjahn. Bochum: Brockmeyer, pp. 1–10.

Altmann, Gabriel and Reinhard Köhler (1996). "'Language forces' and synergetic modelling of language phenomena". In: *Glottometrika 15*. Ed. by Peter Schmidt. Trier: WVT, pp. 62–76.

Antić, Gordana, Emmerich Kelih, and Peter Grzybek (2006). "Zero-syllable words in determining word length". In: *Contributions to the Science of Text and Language*. Ed. by Peter Grzybek. Dordrecht: Springer, pp. 117–156.

Buk, Solomija and Andrij Rovenchak (2007). "Statistical parameters of Ivan Franko's novel *Perekhresni stežky (The Cross-Paths)*". In: *Exact Methods in the Study of Language and Text*. Ed. by Peter Grzybek and Reinhard Köhler. Berlin, New York: de Gruyter, pp. 39–48.

Čech, Radek (2015). "Text length and the lambda frequency structure of the text". In: *Sequences in Language and Text*. Ed. by George K. Mikros and Ján Mačutek. Berlin, Boston: de Gruyter, pp. 71–88.

Čech, Radek (2016). *Tematická koncentrace textu v češtině*. Praha: UFAL.

Čech, Radek, Ioan-Iovitz Popescu, and Gabriel Altmann (2011). "Word length in Slovak poetry". In: *Glottometrics* 22, pp. 44–56.

Cramer, Irene M. (2005). "Das Menzerathsche Gesetz". In: *Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: de Gruyter, pp. 659–688.

Crystal, David (2008). *A Dictionary of Linguistics and Phonetics*. Malden (MA): Blackwell.

Gerlach, Rainer (1982). "Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie". In: *Glottometrika 4*. Ed. by Werner Lehfeldt and Udo Strauss. Bochum: Brockmeyer, pp. 95–102.

Grzybek, Peter (2013). "Historical remarks on the consonant-vowel proportion — from cryptoanalysis to linguistic typology. The concept of phonological stoichiometry (Francis Lieber, 1800-1872)". In: *Glottometrics* 26, pp. 96–103.

Kelih, Emmerich (2010). "Parameter interpretation of Menzerath law: Evidence from Serbian". In: *Text and language. Structures, Functions, Interrelations, Quantitative Perspectives*. Ed. by Peter Grzybek, Emmerich Kelih, and Ján Mačutek. Wien: Praesens, pp. 71–79.

Kelih, Emmerich (2012). "On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts". In: *Synergetic Linguistics. Text and Language as Dynamic Systems*. Ed. by Sven Naumann, Peter Grzybek, Relja Vulanović, and Gabriel Altmann. Wien: Praesens, pp. 67–80.

Köhler, Reinhard (2005). “Synergetic linguistics”. In: *Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: de Gruyter, pp. 760–775.
Mačutek, Ján, Radek Čech, and Jiří Milička (2017). “Menzerath-Altmann law in syntactic dependency structure”. In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Ed. by Simonetta Montemagni and Joakim Nivre. Linköping: Linköping University Electronic Press, pp. 100–107.
Mačutek, Ján, Jan Chromý, and Michaela Koščová (2019). “Menzerath-Altmann Law and prothetic /v/ in spoken Czech”. In: *Journal of Quantitative Linguistics* 26.1, pp. 66–80. DOI: 10.1080/09296174.2018.1424493.
Mačutek, Ján and Andrij Rovenchak (2011). “Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length”. In: *Issues in Quantitative Linguistics 2*. Ed. by Emmerich Kelih, Victor Levickij, and Yuliya Matskulyak. Lüdenscheid: RAM-Verlag, pp. 136–147.
Mačutek, Ján and Gejza Wimmer (2013). “Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics”. In: *Journal of Quantitative Linguistics* 20.3, pp. 227–240. DOI: 10.1080/09296174.2013.799912.
Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.
Mitchell, David (2015). “Type-token models: A comparative study”. In: *Journal of Quantitative Linguistics* 22.1, pp. 1–21. DOI: 10.1080/09296174.2014.974456.
Wimmer, Gejza (2005). “The type-token relation”. In: *Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: de Gruyter, pp. 361–368.
Wimmer, Gejza, Gabriel Altmann, Luděk Hřebíček, Slavomír Ondrejovič, and Soňa Wimmerová (2003). *Úvod do analýzy textov*. Bratislava: Veda.

https://doi.org/10.51305/ICL.CZ.9788076580336.02
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

Description of Postdata Poetry Ontology V1.0

María Luisa Diez Platas
International University of La Rioja, Spain
marialuisa.diez@unir.net
0000-0002-9416-3147

Helena Bermúdez
Université de Lausanne, Switzerland
helena.bermudez@linhd.uned.es
0000-0002-8627-1367

Salvador Ros
POSTADATA Project. UNED, Spain
sros@scc.uned.es
0000-0001-6330-4958

Elena González-Blanco
POSTADATA Project. UNED, Spain
elena@coverwallet.com
0000-0002-0448-1812

Oscar Corcho
Ontology Engineering Group, UPM, Spain
ocorcho@fi.upm.es
0000-0002-9260-0753

Omar Khalil Gómez
POSTADATA Project. UNED, Spain
omar.khalil@dia.uned.es
0000-0002-4128-7074

Laura Hernández-Lorenzo
POSTADATA Project. UNED, Spain
laura.hernandez@scc.uned.es
0000-0003-3489-2193

Mirella De Sisto
POSTADATA Project. UNED, Spain
mdsisto@scc.uned.es
0000-0002-0899-5976

Javier de la Rosa
POSTADATA Project. UNED, Spain
versae@scc.uned.es
0000-0002-9143-5573

Álvaro Pérez
POSTADATA Project. UNED, Spain
alvaro.perez@scc.uned.es
0000-0001-5897-1246

Aitor Diez
POSTADATA Project. UNED, Spain
adiazm@scc.uned.es
0000-0002-2095-4827

José Luis Rodriguez
POSTADATA Project. UNED, Spain
jl.rodriguez@patrimonionacional.es
0000-0003-0888-0084

Abstract

One stream of work in the digital humanities focuses on interoperability processes and the description of traditional concepts using computer-readable languages. In the case of literary studies, there has been some research into these topics, but the complexity of the knowledge domain remains an issue. This complexity is based on the different interpretations of concepts in different traditions, the use of isolated and private databases, unique applications of language and, thus, the richness of poetic information. All of this suggests the need to explore new options to represent the complexity in computer-readable language. This paper presents an ontology network designed to capture poetry domain knowledge. The ontologies in question relate to poetic works and their structural and prosodic components.

1 Introduction

The Poetry Standardization and Linked Open Data Project, POSTDATA, aims to provide a means for European poetry researchers to publish and access semantically enriched data. To achieve this goal, it was necessary to develop a poetry ontology. This ontology attempts to enhance interoperability in the European poetry research community and capture the concepts and properties that define the domain of European poetry knowledge. The development of the ontology began with an attempt to define a domain model of poetry based on an analysis of 23 poetry repertories (i.e. poetry research databases) (Curado Malta, González-Blanco, et al. 2016; Postdata ERC project 2021). These repertories were selected because of their relevance, availability in digital format (i.e. all are implemented in databases) and the rich sample they provided of multilingual poetry. They ranged from the classical period (e.g. Pedecerto[1]) to the modern era (e.g. Corpus of Spanish Golden-Age Sonnets[2]) to the Middle Ages (e.g. Cantigas de Santa Maria for singers.[3]) This first step allowed us to identify the most significant identities and properties that define a poetic work (i.e. a poem), taking into account the different traits related to this literary genre. The result was a domain model that reflected poetry's complexity and heterogeneity. We then transformed this domain model into an ontology network, which would allow for its effective and extensive use in computational frameworks as a Post-data computer-aided annotation tool. In this paper, we present the first version of our network's four most significant ontologies (i.e. version 1.0).[4] These ontologies relate to poetic works, their structural and prosodic components and information about relevant dates. To populate the ontologies, we incorporated the ontology definitions into OMEKA, a framework that facilitates the use of these ontologies in research tasks. This article is structured as follows: In Section 2, we present some previous results related to ontologies of literature, especially of the poetry domain. Section 3 describes the methodology that we used to develop our ontologies. Section 4 presents a detailed description of the most relevant ontologies that we created. Finally, Section 5 outlines our conclusions and directions for future work.

1 http://www.pedecerto.eu
2 https://github.com/bncolorado/CorpusSonetosSigloDeOro
3 http://www.cantigasdesantamaria.com/
4 http://postdata.linhd.uned.es/results/network-of-ontologies/

2 Related works

The first attempt to build a poetry ontology took place within the ReMetca project (González-Blanco and Rodríguez 2016). The results of that project were used to define the TEI-Verse module.[5] However, TEI's tags did not totally capture the structural richness of poetic traditions and alternative approaches remained available. If, for example, we understood the poetic work as a cultural heritage item, we might represent the knowledge associated with it by using the CRM-CIDOC ontology, the Conceptual Reference Model (CIDOC-CRM),[6] which formally describes cultural heritage concepts and relationships. Similarly we might apply the Functional Requirements of Bibliographic Records (FRBR) ontology[7] and FRBRoo,[8] which offer perspectives based on bibliographic and authority records (i.e. the standardised names for people and corporate bodies) (Tillett 2005). These ontologies might cover descriptive aspects of poetic works and the forms of their expression and manifestation; for a description of the the poetic work, its expression and manifestation, see (David et al. 2019; *Home | FRBRoo* 2021). However they could not reflect structural aspects of the works or provide any literary analysis or prosody. In other words, they contained no modelling information for the analysis of textual features.

Aside from examining these well-known ontologies in the digital humanities, we also completed a general search for potentially relevant ontologies in more standard ontology repositories such as Linked Open Vocabularies,[9] Open Metadata Registry[10] and the Basel Register of Thesauri Ontologies & Classifications.[11] We concluded that despite specific efforts to model the poetry domain and the possibility of reusing some foundational ontologies to deal with poetry, these tools were only relevant to limited features of the poetry domain. This situation confirmed the need to create a new, comprehensive ontology of poetry as a literary genre.

3 Ontology network methodology

The first step in this work was to build a conceptual domain model of European poetry based on an accurate picture of the knowledge domain. A conceptual model is a representation of the knowledge domain created from concepts and properties and their relationships. For this purpose, we analysed a set of 23 digital repertories (Curado Malta, Centenera, et al. 2017; Curado Malta, González-Blanco, et al. 2016, 2020; Postdata ERC project 2021) that had been selected for their representation of different poetry traditions, languages, prosodic systems and cultures (Postdata ERC project 2021). These repertories arose from research projects and, thus, contained information that had been gathered or

[5] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html
[6] http://www.cidoc-crm.org/cidoc-crm
[7] http://purl.org/vocab/frbr/core#
[8] http://www.cidoc-crm.org/frbroo/home-0
[9] https://lov.linkeddata.es/dataset/lov/
[10] http://metadataregistry.org
[11] http://www.bartoc.org/

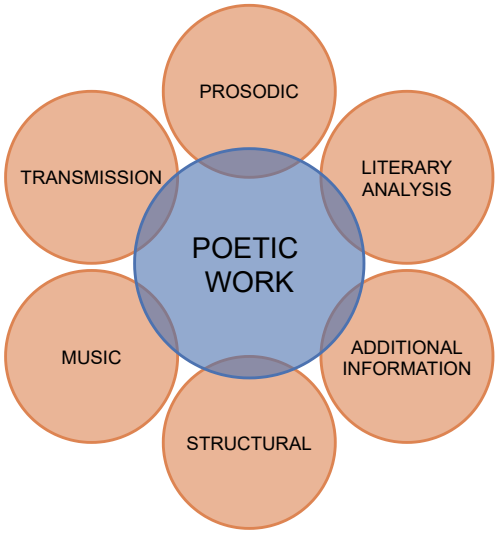


Figure 1: Areas of knowledge in the domain model

generated by experts and had more reliable and robust content, categories and structures. To build the ontology, we applied the NEON methodology approach (Suárez-Figueroa 2010). The latter defines different working scenarios. Once the sources were selected, we applied scenario 2 (i.e. reusing and re-engineering non-ontological resources). According to this scenario, we began working with the conceptual structures (database structures) of the different repositories. Here we used a reverse engineering approach, i.e. moving from the conversion of the database schema to the ontology modelling process. All the concepts were extracted and compared so that we could build a common model out of them (Postdata 2020a,b).

Based on this process, we developed the complete European Poetry Logical Domain Model (EP-DM). This work required us to deal problems related to the potential inconsistency of concepts. These problems had previously been addressed by experts in the poetry domain (Bermúdez-Sabel, Curado Malta, et al. 2017; Bermúdez-Sabel, Díez Platas, et al. 2019; González-Blanco, Rio Riande, et al. 2016). The final conceptual model included both descriptive and bibliographic aspects of the poetic works. It also included information about textual transmission, prosodic,literary and rhetorical features, poetic structures, significant publication elements and relationships with music(Curado Malta, González-Blanco, et al. 2016; Malta et al. 2017) (see Figure 1).

The EP-DM was highly complex because of the number of concepts and properties that had to be included in this model of poetry knowledge's domain. It is worth remembering here that the formal research process in the humanities is open-ended; each investigation focuses on different issues in the field and interprets them in different ways. The size of this first approximation meant that our model lacked usability. We therefore decided to build an ontology

network to represent information about European poetry. Our criteria for building each of the ontologies in the network were as follows:

- The classes, relationships and axioms of the ontology had to be thematically related or else necessary to complete the semantics of another ontology entity. The underlying semantics of each class had to relate to the area of knowledge.
- Weak coupling was required between ontological modules. Each ontology was built as a self-contained module that was related to the other ontologies through a small number of relationships.
- Each module had to be highly cohesive. In particular, the module had to contain the maximum number of class properties in order to ensure a highly cohesive ontology. As such, the ontology should be able to function while avoiding coupling with other ontologies as much as possible.

We completed the development process using an iterative-incremental model. Each ontology was built based on the principles of reusing ontologies, aligning vocabularies and properties to facilitate development, improving the semantic understanding of entities and facilitating interoperability Figure 2.

After each iteration, the OWL specification (*OWL - Semantic Web Standards* 2021) was obtained for each ontology. Eventually, to define property ranges, we identified a set of controlled vocabularies. These were specified using the data model of the Simple Knowledge Organization System (SKOS).[12] These controlled vocabularies allowed us to establish the standard terminologies that are used by the scientific community. The result was ontology version 1.0.

4 Ontology descriptions

This section describes Postdata ontology V1.0. As noted above, this version of the ontology network was formed from the four ontologies we had already developed: the postdata-core ontology, the postdata-prosodic ontology, the postdata-structural ontology and the postdata-dates ontology. These ontologies were the main outcomes of the Postdata project. A literary and transmission ontology will be completed before the end of the project.

4.1 Postdata-core ontology

This is the main ontology for poetic representation,[13] and its prefix is pdcore (i.e. poetry domain-core). It provides information about poetic works and their manifestation. A poetic work (i.e. poem) can be represented through its different manifestations or versions in the poetry domain. In literature, it is also common to find sets of poems grouped together, for example, in a book. These scenarios are represented through three main classes and their relationships: pdcore:PoeticWork, pdcore:Redaction and pdcore:Ensemble (see Figure 3).

[12] https://www.w3.org/2004/02/skos
[13] http://postdata.linhd.uned.es/ontology/postdata-core/

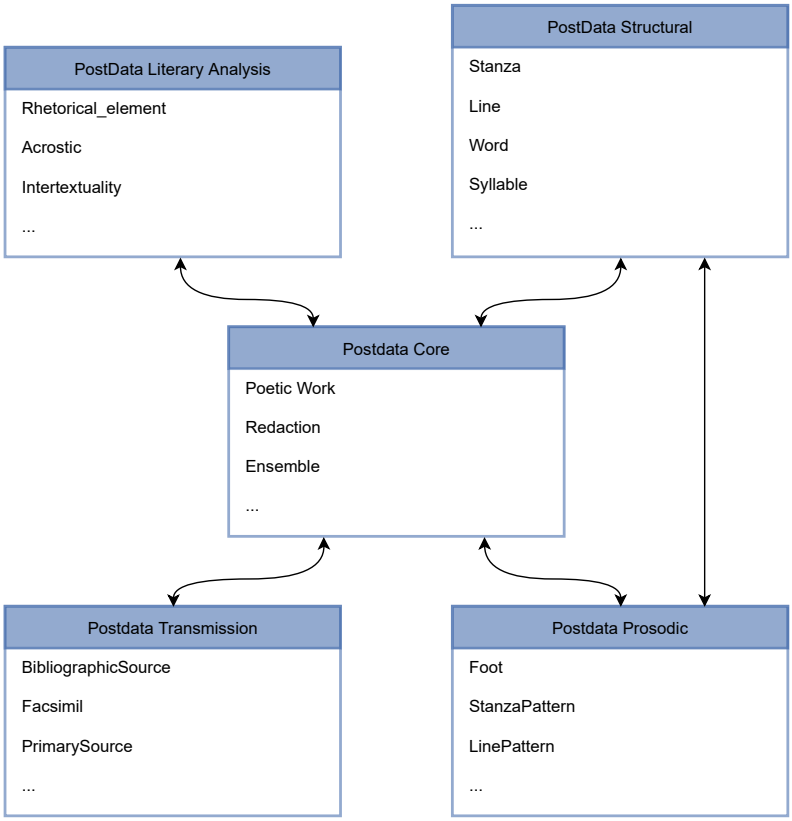


Figure 2: Network of ontologies version 1.0

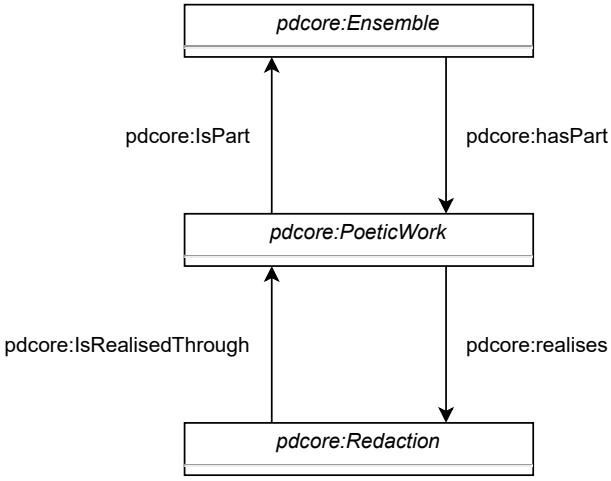


Figure 3: Postdata-core ontology Version 1.0

These classes are special FRBRoo classes.

- The PoeticWork class models the abstract concept of artistic creation. These creations must be in verse (poems, plays or songs), and their properties represent the descriptive metadata of the poetic work (title, abstract, creator or author and creation date). This class is implemented as a subclass of frbroo:F1 class.
- The Redaction class is a subclass of frbroo:F22 class, and it models editions of a poetic work. Each version of a poetic work is a redaction.
- The Ensemble class is a subclass of frbroo:F17 class, and it enables the modelling of an Ensemble as a collection of poetic works (i.e. a book of songs or collection of poems).

Besides these main classes, we considered including classes which, although not specific to poetry, model a transversal knowledge of the poetic domain and provide the relevant information. Some examples of these classes are:

- pdcore:Person and pdcore:Organisation. This class models the agents that participate in the poetic work and the redaction according to their different roles.
- pdcore:Place and pdcore:Event. This class represents places of origin and references to events and locations.

One aspect of a literary work that needed to be addressed was its authorship and the roles played by related agents (see Figure 4). To model this knowledge, we used the design pattern agentRole[14] and defined the class pdcore:Role and its subclass pdcore:CreatorRole. The pdcore:CreatorRole class is useful for dealing with authorship because it can support the representation of:

- Multiple authors through the multiple cardinality assignation of the pdcore:hasCreator property.
- An anonymous author using the pdcore:isAnonymous boolean property.
- Wrongful attributions (i.e. cases where a work was written by another author) through the pdcore:isWrongAttribution property with a Boolean range.

We also identified a set of controlled vocabularies which were used in some of the class properties as ranges.

- In the PoeticWork class, we identified genre, poeticType and authorEducationLevel.
- In the Person class, we identified gender, literaryPeriod, school, socialStatus and religiousAffiliation.

[14] http://www.ontologydesignpatterns.org/cp/owl/agentrole.owl

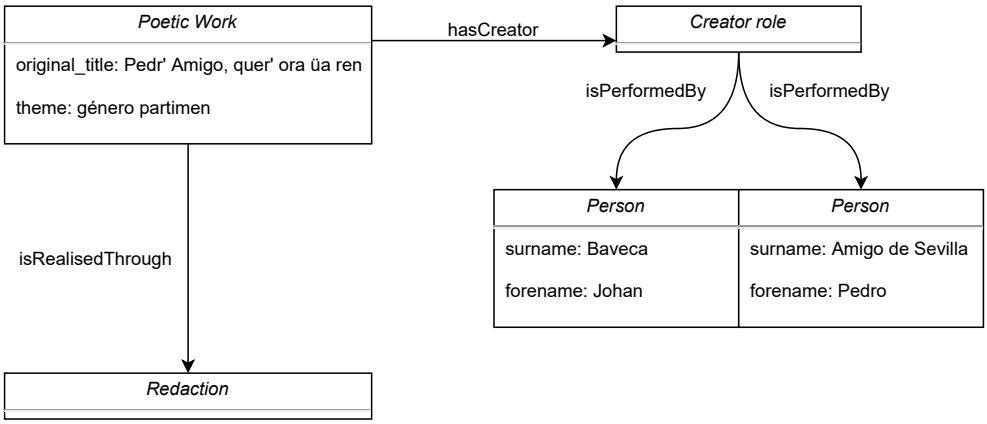


Figure 4: Example: Authorship of a poetic work

- In the Role class, we identified roleFunction and typeOfCharacter.
- In the CreatorRole class, we identified typeOfDesignation.
- In the Redaction class, we identified typeOfTextualElement.

4.2 Postdata-structuralElements ontology

The postdata-structuralElements ontology[15] contains all the information related to the structural elements of a redaction. The ontology prefix is pdstruct. Each redaction of a poetic work is organised into lines and stanzas. The structuralElements ontology defines two classes: pdstruct:OrderedLineList and pdstruct:OrderedStanzaList. These classes are related to the pdcore:Redaction class using the pdstruct:hasLineList property or the psdtruct:hasStanzaList property. Since a stanza is a list of lines, a line is a list of words and punctuation marks, and a word is composed of syllables, we defined five more classes to complete the ontology. These were pdstruct:Line, pdstruct:Stanza, pdstruct:Word, pdstruct:Syllable and pdstruct:Punctuation.

- A line is a unit of verse that usually ends in a visual or typographic break and is characterised by its length and metre.
- A stanza is a group of lines. Usually, this grouping forms the basic recurring verse unit of a poem.
- A word is a list of syllables.
- A syllable is a single unit of speech sound and may have written or spoken forms.
- Punctuation refers to punctuation symbols.

[15] http://postdata.linhd.uned.es/ontology/postdata-structuralElements/

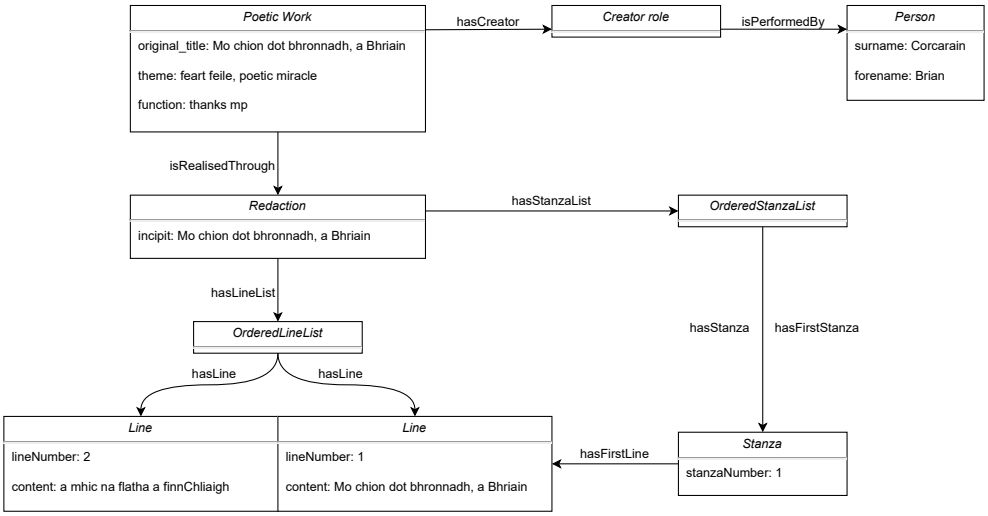


Figure 5: Example: List of lines and stanzas in a structural ontology

Based on Ordered List ontology (olo) semantics, these classes are subclasses of olo:Slot since they may each be understood as a slot in an ordered list. In the first three cases, they are subclasses of olo:OrderedList since they represent a list of ordered elements Figure 5.

In this ontology, four controlled vocabularies were identified and used in some of the class properties as ranges.

- In Stanza, we identified typeOfStanza and typeOfStanzaEdition.[16]
- In Word, we identif partOfSpeech.
- In Syllable, we identified nucleusType.

4.3 Postdata-prosodicElements ontology

This ontology[17] contains the classes and properties required to structure the information extracted from the prosodic analysis of a poetic work. The prefix for this ontology is pdprosodic. The prosodic analysis of a poetic work provides information about the poem's metrical patterns (see Figure 6). These patterns are defined at three levels: poem, stanza and line. This ontology imports the postdata structuralElements ontology since the latter's content relates to the metrical patterns of the line, the stanza and the poem. On this basis, we define three classes to represent these metrical patterns: the pdprosodic:LinePattern, the pdprosodic:StanzaPattern and the pdprosodic-WorkPattern. The three classes belong to the pdprosodic:Pattern hierarchy.

[16] A description of all these properties is available at https://cutt.ly/Dx0BDkw
[17] http://postdata.linhd.uned.es/ontology/postdata-prosodicElements/

- The LinePattern models the metrical pattern of the line. Some important properties related to the line pattern are:
  - pdprosodic:accentedVowels: This represents stressed vowels in the order in which they occur in the text.
  - pdprosodic:countingMetricalScheme: This represents the metrical scheme based on the number of syllables.
  - pdprosodic:grammaticalStressPattern: This represents patterns based on the position of expected stresses according to grammatical rules, including the distribution of weak and strong positions.
- The StanzaPattern summarises specific properties of the stanza as a rhyme scheme. One of the most common conventions is the use of letters to identify rhyming lines.
- The WorkPattern shares some of the properties defined in the LinePattern and StanzaPattern classes but it also presents certain properties as pdprosodic:presentRhymeMatching. This allows a poetic work to be categorised according to the extent of matches between different rhyming sounds (i.e. based on assonance or consonance). This property also applies to the StanzaPattern class.

These ontologies were enriched by adding more classes that store prosodic analysis data. These classes included:

- pdprosodic:Rhyme: This represents the repetition of similar-sounding words at the end of lines in poems or songs.
- pdprosodic:Foot: This is the unit of poetic metre used in most Indo-European poetic traditions, including English syllabic verse and Ancient Greek and classical Latin poetry.
- pdprosodic:metricalEncoding: This is the notation used to represent a metrical pattern; for example,the plus sign is used to encode a strong syllabic position.

In this ontology, controlled vocabularies are of special interest because they represent a normalised form of the values of prosodic properties. We defined 12 vocabularies.

- In Patterns, we identified clausulaSchemeType, metricalCategory, metricalComplexity, metricalContext, rhymeDispositionType and versificationType.
- In Line, we identified Stanza: feetType and metricalType.
- In Foot, we identified FootDivision: clausula, footType and footUnitType.
- In RhymeMatching, we identified typeOfRhymeMatching.
- In Metaplasm, we identified typeOfMetaplasm, which refers to dieresis, syneresis, synalepha and hiatus.

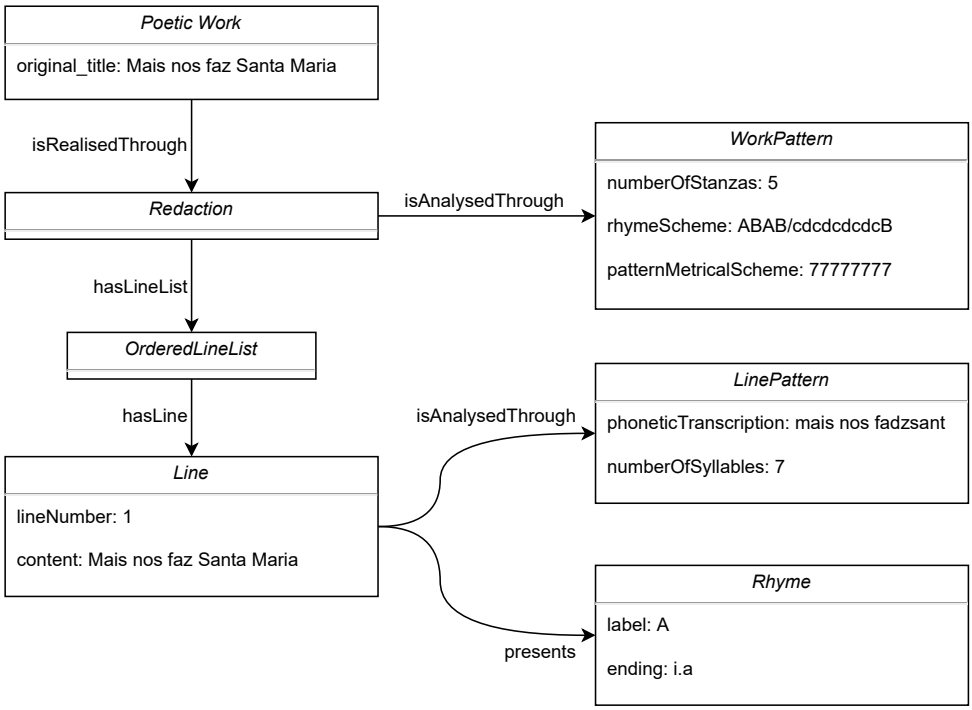


Figure 6: Example: WorkPattern and LinePattern classes in prosodic ontology

4.4 Postdata-dates ontology

Depending on the composition period, it may be difficult to date a poetic work or its versions with any precision. Moreover, in ancient or anonymous publications, it is often impossible to determine a composition date. It may be necessary to establish ranges or suggest a likely date. This problem arises when the form of a text's transmission or preservation does not support tracing the composition date. To address these situations, we have proposed an independent and reusable ontology for the literary or heritage domain that covers special dating needs Figure 7. In this ontology, two classes are provided:[18]

- pddates:DateEntity represents a temporal entity associated with the poetic work, its manifestations or an event.
- pddates:DateExpression forms the basis of the class hierarchy. This class and its subclasses provide different modes for representing a date related to the work's creation or relevant event associated with an entity.

[18] http://postdata.linhd.uned.es/ontology/postdata-dates/

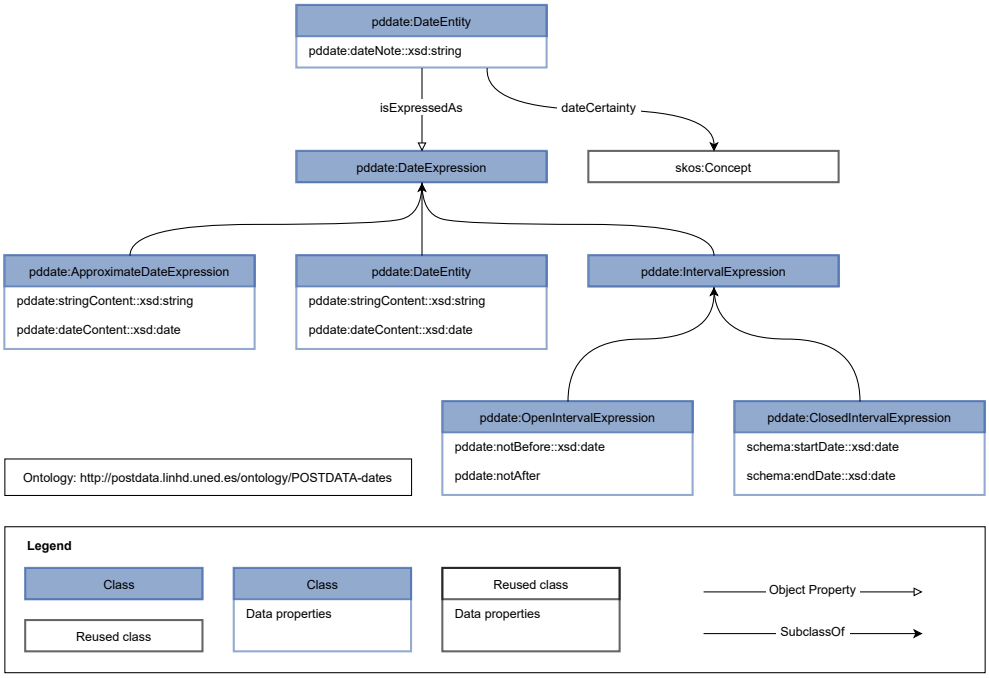


Figure 7: Postdata-dates ontology.

5 A sample application of the Postdata ontologies

To demonstrate the usefulness, versatility and user friendliness of the Postadata ontology V1.0, we present an example. This concerns the song “Mais nos faz Santa María a séu Fillo perdõar,” written by Alfonso X, el Sabio. An analysed redaction is extracted from “Cantigas de Santa Maria for Singers” (http://www.cantigasdesantamaria.com), which was created by Andrew Casson. The RDF implementation presents the poetic work, whose author is Alfonso X, along with the redaction prepared by Casson. The date is shown as an open interval since it is not known exactly. An extract of the structure is also given in stanza and line form. One point that should be highlighted is that this poetic composition is a song with a refrain. The refrain is the first stanza, but it repeats after each of the others. For this reason, the refrain has no assigned number, and its lines are indicated as part of the stanza. Finally, we show the work’s patterns, i.e. the stanzas and the lines.

The relationship between the poetic work and the redaction is represented through the object property pdcore:isRealisedThrough. The object property pdcore:hasCreator is used to denote the author of the poetic work and the creator of the redaction, who are both a type of agent(Person). All these items are related to pdcore:CreatorRole objects.

As such, the author of the poetic work is the person who has these features:

- Name (pdcore:name) – “Alfonso X”

- Nickname (pdcore:additional Name) – “el Sabio”
- A sentence linked to the person’s name indicating his position (pdcore:nameLink) – “Rey de Castilla”
- A link to a Virtual International Authorities File (VIAF) (schema:url) – http://viaf.org/viaf/66476694/#Alfonso_X,_Rey_de_Castilla

In contrast, the creator of the redaction is the person who has these features:

- Name (pdcore:name) – “Andrew”
- Surname (pdcore:additional Name) – “Casson”
- A Link to a personal website (schema:url) – https://independent.academia.edu/AndrewCasson

The date of the poetic work is shown as an open interval since it is not known exactly. It is associated with the poetic work through the object property date that refers to pddates:DateEntity. This is expressed as a pddates:OpenInterval with the following properties:

- notBefore – “1270-01-01”
- notAfter – “1282-12-31”

  In addition, because this work is a piece of medieval literature that is difficult to date, a note is created about who identified the date and how this took place:

- Note (pddate:dateNote) – “Dates set by Walter Mettmann.....”

In addition to the URL and the creator, the redaction is described in the structure of the text that it presents. In this regard, the redaction structure is a list of stanzas (pdstruct:OrderedStanzaList) that relate to the redaction via the property pdstruct:hasStanzaItem from the ontology of structural elements. The stanzas are described by the following properties:

- Stanza number, a positive integer indicating the stanza’s position in the list (pdstruct: stanzaNumber)
- If the stanza is a refrain, then
  - Is Refrain (pdstruct:isRefrain) – “true”
- If the stanza is not a refrain but must be followed by a refrain of which only one instance will be created
  - Is Refrain Omitted (pdstruct:isRefrainOmitted) – “true”
- The next stanza (pdstruct:nextStanza)
- The first and last line of the stanza (pdstruct:hasFirstLine, pdstruct:hasLastLine).

The lines of the text and other features are itemised below:

- Text of the line (pdstruct:content) – “a seu Fillo perdõar,”
- Position of the line (pdstruct:lineNumber) – “1”
- First and last line of the stanza (pdstruct:hasFirstLine, pdstruct:hasLastLine).

Prosodic features are dealt with at different levels. First, prosodic analysis of the redaction, stanza and line with the property (pdprosodic:isAnalysedThrough) take place respectively through the pdprosodic:WorkPattern object, pdprosodic:StanzaPattern object and pdprosodic:LinePattern object. The patterns are described as follows:

- pdprosodic:WorkPattern
  - If the redaction has a refrain (pdstruct:hasRefrain) – “true”
  - The metric scheme (pdprosodic:countingMetricalScheme) – “77777777”
  - The rhyme scheme (pdprosodic:rhymeScheme) – “ABAB/cdcdcdcdcB”
  - The number of stanzas (pdprosodic:numberOfStanzas) – “5”
- pdprosodic:StanzaPattern
  - The metric scheme (pdprosodic:countingMetricalScheme) – “7’ 7 7’ 7”
- pdprosodic:LinePattern (different for each line)
  - The metric scheme (pdprosodic:countingMetricalScheme) – “7”’
  - Phonetic transcription (pdprosodic:phoneticTranscription) – “;mais nos fadzsanta maRi.a”.

The rhyme is also represented by the pdprosodic:Rhyme class object associated with the line. Its properties include:

- The label (pdprosodic:label) – “A”
- The associated phonemes (pdprosodic:ending) – “i.a”

6 Conclusions and future work

In this paper, we have presented Postdata ontology V1.0., which was developed as part of the POSTDATA ERC Project.[19] This version of the ontology takes the form of a network that is comprised of four ontologies: the core ontology, prosodic ontology, structural ontology and date ontology. The first three ontologies represent the poetic work and its essential properties including prosodic elements. The fourth ontology contains information about dates. For literary

[19] http://postdata-prototype.linhd.uned.es/ontology.php

works, this information is particularly complex. The ontologies were published based on the best practices and recommendations for Linked Data vocabulary publishing.

Since developing this first version, we have begun to map poetry databases and repertories onto the ontology with the aim of populating the ontologies and sharing the information in an interoperable RDF format. We have also incorporated the ontology definitions into OMEKA[20] in the hope that this will be a straightforward and easy way to populate the ontology. This cultural heritage collection platform facilitates our support of researchers so they can use the ontologies effortlessly. On the other hand, its Open Source licence allows us to adapt the platform to the scientific description of the primary sources of the poetic texts, which can be found in archives and university libraries. The descriptive standards for these materials, which are reflected in the POSTDATA ontology network, respond to the needs of these institutions, who are the most regular users of the OMEKA system. Based on these processes, we plan to review the ontology descriptions and controlled vocabularies and take into account additional comments from the scholarly community. In this way, we should improve the representative capacity of these tools. It should not be forgotten that an ontology is only useful so long as it can accurately describe knowledge in the domain.

References

Bermúdez-Sabel, Helena, Mariana Curado Malta, and Elena González-Blanco (2017). "Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts". In: *Language, Data, and Knowledge*. Ed. by Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann. Lecture Notes in Computer Science. Springer International Publishing, pp. 156–165. ISBN: 978-3-319-59888-8.

Bermúdez-Sabel, Helena, María Luisa Díez Platas, Salvador Ros Muñoz, and Elena González-Blanco (July 15, 2019). "Towards a common poetry model: challenges and solutions". In: Digital Humanities Conference 2019 (DH2019), Utrecht. DOI: 10.5281/zenodo.3335509. URL: https://www.zenodo.org/record/3335509 (visited on 07/30/2019).

Curado Malta, Mariana, Paloma Centenera, and Elena González-Blanco (2017). "Using Reverse Engineering to Define a Domain Model: The Case of the Development of a Metadata Application Profile for European Poetry." In: *Developing Metadata Application Profile*, pp. 146–180. DOI: 10.4018/978-1-5225-2221-8.ch007.

Curado Malta, Mariana, Elena González-Blanco, Clara Martínez Cantón, and Gimena del Rio Riande (2016). "Digital repertoires of poetry metrics: towards a linked open data ecosystem". In: *Proceedings Digital Humanities and Digital Curation http://ceur-ws.org /Vol-1764/1.pdf*.

[20] https://omeka.org/

Curado Malta, Mariana, Elena González-Blanco, Clara Martínez Cantón, and Gimena del Rio Riande (2020). "Common Conceptual Model for the Study of Poetry in the Digital Humanities". In: *Proceedings of Digital Humanities https://dh2017.adho.org/abstracts /148/148.pdf* McGill University.

David, Ian and Newman, Richard (July 30, 2019). *Expression of Core FRBR Concepts in RDF*. URL: http://vocab.org/frbr/core (visited on 07/30/2019).

González-Blanco, Elena, Gimena del Rio Riande, and Clara Martínez Cantón (June 15, 2016). "Linked Open Data To Represent Multilingual Poetry Collections. A Proposal To Solve Interoperability Issues Between Poetic Repertoires". In: *Proceedings 5th Workshop on Linked Data in Linguistic*. Workshop on Linked Data in Linguistics (LDL-16) Managing, Building and Using Linked Language Resource located at 10th Cnference on Language Resources and Evaluation May 24 Portoro¸ Slovenia. Paris: ELRA, pp. 77–80. DOI: 10.5281/zenodo.60767. URL: http://e-spacio.uned.es/fez/eserv/bibliuned:363-Egonzalez3/LREC2016Workshop_LDL2016_Gonzalez_Blanco.pdf (visited on 12/05/2018).

González-Blanco, Elena and José Luis Rodríguez (2016). "ReMetCa: a TEI based digital repertory on Medieval Spanish poetry". In: *E-espacio UNED*.

Home | FRBRoo (2021). URL: http://www.cidoc-crm.org/frbroo/home-0 (visited on 03/08/2021).

Malta, Mariana Curado, Elena González-Blanco, Clara Martínez Cantón, and Gimena del Rio Riande (2017). "A Common Conceptual Model For The Study of Poetry In The Digital Humanities". In: *12th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2017, Montréal, Canada, August 8-11, 2017, Conference Abstracts*. Ed. by Rhian Lewis, Cecily Raynor, Dominic Forest, Michael Sinatra, and Stéfan Sinclair. Alliance of Digital Humanities Organizations (ADHO). URL: https://dh2017.adho.org/abstracts/148/148.pdf.

OWL - Semantic Web Standards (2021). URL: https://www.w3.org/OWL/ (visited on 03/08/2021).

Postdata (2020a). "Network of ontologies". In: *http://postdata.linhd.uned.es*.

Postdata (2020b). "POSTDATA Repertories". In: *https://cutt.ly/AhdZ12J*.

Postdata ERC project (Apr. 8, 2021). *POSTDATA Repertories*. Google My Maps. URL: https://www.google.com/maps/d/viewer?mid=15MAs3lVHlOk-eWUfBWXBP_prHbE (visited on 04/08/2021).

Suárez-Figueroa, Mari Carmen (June 25, 2010). "NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse". PhD thesis. Facultad de Informática (UPM). 288 pp. URL: http://oa.upm.es/3879/ (visited on 09/27/2018).

Tillett, Barbara (2005). "What is FRBR? A conceptual model for the bibliographic universe". In: *Aust. Libr. J* 54, pp. 24–30. DOI: 10.1080/00049670.2005.10721710.

https://doi.org/10.51305/ICL.CZ.9788076580336.03
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

The Automatic Quantitative Metrical Analysis of Spanish Poetry with Rantanplan: A Preliminary Approach

Laura Hernández-Lorenzo
Digital Humanities Innovation Lab, UNED, Spain
laura.hernandez@scc.uned.es
0000-0003-3489-2193

Mirella De Sisto
Digital Humanities Innovation Lab, UNED, Spain
mdesisto@scc.uned.es
0000-0002-0899-5976

Álvaro Pérez
Digital Humanities Innovation Lab, UNED, Spain
alvaro.perez@scc.uned.es
0000-0001-5897-1246

Javier de la Rosa
Digital Humanities Innovation Lab, UNED, Spain
versae@gmail.com
0000-0002-9143-5573

Salvador Ros
Digital Humanities Innovation Lab, UNED, Spain
sros@scc.uned.es
0000-0001-6330-4958

Elena González-Blanco
School of Human Sciences and Technology,
IE University, Spain
egonzalezblanco@faculty.ie.edu
0000-0002-0448-1812

Abstract

In this paper, we present a quantitative approach to Spanish poetry and versification based on the application of our own automatic metrical tool, Rantanplan, to the complete poetic works of four early modern Spanish poets. All of the poetry of these four representative authors—Garcilaso de la Vega (1503–1536), Fernando de Herrera (1534–1597), Luis de Góngora (1561–1627), and Lope de Vega (1562–1635)—was automatically processed and stress positions were extracted. Thanks to the development of a new stanza identification feature of Rantanplan, we were able to detect metrical structures as well. By completing a quantitative analysis of the stress positions, line lengths, and stanzas used by each author, we aim to model their complete metrical profiles.

1 Introduction

Applications of quantitative methodologies to poetry analysis have grown significantly in recent years driven by the emergence of automatic tools for verse

analysis. Although there have been fewer applications of these techniques to Spanish poetry, quantitative studies have already been carried out of the stress positions in sonnets and romances. We are, however, not aware of the existence of any computational studies that analyze stanza information extracted from the complete poetic works of an author.

In this article, we present a quantitative metrical analysis of four corpora of works by individual Golden Age Spanish poets. Using our own automatic scansion and syllabification tool, Rantanplan (De la Rosa et al. 2020a), which includes a new feature for stanza identification, we undertake the metrical annotation of the complete works of four canonical Spanish poets. These authors, Garcilaso de la Vega (1503–1536), Fernando de Herrera (1534–1597), Luis de Góngora (1561–1627), and Félix Lope de Vega (1562–1635), are selected based on the availability of their texts and their significance in this period. By studying and analyzing each poet's metrical choices related to stress positions, line lengths, and stanzas, we aim to model their complete metrical authorial profiles.

The paper is structured as follows: after this introduction, Section 2 contextualizes our research as part of an ongoing quantitative metrical investigation of poetry with a special focus on Spanish works. Section 3 then introduces our study, the dataset that we use, the process by which it was collected (Section 3.1) and the methodology that we employ (Section 3.2). Section 4 includes a quantitative metrical analysis which presents and discusses our results related to the stress positions (Section 4.1), line lengths (Section 4.2), and stanzas used by these authors (Section 4.3). Finally, our conclusions and proposals for future work are presented in Section 5. We also include an appendix of complementary materials.

2 Quantitative Approaches to (Spanish) Poetry Research

Since the emergence of what is today known as the digital humanities (Schreibman et al. 2004), traditional literary research has benefited from the application of new methods and techniques. In this regard, there has been a notable rise in quantitative approaches to the study of literary records, especially in light of theories of distant reading by Moretti (2003, 2013) and macroanalysis by Jockers (2013). As Burrows writes,"computer-assisted textual analysis can be of value in many different sorts of literary inquiry, helping to resolve some questions, to carry others forward, and to open entirely new ones" (Burrows 2004).

Turning to poetry, quantitative methodologies and Natural Language Processing techniques have been applied far less often than is the case with narrative texts. At first glance, the former genre has several limitations; among the most important are the generally shorter length of poetic texts and their subjective dimensions. Despite this, there has recently been significant scholarly attention to the automatic analysis of poetic works. We refer here in particular to several academic studies that apply these techniques, conferences and monographs dedicated to quantitative approaches to poetry studies (Plecháč et al. 2019), and a growing number of tools specifically developed for verse analysis. The latter include Natural Language Processing tools, which perform the automatic

scansion and syllabification of poems. However, to the best of our knowledge, there has not yet been any exploration of quantitative digital approaches based on a combination of line and stanza information.

Regarding literary traditions, most quantitative and automatic verse studies have addressed English language poetry, while fewer studies and tools are available for other European languages. In the case of Spanish poetry, several automatic scansion and syllabification tools have been developed since the first system built by Gervás (2000). Of these, Zeuscansion (Agirrezabal et al 2013), the ADSO Scansion system (Navarro-Colorado 2017), and Rantanplan (De la Rosa et al. 2020a) are the most recent. This rise in automatic tools has also encouraged quantitative metrical studies of Spanish poetry. These have mainly been carried out within ADSO[1] (Navarro-Colorado 2015, 2016) and the POSTDATA project[2] (Ruiz Fabo et al. 2020). Additional work has taken place within "Verse Rhythm in Golden Age Spanish Poetry: Lope de Vega and Luis de Góngora's Romances," a quantitative metrical project that does not use digital methods at the University of Neuchâtel:[3] (Llamas Martínez 2018; Sánchez Jiménez 2017).

The ADSO project collected Golden Age Spanish sonnets from 52 different poets. These sonnets were automatically annotated with metrical information. Based on the analysis of these texts, lead researcher Navarro-Colorado drew conclusions about each author's preferred rhythmic patterns when writing sonnets (Navarro-Colorado 2015, 2016). As part of the studies conducted in the POSTDATA project, a larger corpus of Spanish sonnets was collected. This featured poems from a broader time period —from the 15th to the 19th centuries—and drew on the work of more than 1000 authors. Researchers were able to use this corpus to analyze metrical patterns in post-Golden Age sonnets, resulting in some interesting findings (Ruiz Fabo et al. 2020). Sánchez Jiménez (Sánchez Jiménez 2017) and Llamas (Llamas Martínez 2018) have also studied a corpus of Lope de Vega and Góngora romances to detect metrical differences between authors working in this subgenre. In short, quantitative metrical research in Spanish poetry has mostly focused on sonnets and romances and neglected the analysis of other subgenres. Similarly it has not considered the entire poetic works of any author.

3 Our Study

This study aims to contribute to the growing number of quantitative metrical studies about the metrical choices of four canonical early modern Spanish poets. To this end, we explore the complete metrical profile of each author

[1] ADSO stands for "Análisis distante del soneto castellano de los Siglos de Oro" ("Distant Reading Approach to Golden Age Spanish Sonnets"). Led by Borja Navarro Colorado, this project was funded by a Fundación BBVA Digital Humanities grant. For more information, please visit the project website: http://adso.gplsi.es/index.php/en/adso-project/ [accessed: 19/11/2020].

[2] POSTDATA stands for "Poetry Standardization and Linked Open Data". Led by Elena González-Blanco, this project received funding from the European Research Council. For more details, visit http://postdata.linhd.uned.es/ [accessed: 19/11/2020].

[3] This project was led by Antonio Sánchez Jiménez and funded by the Fonds National Suisse de la Recherche Scientifique. For more information, please see: http://www.snf.ch/fr/encouragement/projets/projets-toutes-les-disciplines/Pages/default.aspx [accessed: 19/11/2020].

without restricting our analysis to any specific subgenre. This approach is made possible by the development of Rantanplan (De la Rosa et al. 2020a), which performs metrical analysis not only on hendecasyllables but on every type of Spanish verse. Moreover, Rantanplan's new feature of stanza and structure identification enables us to analyze the entire range of strophic and structural usage for each author.

3.1 Dataset

As an essential first step in this project, we obtain the complete poetic works of several early modern Spanish poets. In this regard, one of the limitations of quantitative approaches to Spanish literature has been the scarcity of digitized texts in suitable form. Both digital editions and repositories are lacking. As a result, our selections are based on the relevance of each poet's works in the early modern Spanish period and their digital availability. Our study uses the complete poetic works of four such authors: Garcilaso de la Vega (1503–1536), Fernando de Herrera (1534–1597), Luis de Góngora (1561–1627), and Lope de Vega (1562–1635). Garcilaso and Herrera are considered Renaissance poets while Góngora and Lope are major representatives of the Baroque style. In the case of three of these authors, their entire poetic corpus is available online; the corpus of the fourth poet (Herrera) has not yet been published. For Garcilaso and Lope, we use the source texts provided by Wikisource[4] and Cervantes Virtual library respectively.[5] For Góngora's poetry, we use the *gongocorpus*, a TEI-xml format corpus provided by Averell,[6] a tool developed within the POSTDATA project (De la Rosa et al. 2020b).[7] For Herrera's corpus, we use the complete poetic works digitized within Hernández-Lorenzo (2020).[8]

In order to prepare our texts for automatic metrical annotation —and to ensure there are no orthographic issues which could affect the analysis—, we standardize the orthography to contemporary Spanish and remove character names from dialogues along with rare symbols. We also exclude texts or fragments in languages other than Spanish from our analysis.

3.2 Methodology

To metrically analyze our corpora of early modern Spanish poets, we rely on digital humanities quantitative methodologies, as described in Section 2. We arrange for automatic scansion and syllabification with the open-source Python package Rantanplan (De la Rosa et al. 2020a). We also use Rantanplan's new

[4] Available at: https://es.wikisource.org/wiki/Autor:Garcilaso_de_la_Vega [accessed: 20/11/2020].
[5] Available at http://www.cervantesvirtual.com/obra-visor/poesias-liricas–0/html/ff775820-82b1-11df-acc7-002185ce6064.html [accessed: 19/11/2020].
[6] Available at https://github.com/linhd-postdata/averell [accessed: 19/11/2020].
[7] These source texts are in fact the digital editions of Góngora's poems prepared in the acclaimed POLEMOS project: http://obvil.sorbonne-universite.site/corpus/gongora/ [accesed: 11/11/2020].
[8] These texts were digitized using the most complete and widely respected edition of Herrera's poems (Herrera 1975). As some of his works raise questions about authorship and others contain variants of the same text, we limit our analysis to poems and versions of texts published in the author's lifetime in *Algunas obras* (Herrera 1582).

Author	Poems	Words
Garcilaso	60	26.182
Herrera	91	18.685
Gongora	320	74.564
Lope	51	21.707

Table 1: Corpora of authors in our study including the number of poems and words in their complete poetic works

stanza and structure identification feature, which works as follows: leveraging information about the metrical syllables and stressed endings in each line, our algorithm assigns a customizable window to each rhyme per line. Both metrical lengths and rhyme patterns are then fed to a pattern-matching algorithm which uses recurring expressions to search for matches within a growing knowledge base of structures and stanzas. The most comprehensive match is then returned in both machine and human-readable formats, enabling scholars to pursue further investigations.

We annotate each plain text corpus in two different formats: the first contains only stress information while the second combines stress and stanza annotation. We then use a Python script to extract and calculate quantitative metrical data for each author and compare the results, considering stress position, verse length, and stanza types.

4 Quantitative Metrical Analysis

A quantitative metrical analysis can give us a broad picture of a poet's metrical choices and help determine whether they arise from a particular poetic school or from individual preferences. In this paper, we first compare the distribution of line stresses across the complete poetic corpora of Garcilaso, Herrera, Góngora, and Lope. By analyzing recurring stress patterns and the proportion of particular stress positions in each corpus, we determine the authorship features of the four poets. We then turn to line lengths to assess how these differ for the individual authors; these characteristics can also help identify some authorship traits. Lastly, we analyze stanza use in the poetic works of the four authors. A quantitative analysis of the distribution of stanza types across the corpora can improve our understanding of these authors' use of different poetic forms.

4.1 Stress Positions

This section analyzes the distribution of stressed metrical positions across the four corpora. We refer here to metrical positions and not to syllables because certain phenomena may affect the one-to-one correspondence of these two elements. The use of a synaloepha, an extremely common feature in Spanish poetry, may, for example, cause two syllables to count as one metrical position.

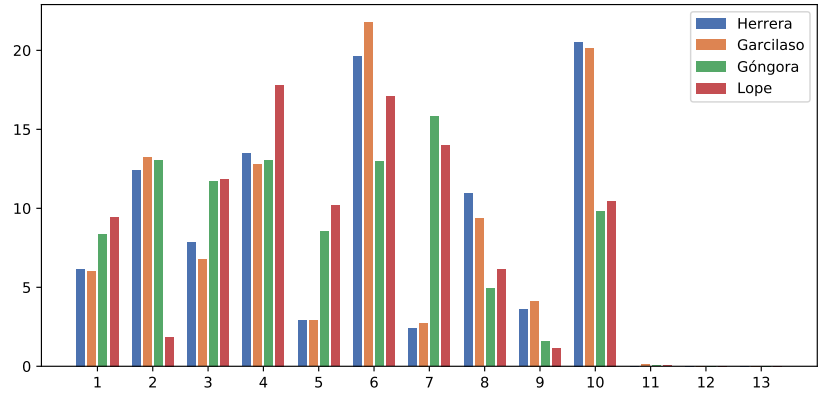


Figure 1: Distribution of stressed metrical positions per line in the complete poetic works of Garcilaso, Herrera, Góngora, and Lope

Rantanplan (De la Rosa et al. 2020a) detects these phenomena and calculates the number of metrical positions accordingly.

On the one hand, this comparison of the distribution of stress positions across the corpora highlights the poetic styles used by these authors and the literary schools to which they belong. On the other, the data reveal these poets' individual preferences regarding metrical position stress; these preferences are not strictly related to the style that they follow.

As shown in Figure 1, based on their stress patterns, the four authors can be split into two groups that respectively represent the Renaissance and the Baroque styles: Garcilaso and Herrera show similar patterns and are both Renaissance figures; Góngora and Lope constitute the Baroque group.

The most striking difference between the two groups relates to the frequency of stressed syllables in the seventh and tenth positions. Góngora and Lope often place a stress in the seventh position, while Garcilaso and Herrera tend to avoid this; the opposite scenario applies in the tenth position, which is strongly stressed in the corpora of Garcilaso and Herrera but is not similarly prominent in the works of the other two authors. Besides showing a marked tendency to stress the tenth position, the Renaissance group are also prone to stressing the sixth position. The co-occurrence of stresses in the sixth and tenth positions suggests that Garcilaso and Herrera mostly composed hendecasyllables *a maiore*. The popularity of hendecasyllables in these two authors' corpora is also evident when considering line-length alone, as we will see in Section 4.3. The predominance of the *a maiore* pattern corroborates the findings of previous traditional and computational studies of the Renaissance hendecasyllable (Domínguez Caparrós 2014; Henríquez Ureña 1919; Navarro-Colorado 2015, 2016; Ruiz Fabo et al. 2020). In fact, there is a clear preference for patterns based on the 6-10 scheme (e.g. 2-4-6-10, 2-6-10) among Renaissance poets (for a detailed account, see Navarro-Colorado (2016)). This preference decreased in Baroque era when the most frequent pattern was 2-4-8-10 (Navarro-Colorado 2016).

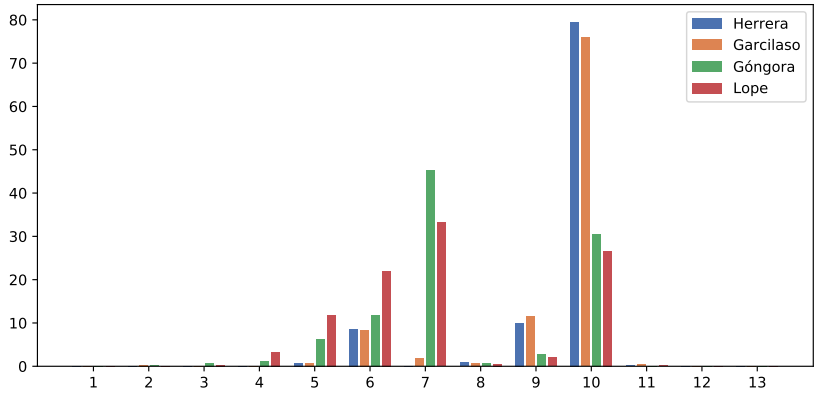


Figure 2: Distribution of line lengths based on the number of metrical positions in the poetry of Garcilaso, Herrera, Góngora, and Lope

By determining the stress pattern for individual authors, we can also identify distinctive features in their work. Lope's poetry stands out for its avoidance of a stress in the second metrical position in favor of one in the third position, while Góngora and the two Renaissance poets seem to prefer to stress the second position. This difference between Góngora and Lope can be explained by considering that, as Navarro-Colorado (2016) observe, the stress in the third position gradually became more widespread during the Baroque period and was a particularly prominent trait in the later stages of this era. Since Góngora is one of the first Baroque poets, this practice is not yet visible in his verse, while it is evident in a later poet like Lope.

A further trend that distinguishes Lope from Góngora is that the former prefers to stress the sixth position while the latter opts more often for a stress in the seventh position.

Finally, even though Garcilaso and Herrera turn out to have fairly similar metrical patterns, they each have some distinctive traits: for Garcilaso, the second, sixth, and ninth positions are stressed slightly more frequently than they are for Herrera; on the other hand, the third, fourth, and eighth positions are stressed more often for Herrera than they are for Garcilaso.

4.2 Line Lengths

Measuring the distribution of line lengths allows us to identify the most represented poetic forms in each of the corpora. Line lengths are calculated based on the number of metrical positions per line. As shown in Figure 2, we can identify two groups based on these line-length distributions: Renaissance and Baroque poets.

The clearest distinction between the two groups concerns the tenth metrical position. In the Renaissance group, the vast majority of lines have ten metrical positions, which suggests that these authors mostly wrote hendecasyllables. In contrast, Baroque authors show no preference for any line length; their

poetry contains lines of diverse lengths of which the most common has seven metrical positions. Lope's poetry shows the greatest variation in terms of line length.

4.3 Stanza Use

In terms of stanza detection, we obtain quantitative results by using Rantanplan (De la Rosa et al. 2020a) to calculate the proportion of each type of stanza for each author as well as the total number of stanzas in his work (see the repository table mentioned in the Appendix). Of the 47 known stanza types, 20 are not used by any the four early modern Spanish poets.[9] Some of these types including the *estrofa manriqueña, copla de arte mayor, copla de arte menor, copla real, copla mixta, décima antigua*, and *cuaderna vía* are typical of the medieval period. Another undetected type, the haiku, was not employed until recently in Spanish poetry.[10]

The rest of the stanza types are used by at least one author; only five, i.e. sonnets, couplets, *octavas, serventesios*, and *tercetos* are used by all four poets. Garcilaso and Herrera write *liras*, whereas Góngora and Lope favor *sextetos liras*. In Lope's case especially, this form comprises 17.05% of detected stanzas. The two Baroque poets also use some popular stanza forms such as *seguidillas, septillas* and *romances*, which are not detected among the Renaissance poets. The *terceto* is the most prevalent type for three of the authors, while more than 50% of the stanzas detected in Góngora's work are *cantares*, also known as *coplas*. Another stanza quite typical of the period, the *octava real*, is heavily used by Garcilaso and Góngora but not by Herrera and Lope. These results accord with our earlier observations about stress positions and line lengths. In particular, the popular stanza forms of the Baroque era and the *cantares* are non-hendecasyllabic. All this correlates with the reduced use of this type of verse by Góngora and Lope.

Of the most popular stanza types, sonnets and *tercetos* are of particular interest given their prevalence in the works of many of our authors. It is, thus, worth taking a closer look at the use of these two forms.

Figure 3 compares the relative frequency of sonnets across the works of the four poets. Lope favors this stanza type, which constitutes over 17% of his detected stanzas. Herrera is the next most frequent user with a rate of nearly 15%. In contrast, sonnets are far less common for Garcilaso and especially Góngora, with a proportion below 10% in both cases.

As noted above, the *terceto* is the most frequent stanza type in the work of three of the four poets. However, the extent of its use varies among those authors. While for both Garcilaso and Herrera, nearly 70% of detected stanzas

[9] These undetected types are the *copla de arte mayor, copla de arte menor, copla real, cuaderna vía, décima antigua, endecha real, estrofa manriqueña, estrofa sáfica, haiku, octava mixta, octavilla, ovillejo, quinteto, seguidilla compuesta, seguidilla gitana, septeto lira, sexteto, silva arromanzada, soleá*, and *terceto monorrimo*.

[10] Notably, other stanzas and structures undetected among the four poets include hendecasyllabic stanzas and structures such as the *estrofa sáfica*, *terceto monorrimo*, *quinteto*, *sexteto*, *septeto lira*, *endecha real*, and *silva arromanzada*. Regarding popular stanzas and structures, the *octavilla*, *ovillejo*, *seguidilla compuesta*, *seguidilla gitana*, and *soleá* are also unused.

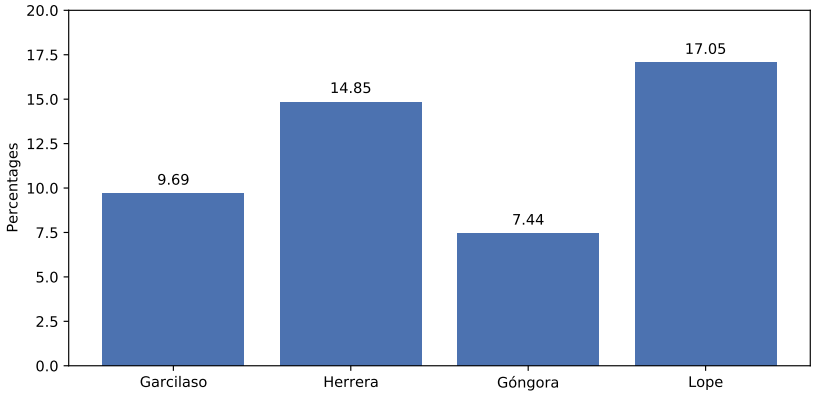


Figure 3: Proportion of sonnets used. Lope relies on sonnets most often (in more than 17% of his detected stanzas) followed by Herrera (almost 15%). Garcilaso and Góngora both employ this structure to a lesser extent (in less than 10% of detected stanzas in both cases).

are *tercetos*, over 78% of Herrera's poetic output is identified as such. On the other hand, this form is far less common for the two Baroque poets. Indeed, *tercetos* account for far less than 50% of Lope's works and do not even represent 3% of the stanzas detected in Góngora's works.

5 Conclusions and Future Work

This paper has applied a quantitative methodology to the metrical analysis of Spanish poetry, with a focus on four representative authors of the early modern period. We have relied specifically on the application of our automatic scansion and syllabification tool, Rantanplan. Thanks to the development of a new stanza identification feature in this software, we were able to perform a complete analysis of the metrical choices of the selected authors. By analyzing quantitative data about stress positions, line lengths, and stanzas, we aimed to model their metrical authorial profiles. Our results confirm previous hypotheses and provide new insights.

The predominance of *a maiore* patterns and the preference for patterns based on a 6-10 scheme correlate with the findings of earlier traditional and computational research about the Renaissance hendecasyllable. In addition, even where differences in stress positions delimit two distinct group of poets, i.e. Renaissance and Baroque figures, there are enough unique features to identify each of the authors.

Regarding line length, our results show that the two Renaissance authors mostly write in hendecasyllables while their Baroque counterparts exhibit a wider range of verse types. This is even more apparent in our quantitative stanza analysis, which confirms that shorter-verse stanzas become more popular among the Baroque poets;*cantares* are, for example, strikingly frequent

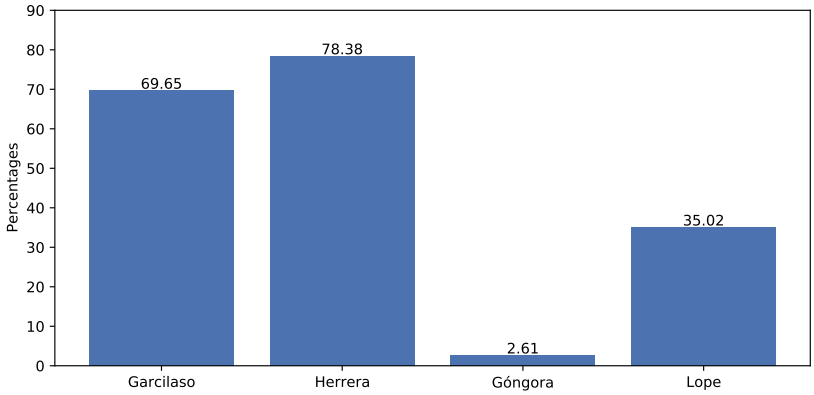


Figure 4: Proportion of *tercetos* used. Herrera most clearly favors this stanza form (in more than 78% of his detected stanzas), followed by Garcilaso (almost 70%). The proportion oftercetos then drops for Lope (to around 35%) and plunges even further for Góngora (to less than 3% of his detected stanzas).

in Góngora's works. Likewise, a close examination of the most popular stanza, the *tercetos*, makes clear that this generally hendecasyllabic stanza was far less common among Baroque than among Renaissance authors.

Significantly, all these findings are restricted to a small number of authors. As such, in future work, we would like to expand the corpora of poets with the goal of corroborating the data obtained in this study with those for other Renaissance and Baroque authors.

Acknowledgments

Research for this paper was made possible thanks to the Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) http://postdata.linhd.uned.es/ led by Elena González-Blanco. This project is funded by the European Research Council (https://erc.europa.eu) (ERC) as part of the European Union's Horizon2020 research and innovation program.

References

Agirrezabal et al (2013). "ZeuScansion: A Tool for Scansion of English Poetry". en. In: *Journal of Language Modelling* 4.1, p. 3. ISSN: 2299-8470, 2299-856X. DOI: 10.15398/jlm.v4i1.102. URL: http://jlm.ipipan.waw.pl/index.php/JLM/article/view/102 (visited on 07/03/2020).

Burrows, John Frederick (2004). "Textual analysis". In: *A Companion to Digital Humanities*. Ed. by S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Blackwell, pp. 323–347.

De la Rosa et al. (2020a). “Fast and Accurate Syllabification and Scansion of Spanish Poetry”. In: *Revista de Procesamiento del Lenguaje Natural* 65.
De la Rosa et al. (2020b). “PoetryLab as Infrastructure for the Analysis of Spanish Poetry”. In: *Proceedings of CLARIN Annual Conference 2020*, pp. 82–87. URL: https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.
Domínguez Caparrós, J. (2014). *Métrica española*. Madrid: UNED.
Gervás, Pablo (2000). “A Logic Programming Application for the Analysis of Spanish Verse”. In: *Computational Logic – CL 2000*. Lecture Notes in Computer Science. Springer. DOI: 10.1007/3-540-44957-4_89.
Henríquez Ureña, P. (1919). “El endecasílabo castellano”. In: *Revista de Filología Española* 6, pp. 132–157.
Hernández-Lorenzo, Laura (2020). “Los textos poéticos de Fernando de Herrera. Aproximaciones desde la estilística de Corpus y la Estilometría”. PhD thesis. Spain: Universidad de Sevilla.
Herrera, F. de (1582). *Algunas obras de Fernando de Herrera*. Casa de Andrea Pescioni.
Herrera, F. de (1975). *Obra poética*. Ed. by J. M. Blecua. Boletín de la Real Academia Española.
Jockers, Matthew L. (2013). *Macroanalysis. Digital Methods and Literary History*. University of Illinois Press.
Llamas Martínez, Jacobo (2018). “Métrica, ritmo acentual y autoría en la poesía española del Siglo de Oro”. In: *Monográfico de Arte Nuevo* 5, pp. 49–214.
Moretti, Franco (2003). “Graphs, Maps, Trees: Abstract Models for Literary History 1”. In: *New Left Review* 24, pp. 67–93.
Moretti, Franco (2013). *Distant Reading*. London: Verso.
Navarro-Colorado, Borja (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*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 105–113. DOI: 10.3115/v1/W15-0712. URL: https://www.aclweb.org/anthology/W15-0712 (visited on 05/17/2020).
Navarro-Colorado, Borja (2016). “Hacia un análisis distante del endecasílabo áureo: patrones métricos, frecuencias y evolución histórica”. In: *Rhythmica*.
Navarro-Colorado, Borja (2017). “A metrical scansion system for fixed-metre Spanish poetry”. In: *Digital Scholarship in the Humanities* 33.1, pp. 112–127.
Plecháč et al., ed. (2019). *Quantitative approaches to versification*. Institute of Czech Literature of the Czech Academy of Sciences.
Ruiz Fabo et al. (2020). “The Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings”. In: *Digital Scholarship in the Humanities*. DOI: https://doi.org/10.1093/llc/fqaa035.
Sánchez Jiménez, Antonio (2017). “Acentos contiguos en los romances de la Arcadia (1598), de Lope de Vega”. In: *Atalanta: revista de las letras barrocas* 5.1, pp. 5–61.
Schreibman, S., R. Siemens, and J. Unsworth (2004). *A Companion to Digital Humanities*. Oxford: Blackwell.

Appendix

To see a complete table of relative values, i.e. percentages, for the results of stanza and structure detection across the four corpora, please visit the repository at https://github.com/linhd-postdata/Plotting_Poetry2020. This repository also contains the corpora used in this project, with the exception of Herrera's texts, which have not yet been published. Herrera's corpus annotations are, however, included.

https://doi.org/10.51305/ICL.CZ.9788076580336.04
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

Rhyme in 16th-Century Hungarian Historical Songs: A Pilot Study

Szilvia Maróthy
RCH Institute for Literary Studies, Hungary
mthy.szilvi@gmail.com
0000-0003-2558-9504

Levente Seláf
ELTE University, Hungary
levente.selaf@gmail.com
0000-0002-6052-9841

Petr Plecháč
ICL, Czech Academy of Sciences, Czechia
plechac@ucl.cas.cz
0000-0002-1003-4541

Abstract

This article presents a computer-based stichometric analysis of 26 Hungarian historical songs from the 16th century. We explore the validity of comments made by Albert Szenci Molnár in 1607 about the poor quality and simplicity of stanza structures in the poetry of previous generations. The study shows how rhyming changed in this poetic genre between 1539 and 1598. In this respect, it is the first work to explore these changes through a quantitative analysis. We find that during the examined period, there was a marked decline in the frequency of rhymes based on the repetition of the same word. At the same time, the tendency to maintain a rhyme across multiple stanzas did not change significantly.

1 Introduction

In the summer of 2020, our team began a three-year research project dedicated to the computer-based analysis of Hungarian poetry of the 16th century.[1] The main purpose of this investigation is to provide a preliminary stylometric and stichometric analysis of this corpus.

The project focuses on the oral and written patterns in literature, a matter that has raised much debate. While most poems of the period are closely related to written culture (or their acrostics and colophons at least suggest this link), they have equally strong rhetorical and stylistic components that connect them to oral performance. In addition,they were treated as a type of

[1] The members of our team are senior researchers Margit Kiss, Szilvia Maróthy, Petr Plecháč, Levente Seláf, and Artjoms Šeļa and students Mária Finta, Villő Vigyikán, and Zoé Zohó-Tóth. The authors warmly thank all the participants for their contributions to this article.

communal poetry by traditional scholarship and early theoretical reflections (see Sir Philip Sidney's comments in his *Defence of Poesie*, also cited in Seláf (2020)). Our research considers both the inherent oral stylistic patterns and their imitations in this literature. We believe that a stylistic and digital metrical analysis may reveal the structure of multilayered editions,including the textual interventions and modifications of various authors and editors (see Vadai (2009) and Schelhammer (2019) on traces of the different redactions of poems in the given corpus). In this paper, we focus exclusively on a stichometric analysis.

One of the main goals of stylometry and computerized metrical analysis is authorial attribution. In the case of our corpus, however, instead of problems related to doubtful attribution, our concern is with whether the stylistic and metrical features and practices of an author, editor, publisher, or genre are homogeneous. In addition, by examining the unity of texts in the same genre or that were edited or reproduced in the same volumes or by the same persons, our project provides feedback on the validity of the stylometric method regarding authorship in the corpus.

This article is not the first study to employ a data-driven analysis of early modern Hungarian poems. While working on the *Repertory of Early Modern Hungarian Verse* (RPHA (Horváth et al. 1979–2020)) in the 1980s and early 1990s, Iván Horváth offered a statistical examination and summary of the core characteristics of early modern Hungarian verse based on the so-called iso rule (Horváth 1991). This arose from the observation that the first verse and the first stanza formally determine all subsequent verses and stanzas. This primarily metrical examination helped researchers understand the homogeneity of the corpus, a certain formal simplicity and uniformity that remained characteristic in Hungarian poetry up until the 19th century. Related studies have analyzed groups of poems that significantly differ from the rest of the corpus from different perspectives. This includes, for instance, poems that start with "ab" rhymes (Horváth 1991; Szigeti 2005), those in couplets (Bognár 2016), and those affected by the "+1 rule" concerning metrical boundaries and regularities in oral and written poetry (Vadai 1991, 2012). Theories have been proposed about characteristic features that counterbalance formal homogeneity, for example, by phonetic means (Horváth 1991, 2006) or by constructing complicated acrostics and anaphoric structures that compensate for monotonous rhyme schemes and syllabic structures (Seláf 2017). Hapaxes and poems that diverge from the homogeneous sample have also undergone thorough examination; their technical uniqueness has partially been explained by foreign, mostly Latin and German, influences (Bognár 2016).

2 The Historical Song in 16th-Century Poetry

The project consists of two major phases, each of which uses an important subcorpus of poetry of the period. The first phase of research examines the most characteristic genre of Hungarian literature: the *historical* (narrative or epic) *song*. Our inquiry focuses mostly on historical songs of the 16th century, but some 17th-century compositions of the same genre that have not previously

sparked much academic interest are also analyzed. The second phase of the project will examine secular lyric poems of the second half of the 16th century. The present study of rhyme patterns in a select sample of history songs marks our first effort to explore this domain.

The historical song is a traditional genre thought to have emerged in the 15th century or earlier.[2] While it cannot be confirmed for certain, it appears likely that the the earliest and most important themes of these songs were the past acts of Hungarians. The first known major poet whose literary output focused on ancient and more recent Hungarian history was Sebestyén Tinódi, who was active in the 1540s and 1550s. In 1554, an entire songbook of Tinódi's work entitled *Cronica* was published under his supervision. The popularity of Tinódi's book and his melodies suggests that his work was an enduring stylistic model for historical songs in Hungary. Although the stanzaic forms used by Tinódi were not of his own invention, his epic songs on historical subjects became the stylistic, musical, and metrical models for later narrative poems on a wide range of themes. These extended from love stories to biblical plots. While more broadly speaking, the historical song genre can be divided into different subgenres based on the themes explored in the songs. From a formal point of view, they are all very similar. Although several genres in the Hungarian poetic tradition reflected trends in European poetry of the Renaissance, especially the Protestant Reformation (e.g. psalm paraphrases), the history song remained popular over an extended period. We can, thus, trace a trajectory from its first surviving specimens in the late 15th century to its gradual disappearance during the 17th century.

3 Historical Criticism of Early Hungarian Versification

In the foreword to his book of psalm paraphrases published in 1607, Albert Szenci Molnár harshly criticized the versification techniques of earlier Hungarian poets. This was the first serious analysis of poetical forms, especially rhyming, in Hungarian literature.

> Az régi Magyar énekekben pedig avagy semmi egyenlö terminátioc nem voltac, avagy tiz versis egy másután mind egy igében ment ki, àhonnan az historias énekekben, számtalan az soc Vala vala vala. Kin az idegen nemzetec az kic ezt láttyác, nem gyöznec eleget rayta nevetni. De hálá Istennec, ez egynéhány esztendökben az mi emberinkis ékesb verseket szoktac irni. (http://magyar-irodalom.elte.hu/gepesk/eloszo/12/12.html)
>
> *In old Hungarian songs, there were either no rhyming line endings or ten consecutive stanzas/verses ended with the same rhyme. That's why we have countless "vala, vala, vala" [as the rhyme word] in historical songs. Foreign nations who see this cannot stop laughing about it. But thank God, in recent years, some of our people have been writing more ornate stanzas.*

[2] *Szabács viadala*, the earliest Hungarian text devoted to a historical event, dates from the last quarter of the 15th century, and its versification bears traces of German literary influence. We assume that there were other historical songs in this period, but we have no information about them.

Szenci Molnár's criticisms have been cited by many recent scholars, who propose diverse interpretations. Some suggest that Szenci Molnár essentially condemned the use of identical rhyme endings in several consecutive (monorhymed) stanzas (Ötvös et al. 2004). While this may be true, it is worth noting that he also criticised the over-use of rhymes based on repetition of the same word, particularly "vala" ('has been'). This pattern was typical of the history songs of Tinódi and his imitators, works which Szenci Molnár believed to lack artistic merit.

In the analysis that follows, we assess the validity of Szenci Molnár's critique and examine the formal and stylistic characteristics of Hungarian epic verse in general.

4 Data and Annotation

For this first phase of this project, we selected 26 poems that dated from the period between 1539 and 1598. In most cases, these works could be dated precisely based on their colophon stanzas.[3] Nineteen of these poems are attributed to specific authors (11 different persons in total), while seven poems are anonymous. The two most represented authors in our corpus were Sebestyén Tinódi, who was active in the first part of the period (with three poems dating from between 1550 and 1553), and István Bogáti Fazakas, who worked toward the end of the period (with seven poems dating from between 1576 and 1598). They are both major figures in the genre, and in a later phase of the project, we intend to expand the corpus to include the rest of Tinódi's history songs. The texts used for our analysis were mostly taken from the digital anthology *Early Modern Hungarian Literature* (Jankovics et al. 2000) and partly from the critical edition of Bogáti's works (Ács et al. 2018). The shortest text was a fragment from *Szilágyi és Hagymási históriája* of 24 stanzas (the lost entire work has an estimated length of 50 stanzas), while the longest one was *Eurialus és Lucretia históriája* (504 stanzas).

Except for one work by Bogáti, each poem in the corpus was isostrophic, and each stanza was isorhymed and isosyllabic.[4] Six of the poems were written in tercets and 20 in quatrains. The different meters used ranged from 10-syllable to 19-syllable lines, with the most frequent structures being monorhymed quatrains of 12-syllable or 11-syllable lines, with eight occurrences of each type. One poem by Bogáti contained a first part with 11-syllable lines and a second part with 12-syllable lines. Overall the limited variety of the corpus reflected the formal homogeneity of 16th-century Hungarian poetry rather well.

The poems in raw TXT were tokenized, lemmatized, and morphologically analyzed by means of the `emtsv` system (Indig et al. (2019) also known as

[3] See the list of selected poems and associated key data on GitHub, listofpoems.csv. The following three works were not precisely datable although they certainly belong to the same period: 1) *Toldi Miklós históriája*; 2) *História egy Argirus nevű...*;and 3) *Szép históriás ének az Telamon királyról*. Since, however, we wished to trace trends in changing rhyming schemes, these three undatable poems were not included in the figures.

[4] This meant that (1) all stanzas had the same syllabic and rhyme structure, (2) every line in a stanza rhymed with the others,and (3)all lines in a stanza were of the same length.

E-magyar; Váradi et al. (2018)). In addition, they were phonetically transcribed using the eSpeak synthesizer. Rhyme recognition was provided by a simple rule-based algorithm, which on inspection turned out to be reliable enough for our needs. Given that a mere match of line-final vowels and/or line-final suffixes was thought sufficient to establish a rhyme in this period, we considered two lines to rhyme with each other if (1) they came from the same stanza and (2) their final vowels and/or their final morphemes were identical.[5] The output was stored in JSON format. Initially, we encountered serious difficulties when analyzing the morphology of rhyming words because the emstv analyzer was incapable of correctly interpreting some old morphological models with the default vocabulary file. For this reason, we used the special vocabulary for Old and Middle Hungarian developed by Attila Novák and Bálint Sass (Novák 2014; Novák et al. 2016). Based on a manual check of a random sample of 300 line-ending words, we estimated accuracy for our data at approx. 0.9.

5 Results

5.1 Rhymes Between Stanzas and Identical Rhymes

Szenci Molnár's critique of the proliferation of same-rhymed endings has been understood by most scholars as a comment on the over-use of same-rhymed endings in consecutive stanzas; this was, for example, the interpretation of Ötvös et al. (2004) while Iván Horváth (2009) held that the monorhymed structure had been the target of Szenci Molnár's criticism. To test these claims, we measured the average sequence length in each poem where a sequence was defined as a group of successive stanzas that shared the same final vowel in most of their lines. These values were compared to a theoretical model constructed by randomizing the order of stanzas in each poem 10,000 times. As Figure 1 shows, the differences between the expected and observed values were not statistically significant for most of the poems ($\alpha = 0.05$). We were therefore unable to confirm any overall tendency to group stanzas based on their line endings. These differences in value seemed to decrease over time. However, given the lack of statistical significance, this observation has rather limited relevance.

On examining the observed values on their own, we noted a number of upper outliers in both the first and second half of the period. These included two poems by Tinódi where the rhyme spanned more than 2.5 stanzas on average (*Zsigmond császárnak históriája*, 1552, RPHA 1495 (2.70) and *Egri históriának summája*, 1553, RPHA 1292 (2.59)) and another three poems by other authors where it continued over more than 2 stanzas (Ráskai Gáspár: *Egy szép história az vitéz Franciskórúl*, 1552, RPHA 0322 (2.07); Bogáti: *Ez világi nagysok zűrzavarról*, 1586, RPHA 1158 (2.05); Bogáti: *Aspasia asszony*, 1587, RPHA 0693 (2.03)). In fact, these results suggest that at least with respect to historical songs, there was no material change in this versification technique in the period preceding Szenci

[5] The following vowels were treated as equivalent: [e] = [ɛ] and [a] = [ɒ]. Vowel length was not taken into account.

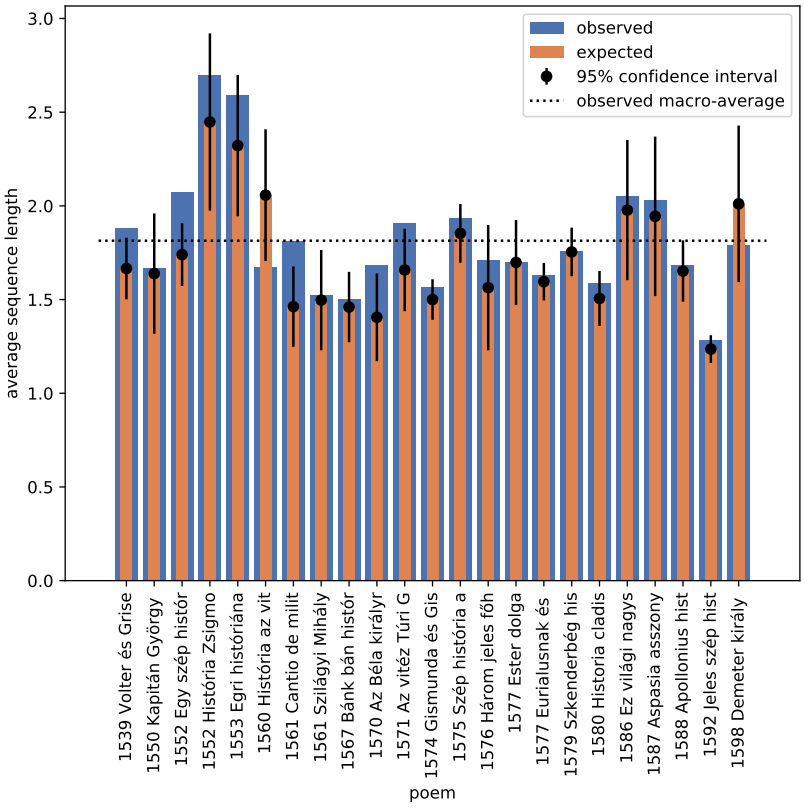


Figure 1: Average observed length of stanza sequences with the same rhyme endings compared with expected length based on a random distribution of all stanzas

Molnár's comment; our analysis of the sample simply did not support such a trend. We must therefore infer that if this change occurred, either it applied to other genres or it related, as Horváth believed, to the technique of building monorhymed stanzas and not the mere use of same rhyme-endings in successive strophes.[6]

However, Szenci Molnár's remark also applied to situations where the rhyme was achieved through the repetition of the very same word. We refer to this phenomenon in this paper as an identical rhyme, and it describes a case where the meaning, spelling, and phonetic form were all the same, e.g.:

> Nagy had vala régen Görögországban,
> Két szomszéd tartomány öszvevívásban,
> Phocis és Thessalia nevek *vala*,
> Ez két had immár szembeszállott *vala*.
> (Bogáti: *Szép história az tökéletes asszonyállatokról*, v. 21–24.)

[6] In our corpus of historical songs, the non-monorhymed Balassi-strophe (6aa7b6cc7b6dd7b) was limited to a single sample from the 16th century. This was Márton Gyulai's *Epicinia* (RPHA-1029).

Sőt az veszedelmet ha ők értenék,
Az asszonynépeket mind ott égetnék,
Ez tanács az egész hadnak mind *tetszék*,
Egy ember lőn, kinek jobb tanács *tetszék*.
(Bogáti: *Szép história az tökéletes asszonyállatokról*, v. 37–40.)

Significantly, Szenci Molnár singled out the rhyme word "vala" as typical of these poorly rhymed poems. From our analysis, it also appeared that the repetition of this word was extremely popular in rhymes in this period. A total of 1167 of the 1696 identical rhymes, i.e. almost 69%, consisted of repetition of "vala") and it seems likely that the authors understood this as a distinct style pattern. Figure 2a shows the proportion of identical rhymes involving "vala"–"vala" rhymes among all the rhymes found in particular poems, while Figure 2b shows the proportion of lines ending with "vala".

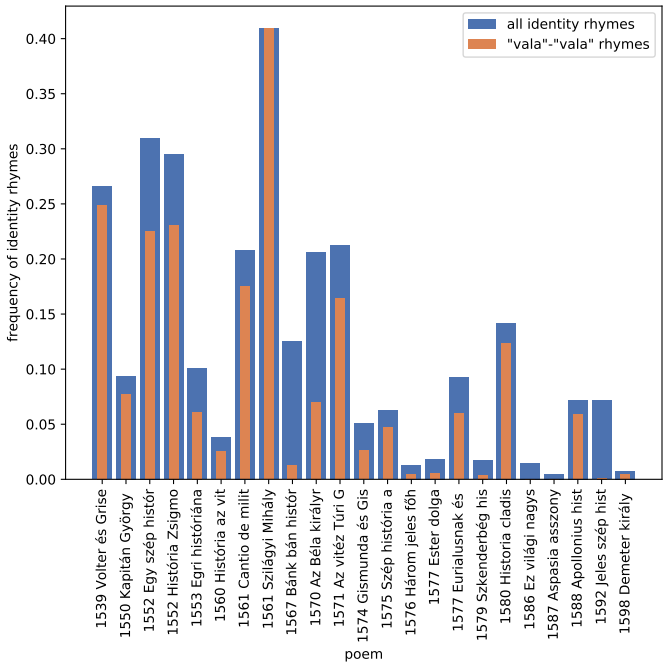
In fact, "vala" may be understood as a special case of *suffix rhyme* (see Section 5.2). The verb form "vala" is the third-person singular form of the verb "to be" in the past perfect tense, but it also functions as a suffix to express the past perfect or the past imperfect tense of other verbs. In that case, it is written separately and follows the conjugated verb (Verb + PERF.3SG be-PAST or IMPERF.3SG be-PAST). Critics have condemned historical songs for their over-reliance on "vala", a tendency that likely results from the genre's general focus on past events and the easy solution these verb forms offer for rhyme.

Figure 2 shows that the use of "vala" rhymes is a salient feature of Tinódi's poems, however it is even more marked in some other poems. It is, for instance, the only identical rhyme appearing in the anonymously authored *Szilágyi és Hagymási históriája*, which has the highest proportion of identical rhymes (41%) of any work in the corpus.[7]

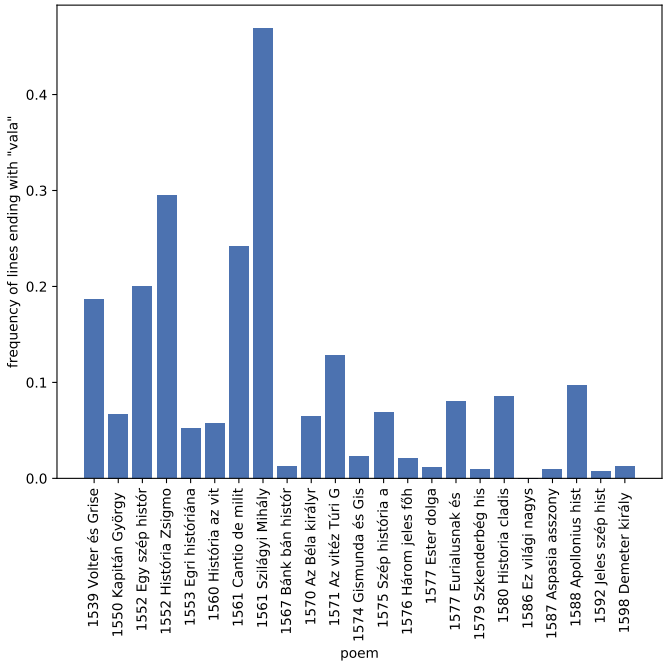
The trend over time is clear: there is an ongoing decrease in the use of identical rhymes, and this decline can also be approximated fairly well as a linear function ($r^2 = 0.41$; Figure 3). The pattern becomes even more striking when we compare the two most represented authors in the corpus: Tinódi, who wrote earlier in the period and employed these rhymes considerably, and Bogáti, who was active at the end of the century and used some identical rhymes in his earliest historical song *Szép história az tökéletes asszonyállatokról*, 1575) but hardly any in his later compositions.

Notably Bogáti not only reduced the percentage of identical rhymes in his poems, but apparently avoided the "vala" rhyme as much as possible, and even extended this stance to the very similar "volna", which is part of the conditional verb form [Cond.NDef.3Sg]. There are no "vala" rhymes whatsoever in *E világi nagy sok zűrzavarról való ének* and only two occurrences of "vala" in a line-ending position in *Aspasia asszony* (the latter has the lowest proportion (0.5%) of identical rhymes in the corpus). However those two instances occur in stanzas that are distant from each other and are, thus, not considered to rhyme with one another. We may conclude that Bogáti made conscious and increasing efforts to eradicate this kind of rhyme from his poetry.

[7] It should be kept in mind, however, that this work is a textual fragment. Only roughly half of the original poem has been preserved.



(a) Identical rhymes



(b) Lines ending with “vala”

Figure 2: Identical rhymes and “vala” rhymes

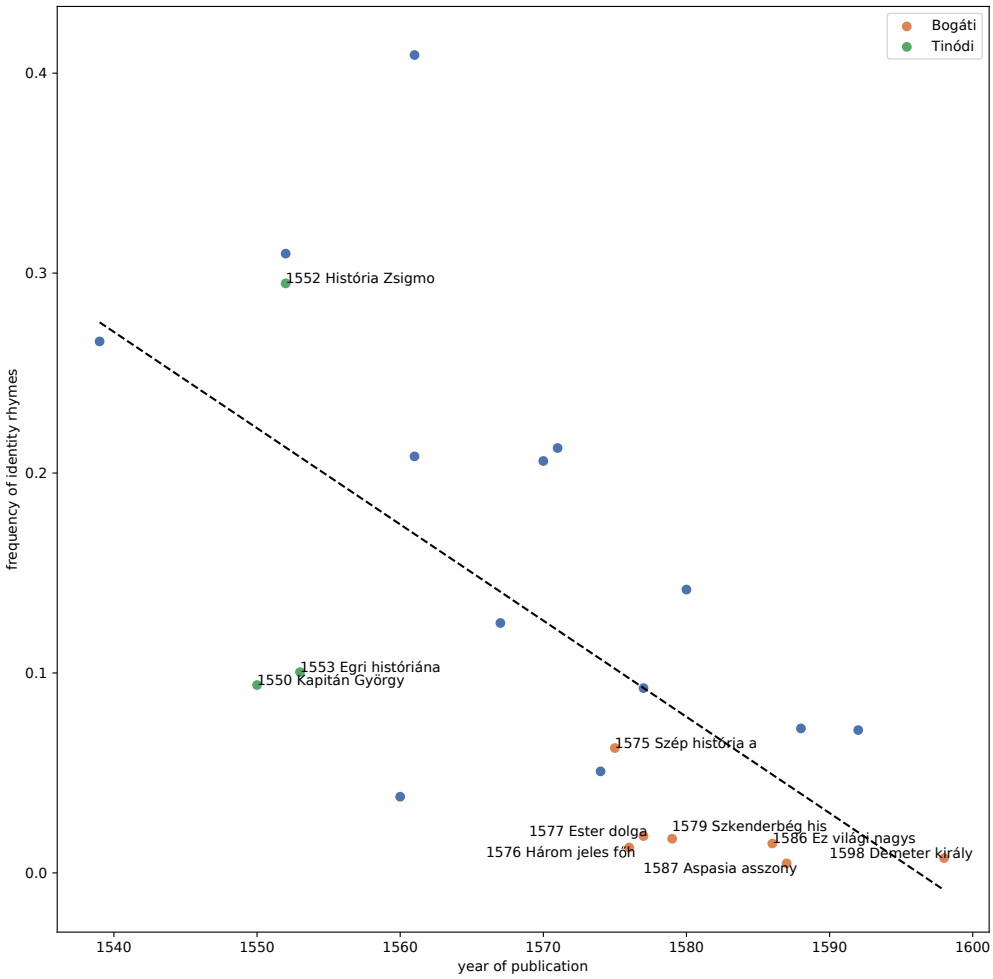


Figure 3: Identical rhymes; linear regression ($r^2 = 0.41$)

5.2 Suffix Rhymes

We understand a suffix rhyme to refer to a pair of rhyming words that each end with the same grammatical suffix.[8] Figure 4 shows the frequency of these rhymes across all the rhymes in the corpus excluding identical rhymes. As the Hungarian language is agglutinative, rhyming with an identical suffix is rather simple. As such, we hypothesized that sophisticated poets would have attempted to avoid this extremely basic method. Three poems had a very high percentage of suffix rhymes: András Valkai's *Bánk bán históriája* (0.65), György Szepesi's *Historia cladis Turcicae* (0.59), and Tinódi's *Kapitány György bajviadala* (0.56). Further assessment of the other characteristics of these poems is

[8] Some authors use the term grammatical rhyme to describe this situation.

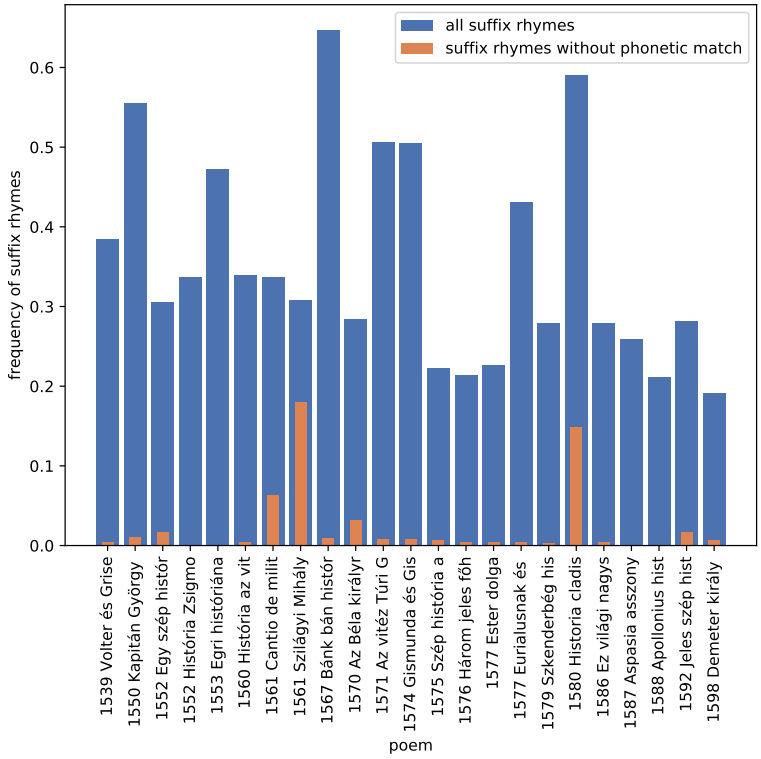


Figure 4: Suffix rhymes

needed, however, to judge whether they showed other weaknesses or managed to compensate for this simplicity.

There is another striking peculiarity of 16th-century Hungarian rhyme that is absent from later poems: some rhyming words have the same grammatical suffix although they differ phonetically. This is true, for example, of -ról/-ről, -ban/-ben, -nak/-nek. The suffixes are phonetically multiform, and the form that corresponds with the phonetic structure of the root is selected. As rhyming by definition requires that (at least) the last vowel in each of two lines be identical, these lines do not rhyme unless we accept that for these authors, the identical grammatical function of the last syllables was an *acceptable and sufficient criterion* for rhyming. Our data show that two poems in the corpus had a surprisingly high number of unmatching suffix rhymes:the anonymously authored *Szilágyi és Hagymási históriája* (0.18) and György Szepesi's *Historia cladis Turcicae* (0.15), which also had the second highest proportion of suffix rhymes of any kind (i.e. including those that matched). We believe that this is an archaic feature of the versification of these poems and relates to their oral character. Significantly these poems differed from the majority of the corpus since they contained no acrostics and had a weaker relationship with writing

and written culture.[9] As mentioned above, *Szilágyi és Hagymási históriája* also has the highest number of identical rhymes, all of which ended with "vala".

5.3 Unrhymed Lines

We also examined the proportion of unrhymed lines in the corpus poems. Unrhymed lines were defined here as lines with a different last vowel than any other line in the same stanza. Again two poems showed significant reliance on the feature: *Cantio de militibus pulchra* (15.5%) and *Szilágyi és Hagymási históriája* (23%). Both these works were written in 1561.[10] *Cantio* has long been connected with oral poetry (Horváth 1984, p. 125), and with a somewhat naive, unschooled poetic tradition. It also contains no acrostics.

Again there was one poem by Bogáti which entirely avoided the feature: (*Három jeles főhadnagyoknak vetélkedése*) had no unrhymed lines, and Bogáti's other poems all contained only a very limited number of these lines.

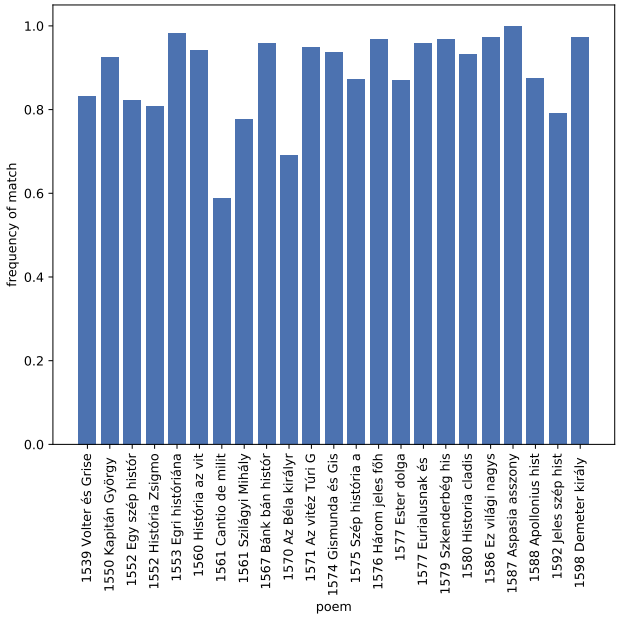
5.4 Consonants and Vowels

Finally, we compared the last two consonant clusters in rhyming lines and the penultimate vowels in rhyming words (Figure 5). We expected that as rhyming evolved, the number of matching vowels and consonants would increase. We reasoned that as rhyming became more and more sophisticated, poets not only avoided using the same rhyming words in the same stanza but also tried to find words that rhymed better phonetically. However, the only trend that we could detect was a decrease in matching penultimate consonant clusters. Tinódi's long poem dedicated to the siege of Eger had by far the most matches (0.33). In contrast, in Bogáti's poetry, the correspondence was very weak (0.04 to 0.1).

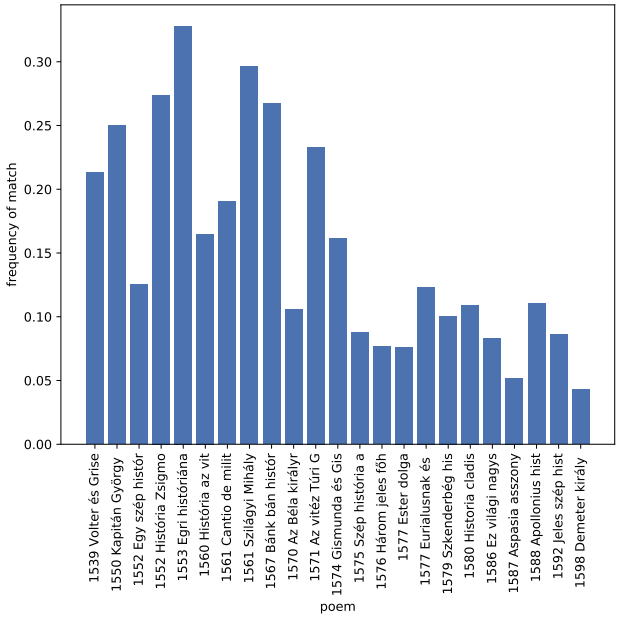
We can only formulate a preliminary hypothesis as to why these rhymes did not become more sophisticated or "richer" over time and why the final and penultimate consonants and the penultimate vowels tended to remain identical. With seven poems in the corpus, Bogáti dominated one end of the time line, and this may be one reason for this curious finding. It may be that because of his various efforts to use complicated acrostics and reduce his reliance on identical rhymes and suffix rhymes, he sometimes needed to settle for rhymes that were weak, with only their last vowel matching. Further investigation and the expansion of the corpus are necessary to assess whether this was a question of personal style or related to more general changes in Hungarian poetry.

[9] In this tradition, acrostics were almost exclusively constructed from the first letters of consecutive stanzas. Fifteen of the 26 poems in our corpus contained acrostics, see listofpoems.csv.

[10] The kinship of these two poems has been observed by previous scholars; see Orlovszky (2009).

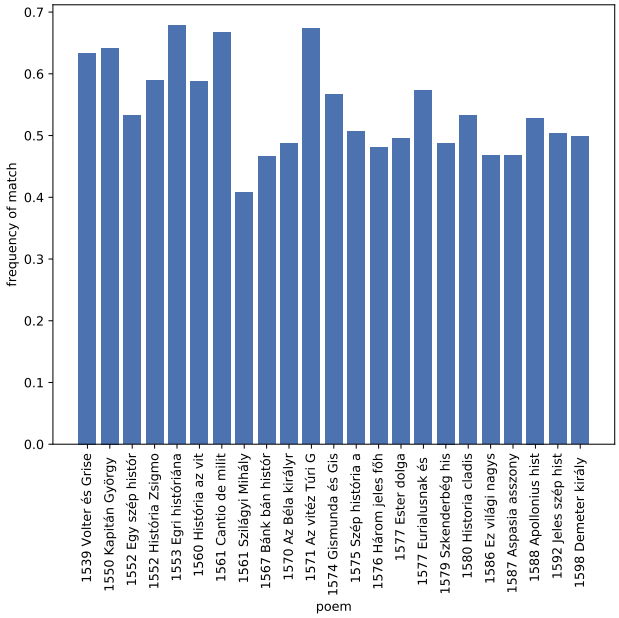


(a) Final consonant cluster



(b) Penultimate consonant cluster

Figure 5: Frequency of matching consonants and vowels. In all cases, identical rhymes and suffix rhymes were excluded from the sample.



(c) Penultimate vowel

Figure 5: Frequency of matching consonants and vowels (cont.)

6 Conclusion

Our analysis revealed a marked change in rhyming techniques in Hungarian historical songs over the examined period. This change took place at different levels and was reflected in an increasing aversion to identical rhymes, unrhyming lines, and phonetically unmatched suffix rhymes. At the same time, based on this corpus alone, we were unable to confirm any change in the use of stanza sequences with the same rhyming syllable by the end of the 16th century. If this shift took place among Szenci Molnár's contemporaries as he professed, then it must have been reflected elsewhere. The expansion of the corpus to include lyric poems may allow us to assess this claim more precisely. It may also reveal distinct and most likely archaic versification patterns in some poems that were less related to literacy than to oral performance and transmission.

As we have noted, the findings of this paper reflect only the first phase of a three-year project. Our research into Hungarian historical songs will proceed on several tracks. We plan to expand the corpus (to 100–150 historical songs) to enable the comparison of subgenres such as songs based on biblical stories and those with more secular content. This expansion will also allow for a better evaluation of personal styles. As a necessity, we will provide a more precise analysis of the grammatical structure of corpus poems. And in the later stages, we will examine the syllabic structure of this verse. Ultimately we hope that this combination of a stylometric analysis and stichometric approach will deliver

a better understanding of the stylistic patterns of the 16th-century historical songs that we have described.

Acknowledgments

The research in this study is supported as a National Research, Development and Innovation Office–NKFIH, OTKA 135631 project. Data and code are available at http://github.com/versotym/oldhun.

References

Ács, Pál, Mihály Etlinger, Balázs Pap, Áron Szatmári, Géza Szentmártoni Szabó, and Edina Zsupán (2018). *Bogáti Fazakas Miklós históriás énekei és bibliai parafrázisai*. Ed. by Pál Ács, Mihály Etlinger, Balázs Pap, Áron Szatmári, Géza Szentmártoni Szabó, and Edina Zsupán. Régi Magyar Költők Tára, XVI. század 13/A. Budapest: Balassi Kiadó. ISBN: 978-963-456-024-1. URL: http://real.mtak.hu/80295/ (visited on 11/27/2020).

Bognár, Péter (2016). *A régi magyar párrímköltészet német vonatkozásai*. Információtörténeti műhely. Budapest: Országos Széchényi Könyvtár.

Horváth, Iván (1984). "Recenzió: Eötvös-füzetek". In: *Irodalomtörténeti Közlemények* 88.1, pp. 124–126. URL: http://epa.oszk.hu/00000/00001/00334/pdf (visited on 11/30/2020).

Horváth, Iván (1991). *A vers: három megközelítés*. 2000 Könyvek. Budapest: Gondolat. ISBN: 978-963-7577-00-0.

Horváth, Iván (2006). *Gépeskönyv*. Budapest: Balassi Kiadó. ISBN: 978-963-506-660-5.

Horváth, Iván (2009). "A magyar vers a reneszánsz és reformáció korában. 1536: Megjelenik két verseskötet". In: *A magyar irodalom történetei*. Ed. by László Jankovits, Géza Orlovszky, and Iván Horváth. Budapest: Országos Széchényi Könyvtár. URL: https://f-book.com/mi/index.php?chapter=0303HORVAMAG.

Horváth, Iván, Zsuzsa Font, Gabriella H. Hubert, János Herner, Etelka Szőnyi, and István Vadai (1979–2020). *Répertoire de la poésie hongroise ancienne, v. 7.0 Beta*. URL: https://f-book.com/rpha/v7/index.php (visited on 11/27/2020).

Indig, Balázs, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai (2019). "One format to rule them all – The emtsv pipeline for Hungarian". In: *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 155–165. DOI: 10.18653/v1/W19-4018. URL: https://www.aclweb.org/anthology/W19-4018.

Jankovics, József, Péter Kőszeghy, and Géza Szentmártoni Szabó, eds. (2000). *Régi magyar irodalmi szöveggyűjtemény II.* Digitális Tankönyvtár. Balassi Kiadó. URL: https://regi.tankonyvtar.hu/hu/tartalom/tkt/regi-magyar-irodalmi-2 (visited on 05/20/2020).

Novák, Attila (2014). “A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 26–31. ISBN: 978-2-9517408-8-4.

Novák, Attila, Borbála Siklósi, and Csaba Oravecz (2016). “A New Integrated Open-source Morphological Analyzer for Hungarian”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. event-place: Portorož, Slovenia. Paris, France: European Language Resources Association (ELRA), pp. 23–28. ISBN: 978-2-9517408-9-1.

Orlovszky, Géza (2009). “A históriás ének: 1574: Megjelenik a Cancionale”. In: *A magyar irodalom története*. Ed. by László Jankovits, Géza Orlovszky, and Iván Horváth. Hálózati kiadás. Budapest: Országos Széchényi Könyvtár. URL: https://irodalom.oszk.hu/villanyspenot/#!/fejezetek/yJ8pg7P3T4yTMgtHhMYEYA.

Ötvös, Péter, László Szilasi, and István Vadai (2004). “Balassi Bálint: Hymni tres ad Sacrosanctam Trinitatem”. In: *Tiszatáj* 58.10, pp. 25–43. URL: http://epa.oszk.hu/00700/00713/00158/pdf (visited on 11/27/2020).

Schelhammer, Zsófia (2019). “Tinódi ismeretlen verseiről”. In: *Verso* 2.3, pp. 5–20.

Seláf, Levente (2017). “Poétiques perpendiculaires: les acrostiches versifiés latins dans la poésie hongroise de la Renaissance”. In: *The Poetics of Multilingualism - La Poétique du plurilinguisme*. Ed. by Patrizia Noel and Levente Seláf. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, pp. 103–120. (Visited on 02/13/2020).

Seláf, Levente (May 2020). “Between Lyric and Epic: The Great Turkish War in German, Italian and Hungarian Ereignisliedern”. In: *Controversial Poetry 1400–1625*. Ed. by Judith Keßler, Ursula Kundert, and Johan Oosterman. Radboud Studies in Humanities 11. BRILL, pp. 61–86. ISBN: 978-90-04-29191-1. DOI: 10.1163/9789004291911_005. URL: https://brill.com/view/book/edcoll/9789004291911/BP000006.xml (visited on 11/29/2020).

Szigeti, Csaba (2005). *Magyar versszak*. Budapest: Balassi Kiadó. ISBN: 978-963-506-623-0.

Vadai, István (1991). “+1: Metrikai határjelölések a régi magyar versben”. In: *Irodalomtörténeti Közlemények* 95.4, pp. 351–369. ISSN: 0021-1486. (Visited on 05/28/2019).

Vadai, István (2009). “A tudósító ének műfaja: 1554: Megjelenik Tinódi Sebestyén Cronicája”. In: *A magyar irodalom története*. Ed. by László Jankovits, Géza Orlovszky, and Iván Horváth. Hálózati kiadás. Budapest: Országos Széchényi Könyvtár. URL: https://f-book.com/mi/index.php?chapter=0405VADAATUD.

Vadai, István (2012). “Szóban kettő - írásban négy: az oralitás metrikájáról”. In: *Doromb: Közköltészeti tanulmányok* 1, pp. 19–41. ISSN: 2063-8175. (Visited on 02/13/2020).

Váradi, Tamás, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, Balázs Indig, Richárd Farkas, and Veronika Vincze (May 2018). “E-magyar – A Digital Language Processing System”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

https://doi.org/10.51305/ICL.CZ.9788076580336.05
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

"Replete with instruction and rational amusement"?: Unexpected Features in the Register of British Didactic Novels, 1778–1814

Juliette Misset

University of Strasbourg, France
jmisset@unistra.fr
0000-0002-0036-8919

Abstract

British didactic novels of the turn of the 19th century have been defined as works of fiction where instruction in moral codes of behavior rather than imaginative elements is the primary focus (Havens 2017, p. 5). My research aims to investigate the generic specificities of such novels by working with the open-source software TXM and AntConc to compare two corpora of novels published between 1778 and 1814 in Britain. These corpora were created using reviews from the *Monthly Review* and the *Critical Review*. Contrary to my hypothesis, a lexical comparison of the two corpora shows that the novels they contain do not materially differ in their use of lexis related to instruction and morality. This leads me to reassess the basis for the early reception of didacticism in these novels. Fruitful new hypotheses are generated using both corpus stylistics and close reading.

1 Introduction

The quote in my title, which comes from an early review of Frances Burney's second novel, *Cecilia* (1782) found in the *Critical Review*, matches recent definitions of the didactic novel as a prevalent subgenre of fiction at the turn of the 19th century. According to Havens (2017), "while didactic novels were allowed imaginative elements, instruction had to remain the primary focus" of the narrative, and this "perpetuated strict moral codes" (pp. 5, 8). Similarly, Wood (2003) argues that the two decades immediately following the French Revolution were "perhaps the most tolerant of overt didacticism in the history of British fiction" and this was in part because of the fear of revolution; as such, the overt didacticism in these novels "coexists with or subsumes aestheticism" (p. 12). While Wood focuses her study on anti-revolutionary female novelists, she states that writers across the political spectrum wrote didactic fiction, which

was "constructed to avoid ambiguity, and to discourage personal and idiosyncratic exegesis" (pp. 64–65). These novels rely on plot "to inculcate particular morals" (p. 68), and the language they use, for example, in their embedded statements and value judgments "implicitly support[s] the text's moral basis" and "indicate[s] the appropriate readerly response" (p. 66). These critics define turn-of-the-19th-century British didactic novels as works with a straightforward moral message that was delivered through the medium of narrative fiction; here instruction came first and amusement second. This is opposed to fiction that foregrounds the kinds of "narratorial 'indirection'" that literary critics have increasingly come to value since the 19th century (Wood 2003, p. 16).

Both Havens and Wood consider didactic novels to have been of cultural—although not necessarily artistic—importance in Britain in the decades surrounding the French Revolution. Nevertheless, while adjacent categories of fiction of the period such as sentimental or domestic novels have been studied at length, the formal specificities of the didactic novel as a subgenre in this narrative landscape are yet to receive much critical attention (Havens 2017, p. 13). This paper aims to investigate elements of what may be called the constitutive register of didactic novels based on Biber's definitions of this notion (2019, p. 16). I use a combination of computer-aided textual analysis and close reading to compare two corpora of novels published between 1778 and 1814 in Britain. As we will see, the former approach yielded very different results from the ones I expected, which forced an abrupt shift in my perspective on this research. I complement the original corpus-based methodology with a corpus-driven one, which leads to stimulating new approaches to the textual basis for these novels' reception.

2 The Corpora

Contemporary reviews of novels of this period found in the *Monthly Review* and the *Critical Review* were used to create the two corpora.[1] This allowed for a systemic approach to the creation of the corpora since both these *Reviews* professed to address all new publications in their book reviews rather than "select[ing] drastically" as the *Edinburgh Review* did at the beginning of the 19th century (Butler 1993, p. 131; Christie 2018, p. 282). In addition, starting with these early reviews made it possible to study the evolution of the reception of didacticism over time and to compare this with the findings of my textual analysis. In order to qualify for inclusion in the didactic corpus, novels had to

1. have been praised by at least one reviewer for their ability to instruct as well as amuse or entertain readers,[2]

2. be set primarily in Britain in the period of their conception, and

[1] Partial reviews of all of the novels published between 1770 and 1799 are available in Raven et al. 2000, while all full reviews for the period 1800–1829 can be found in P. Garside et al. 2004.

[2] Both instruction and amusement were necessary for inclusion in keeping with the *utile et dulce* formula central to 18th-century conceptions of the value of fiction (Wood 2003, p. 15; Millet 2007, p. 43).

1778	*Evelina*	Frances Burney
1778	*Munster Village*	Lady Mary Hamilton
1782	*Cecilia*	Frances Burney
1788	*Mary, A Fiction*	Mary Wollstonecraft
1790	*Julia, A Novel*	Hannah Maria Williams
1796	*Hermsprong, or Man as He Is Not*	Robert Bage
1796	*Memoirs of Emma Courtney*	Mary Hays
1798	*Edgar: or, The Phantom of the Castle*	Richard Sicklemore
1798	*Maria; or, The Wrongs of Woman*	Mary Wollstonecraft
1801	*Belinda*	Maria Edgeworth
1801	*The Father and Daughter*	Amelia Opie
1805	*The Nobility of the Heart*	Elizabeth Spence
1808	*Cœlebs in Search of a Wife*	Hannah More
1810	*Romance Readers and Romance Writers*	Sarah Green
1811	*Sense and Sensibility*	Jane Austen
1811	*Self-Control*	Mary Brunton
1813	*Pride and Prejudice*	Jane Austen
1814	*Patronage*	Maria Edgeworth

Table 1: The didactic corpus

3. be available in electronic format.

The earliest work in the didactic corpus is Frances Burney's first published novel, *Evelina* (1778), which was chosen as a chronological starting point based on Burney's importance as a novelist in the last two decades of the 18th century (Havens 2017, p. 8).[3] Eighteen novels fit all the criteria for inclusion; of these, nine were published between 1778 and 1799 and nine between 1800 and 1814 (see Table 1).

A reference corpus was built to provide a representative sample of fiction of the same period to which the didactic corpus could be compared. Here I relied on Mahlberg's (2013) claim that "corpus work is essentially comparative: a text or text extract is compared to an appropriate reference corpus providing a relative norm" (2013, p. 24). The reference corpus is comprised of 18 novels that were noted for their instructive effect in the *Monthly Review* or the *Critical Review* but otherwise have the same characteristics of being set primarily in contemporary Britain and now available in digital form. The reference corpus also features nine novels published between 1788 and 1799 and another nine that appeared between 1800 and 1814 (see Table 2).

Based on Havens' and Wood's claims that didactic novels of the turn of the 19th century predominantly rely on straightforward language to perpetuate moral norms and codes of behavior, the two corpora were compared using computer-aided textual analysis. In particular, this analysis sought to confirm

[3] Interestingly, while Havens excludes *Evelina* from her discussion because of its "pervasive satire," a critic from the *Critical Review* pronounced the book full of "lessons" leading "to improvement and to virtue" thanks in part to its "useful humour and diverting satire" (*CR* 1778, vol. 46: 203).

1778	*Learning at a Loss; or, The Amours of Mr. Pedant and Miss Hartley*	Gregory Lewis Way
1788	*Emmeline; or, The Orphan of the Castle*	Charlotte Smith
1791	*A Simple Story*	Elizabeth Inchbald
1792	*Anna St. Ives*	Thomas Holcroft
1794	*Caleb Williams*	William Godwin
1795	*Henry*	Richard Cumberland
1796	*Nature and Art*	Elizabeth Inchbald
1798	*Rosamund Gray*	Charles Lamb
1799	*The Vagabond*	George Walker
1804	*Adeline Mowbray*	Amelia Opie
1805	*Fleetwood; or, The New Man of Feeling*	Caleb Williams
1806	*Leonora*	Maria Edgeworth
1806	*The Wild Irish Girl*	Sydney Owenson
1812	*The Son of a Genius*	Barbara Hofland
1813	*The Heroine*	Eaton Stannard Barrett
1814	*Mansfield Park*	Jane Austen
1814	*Discipline*	Mary Brunton
1814	*The Wanderer*	Frances Burney

Table 2: The reference corpus

whether the topic of morality and instruction is a salient marker of the register of the didactic corpus and can be established as the primary reason why these novels were received as didactic upon first publication. In fact, my results completely negated this hypothesis, which led at first to considerable frustration. Eventually, however, it prompted an extremely fruitful reappraisal of the complex links between reception and textual elements, as evidenced by digital tools.

3 Testing the Original Hypothesis

From the outset of my research, I endeavored to trace the themes of morality and instruction in the didactic corpus as an element of textual register (Biber et al. 2019, p. 40). Given Biber's claim that "the words used in a text are to a large extent determined by the topic of the text" (p. 40), keyword analysis was chosen as a means to determine whether these themes were particularly salient in the didactic corpus in comparison to the reference corpus. These keywords were, thus, posited as a potential aspect of the self-evident register of what has been termed "overt didacticism" in novels of the period (Butler 1972, p. 449; Wood 2003, p. 12).[4] My hypothesis was that novels received as didactic would be likely to engage with questions of morality and instruction in

[4] Word frequency is often used to attribute authorship in corpus linguistics and stylistics (Jockers 2013, p. 70; Burrows 2018, p. 724; Szudarski 2018, p. 25), and it has also been used to classify novels in terms of genre (Allison et al. 2011, p. 5).

a direct manner perceivable in terms of vocabulary frequency. This stemmed from the overall critical consensus regarding the porosity between conduct books and novels, and particularly between conduct books aimed at women and novels written by women (Bilger 1998, p. 21; Spencer 1986, p. 142). This seemed particularly pertinent given that 16 of the 18 novels in my didactic corpus were written by women. Based on concepts from 18th-century moral philosophy and definitions from the *Oxford English Dictionary*,[5] I consequently devised a list of words related to morality and instruction to be quantified using the TXM software. The latter allows for the creation of complex Corpus Query Language (CQL) queries, making it possible to combine several terms into a concept or theme to be studied.

Given the influence of Adam Smith's system of moral philosophy on 18th-century thought (Howell 1971, p. 447), I included Smith's "cardinal virtues," namely prudence, benevolence, justice, self-command, and sympathy, in my list of moral terms (Haakonssen 2002, pp. viii, xiii, xx). I added the term "modesty" since it is particularly associated with women in John Locke's *Some Thoughts Concerning Education* (1902, p. 164).[6] "Honor" and "courage" were also added because critics mentioned them in the novels' reviews. Similarly I included "sensibility" and "delicacy" as Hugh Blair considered them the grounds for a "superior moral life" (quoted in Van Sant 1993, p. 5). "Reason" and "passion" made it onto the list as the main contentious forces at play in the exercise of virtue according to Mary Wollstonecraft (2004, pp. 30–31). "Propriety" was included based on Jane Spencer's claims about its growing importance throughout the 18th century and its links to morality and modesty, particularly for women (1986, p. 75). I added "duty" and "conduct" to complement ideas about the behavioral norms that helped define "morality" along with "propriety." The list of terms related to instruction was built using definitions and synonyms from the *Oxford English Dictionary* online.

In order to compare the values yielded by TXM, I used a per-mill approach as well as log-likelihood (LL), a "test [which] helps you determine whether differences in the frequency of words are reflective of the actual variation in language or whether they result from chance occurrences" (Szudarski 2018, p. 27). According to Rayson, Berridge, et al. (2004), "one million words gives sufficient evidence for mid- to high-frequency words" in corpus linguistics studies (p. 1). Since my corpora were respectively 2, 532, 943 and 2,683, 379 words long based on a count by TXM, the log-likelihood test was well-placed to deliver valid results.

The threshold for statistical significance commonly used for statistical measures such as log-likelihood is 5%, which amounts to a critical value of 3.84.[7] In recent years, however, critics have questioned the pertinence of the log-likelihood test if used on its own. It has been argued, for instance, that the test claims to detect too many significant differences when comparing two corpora

[5] *OED Online*. Oxford University Press, June 2021 [accessed 1 July 2021.]

[6] John Locke is directly quoted in several novels from both corpora: Lady Mary Hamilton's *Munster Village* (1778), Hannah More's *Cœlebs in Search of a Wife* (1808), and Maria Edgeworth's *Patronage* (1814) in the didactic corpus, and George Walker's *The Vagabond* (1799), Amelia Opie's *Adeline Mowbray* (1804), and Sydney Owenson's *The Wild Irish Girl* (1806) in the reference corpus.

[7] See, for example, the online calculator https://www.korpus.cz/calc/.

(Bestgen 2017, p. 37). Moreover, log-likelihood is a measure of statistical significance, and "does not by itself inform us of whether the difference between the frequencies [...] carries any descriptive value" (Fidler et al. 2015, p. 227). Effect size statistics may be used to complement research as they "focus [...] on how large the difference between the two frequencies of a word is" (Pojanapunya et al. 2018, p. 145). One example of an effect size metric is the Log Ratio (LR); this is included in Rayson's online calculator, the tool used to process the data set presented in Tables 3 and 4.[8]

Each word category from Tables 3 and 4 includes all the grammatical forms of the lemma that pertain to the central notion. For example, the category "instruction" includes the nominal and verbal lemmata "instruction" and "instruct". Where a grammatical category changed the fundamental meaning of a word, that category was not included in the table, and concordance lines were used to select relevant occurrences of polysemous words based on the context in which they were used. An example is the verb "to conduct," which among other things may mean "to behave" or "to lead."

These tables suggest that the two corpora do not differ materially when it comes to the explicit presence of the topics of morality and instruction, the central elements of the concept of didacticism in fiction at the time (Havens 2017, p. 5).

The results presented in Tables 3 and 4 invalidate the hypothesis tested by this corpus-based approach. Although the differences in the frequency of several terms in Table 3 are shown to be statistically significant, the log-likelihood values remain rather low. It therefore becomes difficult to make any reliable claim about the greater engagement with the topic of morality and instruction of the didactic corpus when compared with the reference corpus. The LR measure corroborates this: if a word has the same relative frequency across the corpora, its LR value is 0; if it is twice as common in the analyzed corpus, its LR value is 1, and every additional point represents a doubling of the ratio (Collins et al. 2020). In Table 3, "prudence" stands out as the term with the highest LL and LR values, but the overall picture suggests that both corpora include the topic of morality to similar degrees. This is also true for Table 4, which shows little difference in the presence of the topic of instruction in the two corpora, as seen in the list of terms. Furthermore, two of the three lemmata with the highest LL and LR values, "edify" and "tutor" are actually over-represented in the reference corpus.

To complement these results, I completed a qualitative review of the endings of all of the novels. My aim was to determine whether language about the moral conclusions to be drawn from the narrative was more prevalent and/or less ambiguous in the didactic novels than in the reference novels. In both corpora, 15 of the 18 novels mention a vice punished and a virtue rewarded in their closing paragraphs; this mirrors the quantitative results in Tables 3 and 4. Even more strikingly, the endings of five of the 18 novels in the didactic corpus are morally ambiguous in some respect. This may be seen, for example, in the last words of Frances Burney's *Cecilia* (1782) about the spendthrift Mrs. Harrel, who does not learn from her first husband's financial ruin and subsequent

[8] Rayson's log-likelihood calculator can be found at http://ucrel.lancs.ac.uk/llwizard.html.

	Didactic corpus 2,532,943 tokens	‰	Reference corpus 2,683,379 tokens	‰	Log-likelihood	Log Ratio
benevolence	270	0.11	307	0.11	-0.72	-0.10
conduct	669	0.26	522	0.19	27.66	0.44
delicacy	479	0.19	406	0.15	10.98	0.32
duty	535	0.21	584	0.22	-0.25	-0.04
hono(u)r	1075	0.42	998	0.37	9.03	0.19
justice	318	0.13	457	0.17	-17.69	-0.44
modest	182	0.07	211	0.08	-0.8	-0.13
moral	264	0.10	291	0.11	-0.22	-0.06
passion	806	0.32	922	0.34	-2.54	-0.11
proper	565	0.22	580	0.22	-0.28	0.05
prudence	365	0.14	226	0.08	41.51	0.77
reason	1314	0.52	1225	0.46	10.37	0.18
self-command	32	0.01	36	0.01	-0.06	-0.09
sensibility	254	0.10	280	0.10	-0.21	-0.06
sympathy	170	0.07	153	0.06	2.14	0.24
vice	203	0.08	212	0.08	-0.02	0.02
virtue	728	0.29	816	0.30	-1.23	-0.08
TOTAL	8229	3.25	8226	3.07	13.86	0.08

Table 3: Lemmata related to morals in both corpora

	Didactic corpus 2,532,943 tokens	‰	Reference corpus 2,683,379 tokens	‰	Log-likelihood	Log Ratio
advise	183	0.072	206	0.077	0.36	-0.09
edify	11	0.004	44	0.016	-19.34	-1.92
educate	338	0.133	262	0.098	14.54	0.45
enlighten	45	0.018	62	0.023	-1.82	-0.38
explain	476	0.188	398	0.148	12.2	0.34
improve	310	0.122	270	0.101	5.55	0.28
inculcate	25	0.099	12	0.004	5.45	1.14
influence	304	0.120	353	0.132	-1.38	-0.13
inform	804	0.317	735	0.274	8.36	0.21
instruct	199	0.079	254	0.095	-3.9	-0.27
learn	498	0.197	625	0.233	-8	-0.24
lesson	94	0.037	139	0.052	-6.34	-0.48
study	360	0.142	367	0.137	0.27	0.06
teach	349	0.138	407	0.152	-1.74	-0.14
tutor	19	0.008	49	0.018	-12.03	-1.28
urge	180	0.071	263	0.098	-11.22	-0.46
TOTAL	4195	1.656	4446	1.657	0.00	0.00

Table 4: Lemmata related to instruction in both corpora

suicide, and far from being punished, "married very soon a man of fortune in the neighbourhood, and quickly forgetting all the past, thoughtlessly began the world again, with new hopes, new connections,–new equipages and new engagements!" (Book 10, chapter 10). In contrast, only one novel from the reference corpus includes a morally ambiguous ending: the cautionary tale promised at the beginning of William Godwin's *Caleb Williams* (1794) turns out to simply be a profession of frankness by the autodiegetic narrator, without any consequences for his perceived vices.

Yet while the tables and analysis of endings show that overall and unexpectedly the novels of both corpora engage with the topics of morality and instruction to similar degrees, a divergence based on gender emerges for three terms from Table 3. In these cases, the LL and LR values respectively show a statistically significant difference in frequency and comparatively greater differences in frequencies within the data. The terms "conduct" and "prudence," which are both over-represented in the didactic corpus, evoke values prevailingly attached to femininity in the period.[9] In contrast, "justice" is a term linked to the traditionally and historically male-dominated world of legal power as the basis of jurisprudence in Adam Smith's moral philosophy (Haakonssen 2002, p. xx). The didactic corpus features works predominantly written by women (16 of the 18 novels) whereas the reference corpus is comprised of 10 works by female authors and eight by male authors, a rough reflection of the gendered distribution of authorship in the period (Mandal 2007, pp. 13, 27). The ratio of female to male protagonists in the novels in each corpus mirrors these proportions, with 15 novels from the didactic corpus and 10 from the reference corpus featuring a female main character. At the same time, some of the male-authored novels in both corpora have a female protagonist, and vice versa. It is therefore not surprising to find that gender plays a role in the differences in vocabulary use within the corpora, especially given that didacticism has often been linked to female authorship (Towsey 2015, p. 33; Havens 2017, p. 13).

4 Looking for Overt Didacticism…

Before pursuing the angle of gender, however, I was prompted by my initial results to look for more lexical and syntactical markers of what has been seen as overt didacticism (Butler 1972, p. 449; Wood 2003, p. 12). Here I moved beyond tracing moral didacticism as a topic to investigating the linguistic features of its register. Wood (2003) has opposed overt didacticism to the kinds of indirection found and valued, for instance, in the works of Jane Austen, and she notes the importance of an authoritative narrative voice in making overt didacticism effective (p. 66). Susan Lanser (1992) conceives of an overt authoriality involving narrative voices that engage in extra-representational acts such as "reflections, judgments, generalizations about the world 'beyond' the fiction, direct addresses to the narratee, comments on the narrative process, allusions to other writers and texts" (pp. 16–17). She also contrasts this explicit authoriality with forms of indirection such as "free indirect discourse, irony, ellipsis, nega-

[9] See, for example, Butler 1987, p. 122; Mellor 1993, p. 40; Spencer 1986, p. 75.

tion, euphemism, [and] ambiguity," which are characteristic of Austen's novels (p. 62). Lesser-known female novelists of the 18th century, she suggests, tended to engage in this overt authoriality even though women novelists consistently received more praise when their authoriality was most covert (pp. 66, 78).

Given that my didactic corpus consists mostly of works which have not entered the literary canon, I was led to investigate whether overt authoriality is a marker of overt didacticism as the register of these novels. These findings were then compared with those for the reference corpus. Again, this did not yield the expected results but instead showed that an authoritative tone is not a prevalent feature of moral didacticism as this was received in novels at the turn of the 19th century. My investigation, for example, of direct addresses to readers (DAR) in the prefaces and main texts of the novels in both corpora found that this device is by no means specific to the novels of the didactic corpus. Moreover, rather than being a strategy to ensure readers' ideological assent, in both the corpora, DAR marks an attempt to negotiate the places of the author, the reader, and the critic in relation to one another at a time when the novel was in the process of becoming a legitimate genre (Misset [forthcoming]).

My study of DAR also showed that the quintessential didactic novel of the period, Hannah More's *Cœlebs in Search of a Wife* (1808) is far from representative of didactic fiction in general.[10] While the addresses to readers in the other novels of both corpora are voiced by the narrator to extradiegetic readers, 10 of the 11 DAR in *Cœlebs* take place through intradiegetic dialogue, as shown in Table 5.

All the characters featured in Table 5 are firmly established as sound moral authorities in the novel, and eight of the 11 occurrences directly address the question of moral improvement through various kinds of reading material. This was a unique finding in the context of the DAR in the other novels in both corpora. Instead of confirming moral didacticism as a unifying trait of DAR, I therefore found their use as vehicles for overt didacticism to be an exception.

5 …and Finding Gender and Class Bias

The results of the corpus-based approach proved crucial for reappraising ideas about overt didacticism as a fictional register. Nevertheless more information was needed in order to move the focus from what didacticism is *not* according to these corpora and, thus, to try to determine what it is. This led me to adjust my method from one that was corpus-based to one that was corpus-driven and so would generate rather than verify hypotheses (Cornby et al. 2016, p. 7). I consequently switched analytic tools from `TXM` to `AntConc`, a concordance tool which, as its creator writes, generates keyword lists that show "which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus, for example, as part of a genre […] study" (Anthony 2019b, p. 7).

[10] *Cœlebs* is known for being an early 19th-century bestseller (Stott 2003, pp. 277, 281) and often cited as an example of didactic fiction of the period (Kelly 2018, p. 198; Wood 2003, p. 66; Mandal 2007, p. 95).

Cœlebs	however, will inevitably dazzle the feeling	reader	, till it produce the common effect of
Mr S	simplicity, as far removed from the careless	reader	of a common story, as from the declamation
	Lady Belfield, who, though not new to the	reader	or the writer, were new at Stanley Grove.[a]
Mrs S	the omnipotence of love, that the young	reader	was almost systematically taught an
Cœlebs	it may and does tire the patience of the	reader	, yet it never leaves him ignorant; and of
Mr S	human nature is corrupt; that the young	reader	is helpless, and wants assistance; that he is
Sir J	fancy, nor to extinguish a taste for them in	readers	. “ ” Show me any one instance of good that
Cœlebs	elevation of fancy led Milton, or Milton his	readers	? Into what immoralities did it involve
Cœlebs	of the living minstrel of the LAY? What	reader	has Mason corrupted, or what reader has
Cœlebs	What reader has Mason corrupted, or what	reader	has Cowper not benefitted? Milton was an
Sir J	communicated, a hundred thousand	readers	caught, the infection. Sentimentality was

Table 5: DAR in *Cœlebs in Search of a Wife*[b]

[a] This is the only occurrence which is not reported speech.
[b] This table was created using concordance lines generated by `TXM`

AntConc's keyword list ranks words according to their "keyness," which is measured using log-likelihood—the higher a word's keyness score, the more characteristic it is of the corpus as compared to the reference corpus. At the same time, the tool takes into account the difference in the size of the corpora it compares.[11] The keyword lists may also be made to incorporate Log Ratio as a measure of effect size.

In order to compare the novels in the didactic and the reference corpora using AntConc, the texts were lemmatized manually using TreeTagger. This merged all the different inflections of the same lemma into one term. The results of this method corroborate those of the corpus-based approach insofar as the terms found to be statistically overused in one corpus or the other using the LL calculation website and the TXM frequency values appear as keywords. "Prudence," for instance, (rank 481, LL 26.09, LR 0.7905) appears in the keyword list when the didactic corpus is set as the primary corpus, while "justice" (rank 265, LL 28.95, LR 0.5977) occurs in the list when the reference corpus is set as the primary corpus.[12]

The most striking feature of the keyword lists generated through AntConc's comparison of didactic and reference corpora is the gendered divide. This is clear from the frequency of female pronouns and nouns, and it corroborates the findings of the corpus-based study using TXM. When the didactic corpus is compared to the reference corpus in AntConc, the highest-ranking keywords that are not characters' names are "lady" (rank 12, LL 968.62, LR 0.8949), "she" (rank 25, LL 708.62, LR 0.3463), "her" (rank 32, LL 555.65, LR 0.2492), "ladyship" (rank 64, LL 332.43, LR 1.6632), and "daughter" (rank 163, LL 106.28, LR 0.6425). This suggests that the novels in the didactic corpus focus on female characters more than the novels in the reference corpus do. "Lady" and "ladyship" had the highest LR values among these didactic novels, which points to the setting of the novels of the didactic corpus in genteel society more often than the novels of the reference corpus.

Turning to the distribution of "lady" across the different novels in the didactic corpus, we find that the five works with the lowest concentration of the term are Richard Sicklemore's *Edgar, or The Phantom of the Castle* (1798, 6 hits; this is also the only novel in the corpus to focus almost exclusively on male characters); Mary Wollstonecraft's *Maria, or The Wrongs of Woman* (1798, 13 hits) and *Mary,*

[11] At the outset, I was directed to TXM as a tool for my research based on its ability to accommodate complex CQL queries. This appeared particularly useful at the start of my project, which originally aimed to trace the topics of morality and instruction. Although the two tools are not fundamentally different, AntConc was suggested for the subsequent exploratory phase of my research given the ease with which it can compare keyword use between two corpora.

[12] The LL values calculated based on the TXM data and the ones produced by AntConc do not quite coincide. However, this is a common phenomenon when using different tools to study the same corpora (Anthony 2013, p. 149). TXM and AntConc are complex software tools that are programmed differently, and they may make calculations in slightly different ways. This is a drawback of using ready-made software where the user cannot easily access all the settings (Gries 2009, p. 2). More specifically, AntConc can only count one lemma at a time, but I could combine different lemmata derived from the same notion (eg. "moral" and "morality"), as applicable, when working with the TXM data. I could also clean the data when faced with polysemous terms such as "conduct, n." and "conduct, v." Notwithstanding these issues, the different LL values calculated using the two sets of tools do not materially impact the overall results, and the findings correlate.

A Fiction (1788, 13 hits); Mary Hay's *Memoirs of Emma Courtney* (1796, 37 hits); and Amelia Opie's *The Father and Daughter* (1801, 38 hits). The four latter works all feature female protagonists who are gentry women in stories that do not, however, revolve around genteel social life. At the other end of the spectrum, the frequency of "lady" is highest in Maria Edgeworth's *Belinda* (1801) where it records 1554 hits, and most of the 12 remaining novels contain between 200 and 600 hits.

In the reference corpus, "lady" has the highest concentration in novels set among genteel society. These works overwhelmingly feature a female protagonist with only one, Richard Cumberland's *Henry* (1795, 476 hits) concentrating on a male protagonist. On the other hand, the seven novels with the lowest concentration of the term all focus on male protagonists, and in four cases, these characters come from lower walks of life than their counterparts in other novels in the corpora: they are poorer and have to earn their daily bread, whether as subordinates living in a wealthy household (William Godwin's *Caleb Williams*, 1794, 37 hits) or as traders (Elizabeth Inchald's *Nature and Art*, 1796, 61 hits; Barbara Hofland's *The Son of a Genius*, 1812, 20 hits; and George Walker's *The Vagabond*, 1799, 10 hits). The reference corpus, thus, includes novels with a greater variety of social settings and gendered perspectives than the didactic corpus. Female genteel experience may therefore be seen as a defining feature of the didactic corpus and one that is quantifiable through lexical frequency.

Finally, a comparison of the novels of the five authors who appear in both corpora (Jane Austen, Mary Brunton, Frances Burney, Maria Edgeworth, and Amelia Opie) confirms the tendency of the novels received as didactic by *Monthly Review* and *Critical Review* critics to be set in genteel society. At the same time, it highlights that these narratives had to follow a certain pattern. All of the novels by these authors in the corpora center on a genteel female protagonist who is navigating moral questions, but the novelists' word use distinguishes the works in the didactic corpus from the ones in the reference corpus. Terms such as "say," "conversation," "company," "behaviour," "please," and "manner" are over-represented in the novels of the didactic corpus, which illustrates their central focus on the social behavior of genteel life. In contrast, the over-representation in the reference corpus of "my," "self," and "feeling," all of which are fairly evenly distributed among these novels, suggests a greater focus on the personal experiences of their female protagonists.

6 Conclusion

My research on these two corpora is still in progress. Nevertheless this use of quantitative data obtained through `TXM` and `AntConc` has shown the value of interrogating the reception of texts in light of their linguistic features: this method can help determine the part ideology has played in previous readings and categorizations. In the current case, the novels that critics initially received as instructive for readers are the ones that largely support and reinforce the specific norms of behavior and social hierarchies that would become central to the Victorian ethos. This is most notable in the foregrounding of genteel do-

mesticity as a feminine ideal, evocative of the "Angel in the House" (Bilger 1998, p. 85). Crucially, we find that the elements of register specific to didactic novels as a fictional subgenre hinge not on engagement with the topics of morality and instruction or on any tonal authority but rather on the presence of specific kinds of characters—genteel women—whose narrative and moral trajectories center their social interactions rather than their personal experiences and development. These findings have completely redirected my research and form the foundation for my ongoing qualitative work on the narrative trajectories that distinguish the novels of the didactic corpus from those of the reference corpus. As such, they allow for new claims to be made and substantiated about the register and reception of these novels.

References

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore (2011). "Quantitative Formalism: An Experiment". In: Stanford Literary Lab. Chap. Pamphlet 1. URL: https://litlab.stanford.edu/pamphlets/.

Anthony, Laurence (2013). "A Critical Look at Software Tools in Corpus Linguistics". In: *Linguistic Research* 30.2, pp. 141–161.

Anthony, Laurence (2019a). *AntConc*. Version 3.5.8. Tokyo. URL: https://www.laurenceanthony.net/.

Anthony, Laurence (2019b). *AntConc Help Manual*. Tokyo: Center for English Language Education in Science, Engineering, School of Science, and Engineering, Waseda University.

Bestgen, Yves (2017). "Getting Rid of the Chi-Square and Log-Likelihood Test for Analysing Vocabulary Differences between Corpora". In: *Quaderns de Filologia: Estudis Lingüístics* 22, pp. 33–56. DOI: 10.7203/qf.22.11299.

Biber, Douglas and Susan Conrad (2019). *Register, Genre, and Style. Second Edition*. Cambridge: Cambridge UP.

Bilger, Audrey (1998). *Laughing Feminism: Subversive Comedy in Frances Burney, Maria Edgeworth, and Jane Austen*. Detroit, Michigan: Wayne State UP.

Burrows, John (2018). "Rho-grams and rho-sets: Significant links in the web of words". In: *Digital Scholarship in the Humanities* 33.4, pp. 724–747. DOI: 10.1093/llc/fqy004.

Butler, Marilyn (1972). *Maria Edgeworth: A Literary Biography*. Oxford: Clarendon Press.

Butler, Marilyn (1987). *Jane Austen and the War of Ideas*. Oxford: Clarendon Press.

Butler, Marilyn (1993). "Culture's Medium: The Role of the Review". In: *The Cambridge Companion to British Romanticism*. Ed. by Stuart Curran. Cambridge: Cambridge UP.

Christie, William (2018). "Critical Judgment and the Reviewing Profession". In: *The Oxford Handbook of British Romanticism*. Ed. by David Duff. Oxford: Oxford UP.

Collins, Luke C., Elena Semino, Zsófia Demjén, Andrew Hardie, Peter Moseley, Angela Woods, and Ben Alderson-Day (2020). "A linguistic approach to the

psychosis continuum: (dis)similarities and (dis)continuities in how clinical and non-clinical voice-hearers talk about their voices". In: *Cognitive Neuropsychiatry* 25.6, pp. 447–465. DOI: 10.1080/13546805.2020.1842727.
Cornby, Émeline, Yannick Mosset, and Stéphanie de Carrara (2016). *Corpus de textes : composer, mesurer, interpréter*. Lyon: ENS Éditions.
Fidler, Masako and Cvrček Václav (2015). "A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keywork Analysis". In: *Journal of Slavic Linguistics* 23.2, pp. 197–239.
Garside, Peter, Jaqueline Belanger, and Sharon Ragaz (2004). *British Fiction, 1800-1829: A Database of Production, Circulation & Reception*. Cardiff University. URL: http://www.british-fiction.cf.ac.uk/.
Gries, Stefan (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. London, New York: Routledge.
Haakonssen, Knud (2002). "Introduction". In: *Quantitative Corpus Linguistics with R: A Practical Introduction*. Ed. by Adam Smith. Cambridge: Cambridge UP.
Hallowell, Thomas Jewett, Tobias Smollett, and Laurence Hutton (1756–1817). *The Critical Review, or, Annals of Literature*. London: W. Simpkin and R. Marshall. URL: https://catalog.hathitrust.org/Record/008890497/Home.
Havens, Hilary (2017). "Introduction". In: *Didactic Novels and British Women's Writing, 1790–1820*. Ed. by Hilary Havens. New York, London: Routledge.
Heiden, Serge (2010). "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme". In: *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp. 389–398. URL: https://aclanthology.org/Y10-1044.
Howell, Wilbur Samuel (1971). *Eighteenth-Century British Logic and Rhetoric*. Princeton: Princeton UP.
Jockers, Matthew (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, and Springfield: University of Illinois Press.
Kelly, Gary (2018). "The Spectrum of Fiction". In: *The Oxford Handbook of British Romanticism*. Ed. by David Duff. Oxford: Oxford UP.
Lanser, Susan (1992). *Fictions of Authority: Women Writers and Narrative Voice*. Ithaca, NY: Cornell University Press.
Locke, John (1902). *Some Thoughts Concerning Education*. Cambridge: Cambridge UP.
Mahlberg, Michaela (2013). *Corpus Stylistics and Dickens's Fiction*. New York, London: Routledge.
Mandal, Anthony (2007). *Jane Austen and the Popular Novel: The Determined Author*. Basingstoke: Palgrave Macmillan.
Mellor, Anne (1993). *Romanticism & Gender*. New York: Routledge.
Millet, Baudoin (2007). *'Ceci n'est pas un roman': l'évolution du statut de la fiction en Angleterre de 1652 à 1754*. Louvain, Paris, Dudley, MA: Editions Peeters.
Misset, Juliette ([forthcoming]). *'I hope I shall please my readers': Negotiating the Author-Reader Relationship in Two Corpora of British Novels, 1778-1814*.

Pojanapunya, Punjaporn and Richard Watson Todd (2018). “Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis”. In: *Corpus Linguistics and Linguistic Theory* 14.1, pp. 133–167. DOI: doi:10.1515/cllt-2015-0030.

Raven, James and Antonia Forster (2000). *The English Novel, 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles. Volume I, 1770–1799*. Oxford: Oxford UP.

Rayson, Paul, Damon Berridge, and Brian Francis (2004). “Extending the Cochran rule for the comparison of word frequencies between corpora”. In: *JADT*, pp. 1–12.

Rayson, Paul and Roger Garside (2000). “Comparing Corpora Using Frequency Profiling”. In: *Proceedings of the Workshop on Comparing Corpora - Volume 9*. Hong Kong: Association for Computational Linguistics, pp. 1–6. DOI: 10.3115/1117729.1117730.

Spencer, Jane (1986). *The Rise of the Woman Novelist: From Aphra Behn to Jane Austen*. Oxford, UK and Cambridge, Mass.: Blackwell.

Stott, Anne (2003). *Hannah More: The First Victorian*. Oxford: Oxford UP.

Szudarski, Pawel (2018). *Corpus Linguistics for Vocabulary: A Guide for Research*. London, New York: Routledge.

Towsey, Mark (2015). “Women as Readers and Writers”. In: *The Cambridge Companion to Women’s Writing in Britain, 1660-1789*. Ed. by Catherine Ingrassia. Cambridge: Cambridge UP.

Van Sant, Ann Jessie (1993). *Eighteenth-Century Sensibility and the Novel: The Senses in Social Context*. Cambridge: Cambridge UP.

Wollstonecraft, Mary (2004). *A Vindication of the Rights of Woman*. London: Penguin Books.

Wood, Lisa (2003). *Modes of Discipline: Women, Conservatism, and the Novel after the French Revolution*. Lewisbug: Bucknell UP.

https://doi.org/10.51305/ICL.CZ.9788076580336.06
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

Peeking Inside the Rhythmic Possibilities of the Portuguese *Decassílabo*

Adiel Mittmann
Universidade Federal de Santa Catarina, Brazil
adiel@mittmann.net.br
0000-0003-2184-2955

Gabriel Esteves
Universidade Federal de Santa Catarina, Brazil
gabrielesteues@gmail.com
0000-0003-4719-6672

Alckmar Luiz dos Santos
Universidade Federal de Santa Catarina, Brazil
alckmar@gmail.com
0000-0002-7896-0103

Abstract

The Portuguese *decassílabo* has a rich internal structure, which is usually analyzed in terms of the rhythmic patterns revealed by verse scansion. In this article, we aim to explore how the stress of one syllable influences another. In order to achieve this goal, we use raw rhythmic patterns, that is, ones in which stress clashes have not been resolved. To exploit these patterns, we apply three methods: tile plots, indices and graphs. Based on our investigation of a corpus of 24 poets, we find that poets who employ similar rhythmic patterns may not share the same preferences concerning stress clashes and may produce different dependence relations. Because many of the associations among syllables reflect the use of one of the two basic types of *decassílabo*, namely the heroic and Sapphic variants, we also propose verse classifications that are useful for understanding and comparing the works of different poets. The main conclusion of this study is that although the *decassílabo* has certain general features, much room remains for poets to create their own interpretations of the form.

1 Introduction

The *decassílabo* has a long history in the poetry of Portugal and Brazil. Throughout its development, much has been written about its features, but there have been few efforts to analyze the large corpora of works that use the form. Such an analysis may, however, promote a better understanding of the internal structure of the *decassílabo*. While it is common to talk about the rhythmic patterns that poets use, little quantitative data has been produced to show how the stress of one syllable influences the stress of another or how stress clashes are handled by different poets.

This article introduces methods to visualize and analyze the relationships between the different syllables in a verse. These methods use the output produced by Aoidos (Mittmann, Wangenheim, et al. 2016), our automatic scansion tool, in a special way: instead of considering actual rhythmic patterns, they harness information about each syllable's raw stress. As such, stress clashes are left unresolved.

Our methods can be divided into three groups: 1) those that deal with dependence between syllables, 2) those that classify verses into basic categories related to traditional variants (heroic or Sapphic) and 3) those that explore stress clashes. These methods produce percentages and indices, and the dependence relationships among the syllables are also depicted graphically using tile plots and graphs.

Although only the Portuguese *decassílabo* is considered here, there is no reason to believe that these methods would not work equally well with other types of verse and with at least some other languages such as Spanish and Italian.

Throughout this article, we use the following conventions: the smallest number in a table column is underlined; the largest one is shown in bold. Numbers that appear within parentheses were calculated from a small set of verses (fewer than 100) and should therefore be interpreted more carefully. Verses listed in a single block are not sequential in the poems that they were taken from; instead they are merely grouped together for convenience's sake.

This article is structured as follows: The Section 2 describes the corpus used in our analyses. The Section 3 then gives general information about the *decassílabo* and raw rhythmic patterns. The Section 4 describes our methods and the analyses that they produced. Our concluding remarks are given in the Section 5.

2 Corpus

This article analyzes a corpus of 24 works, which were written at various times between the 16th and 21st centuries, as shown in Table 1. Each work was assigned a three-letter code derived from its author's name. In this article. these codes are used to refer to the relevant work or its author.

While most of the works analyzed are single books, some (i.e. `MAT`, `NTA`, `RAB`, `ANJ`) represent the complete works of their respective authors. Some of the works contain *decassílabos* exclusively (this is the case for long epic poems such as `CAM` and `TEI`) while others include many other types of verse. Concerning the latter, this study only considers lines of 10 syllables, and the Table 1 only take these into account. One work, `PIN`, is a translation of Dante's *Divina Commedia* while all the other works were originally written in Portuguese.

The number of *decassílabos* in each work varies greatly, ranging from 1,192 in `EMM` to 62,517 in `GLM`. Since works are analyzed individually, there is no risk of a single work dominating the calculations. For our purposes, the presence in each unit of analysis (i.e. the poet's work) of more than 1,000 lines is more important than the total size of the overall corpus (more than 180,000 lines).

Code	Birth	Poet	Work	Verses
CAM	1524	L. de Camões	*Os Lusíadas*	8,816
MEN	1600	Sá de Meneses	*Malaca Conquistada*	10,637
MAT	1636	G. de Matos	Complete works	4,187
DUR	1722	S. R. Durão	*Caramuru*	6,672
COS	1729	C. M. da Costa	*Obras Poéticas*	5,736
NTA	1740	N. T. de Almeida	Complete works	2,074
GAM	1741	B. da Gama	*O Uraguai*	1,377
GON	1744	T. A. Gonzaga	*Cartas Chilenas*	4,172
JAM	1761	J. A. de Macedo	*O Oriente*	8,694
BOC	1765	M. M. B. du Bocage	*Poesias Eróticas*	3,335
MAG	1811	G. Magalhães	*Suspiros Poéticos*	3,787
PIN	1822	X. Pinheiro	*Divina Comédia*	14,227
DIA	1823	G. Dias	*Os Timbiras*	2,004
RAB	1826	L. Rabelo	Complete works	2,250
FAG	1841	F. Varela	*Anchieta*	8,481
SIL	1854	D. Silveira	*Lises e Martírios*	2,070
EMM	1866	E. de Meneses	*Mortalhas*	1,192
TIG	1822	B. Tigre	*Bromilíadas*	3,406
ANJ	1884	A. dos Anjos	Complete works	5,843
NUN	1897	C. A. Nunes	*Os Brasileidas*	8,502
GLM	1951	G. Mattoso	*Sonnettudo*	**62,517**
COL	<1968	C. A. de Oliveira Leite	*Caxias*	3,866
MNC	<1994	M. N. Costa	*Antônio Conselheiro*	2,265
TEI	<2016	J. C. S. Teixeira	*Famagusta*	9,273
				185,383

Table 1: Poets and works in the corpus used in this article, as ordered by the year of the poet's birth

Most of the poets are well known to mainstream literary critics, however COL, MNC and TEI are exceptions. Their years of birth are unknown to us, and the years shown in Table 1 are the years of publication of their respective works. Those three works are lesser-known epic poems; COL and MNC deal respectively with the lives of Duke of Caxias and Antônio Conselheiro, while TEI is a fantastic retelling of the fall of Famagusta to the Ottomans. While the versification of COL, and especially TEI, is more traditional, MNC uses the *decassílabo* in a distinctive way, as we will see.

Not all the source texts are of equal quality. The XML file read by our scansion tool may contain errors that were introduced in any of several previous stages. The print edition may, for example, have included misprints that could only be "fixed" by comparing different editions; the OCR algorithm may have failed to recognize words properly; human editors may have inadvertently mangled verses when updating the text's orthography. As a consequence, the longer a text has been in our corpus, the greater its reliability; while some texts (such as CAM or GAM) have been improved and corrected over many years, others (such as JAM or COL)are recent additions that have not yet undergone this refinement.

The text of GLM is a very recent addition to the corpus and was particularly challenging to analyze. The author made the text freely available to us as a carefully prepared digital edition, which should have simplified the task of feeding it into scansion tools. However, the text used a non-standard spelling system based on a pre-1943 orthography; substantial effort was therefore required to "update" many words in the work.

3 Background

The Portuguese *decassílabo* consists of 10 syllables which are counted from the first syllable of a verse to its last stressed syllable. Any additional syllables are not counted and typically are not rhythmically relevant. Strictly speaking, the word "decassílabo" may refer to any verse that is 10 syllables long; however, this term usually refers to a more specific type of verse, which should technically be called *decassílabo italiano* based on its alleged Italian origins.[1] This is the kind of *decassílabo* that this article is concerned with.

The *decassílabo* is characterized by a mandatory stress on either the sixth or the fourth syllable. Verses that follow the former type are called *heroic*; those of the latter variety are called *Sapphic*. The sixth-syllable variant is the more important one; no poets known to us object to the stress placement on the sixth syllable, but some do reject the choice of the fourth syllable. A verse whose main stress is on the sixth syllable may also contain a stress on any of its other syllables. However, a verse whose main stress is on the fourth syllable has one further restriction: its seventh syllable cannot be stressed. This means that either the eighth syllable must be stressed (so that the seventh is weakly stressed) or the seventh syllable must have no stress at all. If, however, a 10-syllable verse contains stresses on the fourth and the seventh syllables, then it is usually not considered a *decassílabo italiano* but a verse of Provençal or Iberian origin depending on one's critical framework.[2] Verses that do not follow any of these patterns may also be found in works that are thought to use only *decassílabo*, but typically, however, there are few such variations and they can often be plausibly explained as scansion tool errors or mistakes in the original text.

Dividing verses into heroic and Sapphic variants can be challenging when both the fourth and the sixth syllables are stressed. Consider the following verse by CAM:

> Abrindo a porta ao vasto mar patente,

Readers of this line may place stronger stress on either the fourth ("**por**-ta") or the sixth syllable ("**vas**-to"). There is no simple answer to the question of whether this verse is heroic or Sapphic. The scansion of a given line also depends

[1] Spina (2003), for instance, argues that "when we say that Sá de Miranda introduced the *decassílabo italiano* into the Portuguese poetry, this does not mean that this meter is genuinely Italian; it had already been widely used among troubadours as well as in France from the mid-10th century. What was taken from the Italian *decassílabo* was its caesura."

[2] Bandeira (1997), for instance, calls this variation "decassílabo de gaita galega" while Said Ali (1999) describes it as "hendecassílabo ibérico" and Proença (1955) understands it as a "verso provençal" and so on.

on the preceding and following lines, the internal structure of syllabic groups and the semantic weight of words. There are also cases where one classification may be more likely, but it remains possible that the author preferred the other option:

> Vai de Calisto ao seu contrário Polo,

When reading this verse, also written by CAM, it may be tempting to strongly stress the fourth syllable ("Ca-**lis**-to") and so make it a Sapphic variant; this is especially compelling because of the symmetry between the first two prosodic units, each of which has three unstressed syllables followed by a stressed one. However, the argument could also be made that the author did not use Sapphic verses. Supporters of this view might highlight the rather weak stress on "seu" to read this as a heroic verse. This reading is not inconceivable since verses such as this one can also be found in CAM:

> Lembrando-se do seu passado pranto,

This line can only be plausibly understood as a heroic verse whose strongly stressed sixth syllable has its origins in the weak lexical stress on "seu".

The classification of verses with possible stresses on their fourth and sixth syllables as either heroic or Sapphic requires a judgment call that this article does not attempt to make. As will be seen, when classifying verses, we focus instead on those that are unambiguously Sapphic or heroic.

In usual scansion, the rhythmic patterns that are quantified and analyzed attempt to mimic the sequence of stressed and unstressed syllables that a human reader would produce when reading verses. This study, however, analyzes a different type of pattern, one that does not resolve any kind of stress clash. Consider, for instance, the following two lines by NUN:

> excruciantes vindo ela a contorcer-se,
> do grande rio, vindo essas guerreiras

Both these lines contain a stressed sixth syllable, but in the first case, the stress on "**vin**-do" is weakened since the sixth syllable ("**e**-la") must be stressed by the very nature of the line. In the second case, the stress falls on "**vin**-do" in the sixth position and therefore weakens the stressed syllable in the next word ("**es**-sas").[3] When the patterns of these verses are given as 3-6-10 and 2-4-6-10, these lexical stress clashes are hidden from view. Instead the clashes are always effectively resolved by weakening one of the two successive syllables.[4] In fact, traditional versification techniques prescribe that poets should avoid lexical clashes whenever possible. However, this study considers all lexical stress possibilities, regardless of whether they involve two or more adjacent stressed syllables. As such, the patterns for these verses are understood as 3-6-7-10 and 2-4-6-7-10, respectively, even though this does not reflect the way these lines are read aloud. These patterns are referred to as "raw" and opposed to the patterns that arise when stress clashes are resolved.

[3] It should also be noted that in both cases, the final syllable of "vindo" is joined to the next syllable by a synaloepha so that there are no intervening syllables that would prevent a lexical stress clash.

[4] Said Ali (1999) notes that "the rhythmic movement does not support the collision of two strong syllables pronounced with the very same intensity."

Given the nature of raw patterns, this study often states that a syllable "can" or "may" be stressed. We also refer to rhythmic possibilities since the actual stress patterns are not known before the clashes are resolved.

It is important to state openly that "raw rhythmic patterns" are artificial in the sense that they do not correspond to how a verse is actually read. Indeed, they may contain stress clashes that could be considered unpronounceable. The justification for using raw patterns anyway is that they are a useful signal that can be measured from verses. That this signal is more or less detached from the phonetic reality is not especially relevant so long as the signal is correctly measured and provides useful information.

4 Methods

This section presents the methods used to analyze raw rhythmic patterns in this study. These methods were applied to all works in the corpus, and we describe the results below.

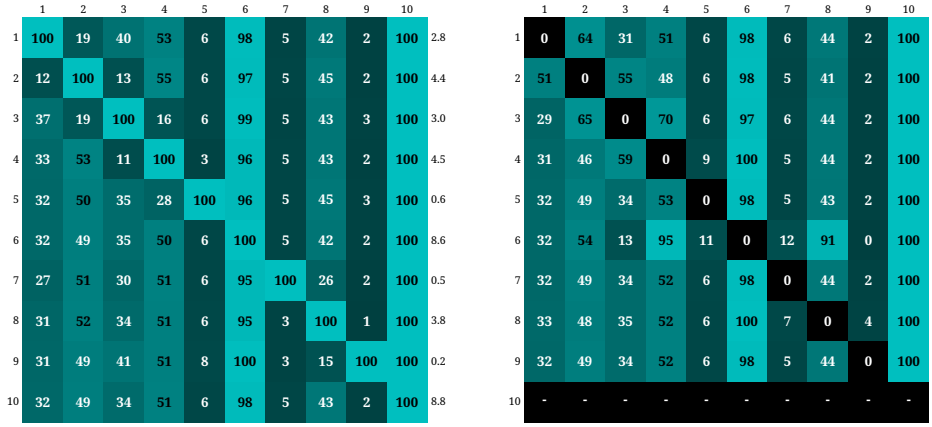
4.1 Dependence

Three kinds of tile plots were used in this study: positive (left-hand column of Figure 1), negative (right-hand column of Figure 1) and contrasting (Figure 2).

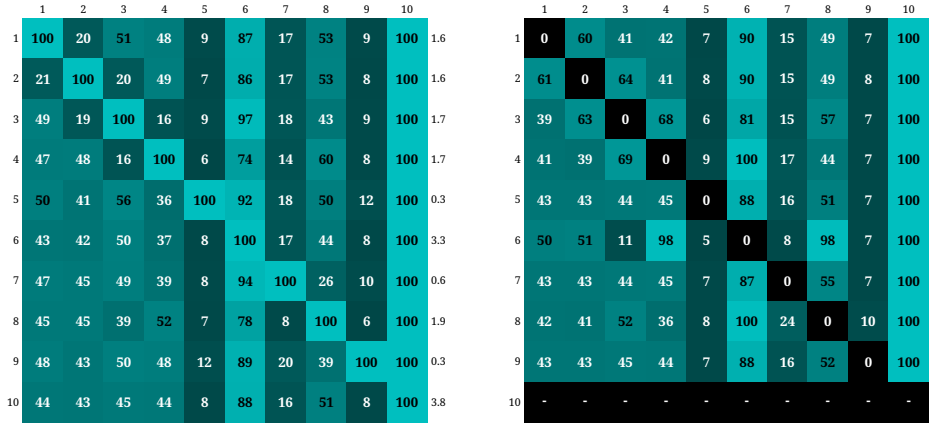
In the *positive* tile plot, each numbered row filters the verses available to include only those that may be stressed on the syllable corresponding to the row number. The total number of verses (in thousands) that have successfully passed through the filter is shown to the right of each row. Row number 1, for example, only contains information about verses whose first syllable may be stressed. The number in each cell is the percentage of verses that may be stressed on the syllable indicated by the corresponding column so that, for instance, the fourth cell in the first row tells us the percentage of verses that may be stressed on the first and the fourth syllables;this figure is relative to the total number of verses that may be stressed on the first syllable. The color of the tiles ranges from dark blue (for lower values) to light blue (for higher values).

The plot is not symmetrical because percentages are calculated relative to the total number of verses considered in each row. In the first positive tile plot of Figure 1, for example, the cell in the first row of the second column and the cell in the second row of the first column have the values of 19% and 12% respectively. Nevertheless they both reflect the same total number of verses—537. The 10th row in a positive tile plot counts all the verses since all those included in the plot are stressed on the 10th syllable. As such, this row gives a summary of the rhythmic possibilities for all these verses.

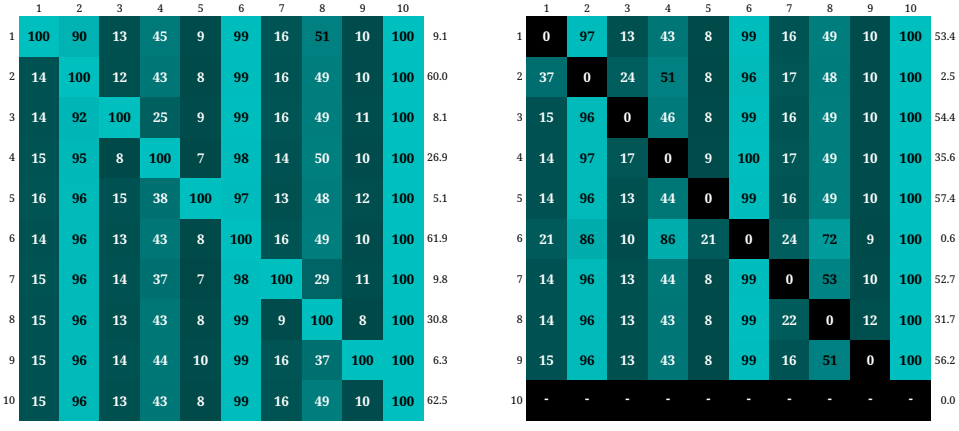
A quick glance at a positive tile plot can reveal interesting patterns. In the case of both `CAM` and `GLM`, the appearance of the even-numbered columns in a brighter shade of blue indicates that these poets exhibit iambic tendencies. It is clear, however, that `GLM` favors this type of verse more frequently. Turning to the last row, we can create a simple *iambicity index* by adding up the percentages



(a) L. de Camões' *Os Lusíadas* (CAM), with 8,816 verses.

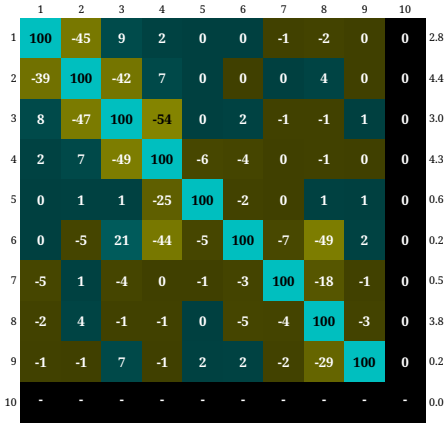


(b) G. Magalhães' *Suspiros Poéticos* (MAG), with 3,787 verses.

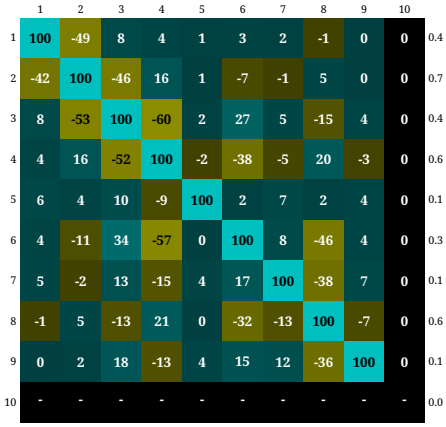


(c) G. Mattoso's *Sonnettudo* (GLM), with 62,517 verses.

Figure 1: Tile plots of rhythmic possibilities. In the plots on the left, each row indicates the percentage of rhythmic possibilities when the corresponding syllable **may** be stressed; the plots on the right show this percentage when the syllable **may not** be stressed.



(a) L. de Camões' *Os Lusíadas* (CAM).



(b) B. da Gama's *O Uraguai* (GAM).

Figure 2: Contrasting tile plots of positive and negative rhythmic possibilities (as seen in Figure 1)

Code	Index	Code	Index	Code	Index	Code	Index
MAG	106.4	FAG	131.3	DUR	144.2	COL	159.9
MNC	114.0	TIG	135.3	PIN	144.7	CAM	162.0
EMM	114.9	BOC	135.5	NTA	150.2	JAM	168.9
MAT	120.2	NUN	137.1	MEN	154.4	DIA	185.2
ANJ	129.4	COS	141.4	SIL	156.4	GON	191.0
RAB	129.8	TEI	142.7	GAM	159.0	GLM	**226.1**

Table 2: Iambicity index for all poets, in ascending order

of syllables in even positions (excluding the 10th syllable) and subtracting those in odd ones; the results are shown in Table 2. The index reflects what can be seen in the plots. It is striking also that GLM stresses the second syllable in the vast majority of cases. This suggests a very peculiar usage of the *decassílabo*.

The *negative* tile plot is similar in most respects to the positive one. The main difference is that each numbered row selects only the verses that may *not* be stressed on the corresponding syllable. When a percentage cannot be calculated due to the lack of any verses in the row, a dash is shown instead of a number. When the percentage is exactly zero, the cell is colored black. The color scheme distinguishes values that are exactly zero from those that have only been rounded to zero. The latter are shown in dark blue instead of black.

The *contrasting* tile plot is calculated by subtracting the negative plot from the corresponding positive plot. Negative values are common in this type of plot and are shown in yellow: the lighter the tile, the more negative the number.

In a contrasting tile plot, each cell contains dependence information. Positive values indicate that if the syllable in the row number position may be stressed,

then the syllable corresponding to the column number also tends to be stressed. Conversely, negative values mean that the other syllable tends not to be stressed.

To some degree, the information in the tile plots makes explicit the relationship between heroic and Sapphic verses. For instance, the negative tile plots for all poets show that when the sixth syllable is not stressed, the fourth and the eighth syllables light up; in other words, when a verse cannot be heroic, it will most likely be Sapphic. But the reverse relationship does not always hold; the negative tile plots for Figure 2 show that when the fourth syllable cannot be stressed, only the poet MAG puts a noticeably increased stress on the sixth syllable. This is due to the larger proportion of Sapphic verses in his work. The reasoning here is as follows: when there are few Sapphic verses, the lack of a stress on the fourth syllable does not mean much for the sixth one. On the other hand, the lack of a stress on the sixth syllable influences the fourth syllable since the only realistic alternative to a heroic verse is a Sapphic one, however few of them there might be.

Among poets who use Sapphic verse more often, a similar effect can be seen regarding the eighth syllable: when it cannot be stressed, a stress on the sixth syllable is more common. Such poets also display other subtler dependence relationships. For instance, Figure 2 shows that the third syllable has a positive influence on the sixth one. In the case of CAM, who uses few Sapphic verses, this influence is minimal. It is, however, very significant for GAM, who employs Sapphic verses more often.

The fifth and the seventh syllables would never be stressed in a realistic reading of a *decassílabo*. This means that when they may be stressed, they will likely be involved in a stress clash that allows them to become weakened in favor of the fourth, sixth or eighth syllables. This can be seen in the positive tile plots for Figure 1: in the rows for the fifth and seventh syllables, the sixth syllable may almost always be stressed as well. The poet GLM shows a marked preference for a stressed second syllable; this is visible in his positive tile plot where a possible stress on the first and the third syllables is very frequently associated with a stress clash with the second syllable.

By adding together the magnitude of the values shown in a contrasting plot, we can measure the general influence of a potential stress in one syllabic position on the possible stress in another position. We call this measure the *dependency index*. When calculating it, we include only cells that have been calculated based on the information from at least 100 verses in both the positive and negative tile plots—the intention is to exclude cells whose values might be too noisy. Table 3 shows the dependency index calculated for all works in the corpus.

The dependence relationship between syllables, as shown in a contrasting tile, can also be visualized as a graph in which nodes represent syllables and edges show their dependence on one another. Graphs for six works are shown in Figure 3. The thickness of the edges reflects the degree of dependence. To reduce the amount of clutter, only cells with an absolute value greater than or equal to 5.0 are represented as edges. Edges have an arrowhead that touches the syllable that is being influenced. The shape of the arrowhead indicates whether the influence is positive (a regular arrowhead) or negative (an inverted

Code	Index	Code	Index	Code	Index	Code	Index
GLM	3.87	MAT	8.70	MEN	9.60	EMM	11.05
TEI	6.33	COS	9.05	BOC	9.77	NTA	11.07
CAM	7.36	MNC	9.06	MAG	10.22	DUR	11.69
COL	8.11	PIN	9.27	ANJ	10.29	JAM	11.82
GON	8.27	RAB	9.51	FAG	10.45	SIL	11.95
TIG	8.50	NUN	9.60	DIA	10.81	GAM	**13.62**

Table 3: Dependency index for all poets, in ascending order

arrowhead). In Figure 3c, for example, one edge joins the sixth and the third syllables; the edge's arrowhead is regular so the relationship is positive. The edge points from the sixth to the third syllable, which means that when the sixth syllable is stressed, the third one tends to be stressed too. This edge is directly based on the sixth row of the third column of the contrasting plot in Figure 2a.

These graphs also highlight certain properties of the corpus. The tendency to avoid stress clashes is clear: there are often (though not always) negative edges between adjacent syllables. We also observe simpler networks among the poets who rely heavily on the heroic *decassílabo*, i.e. (GLM, TEI and CAM). Notably in the work of these poets the sixth syllable is not influenced by the other syllables, presumably because it is almost always stressed anyway.

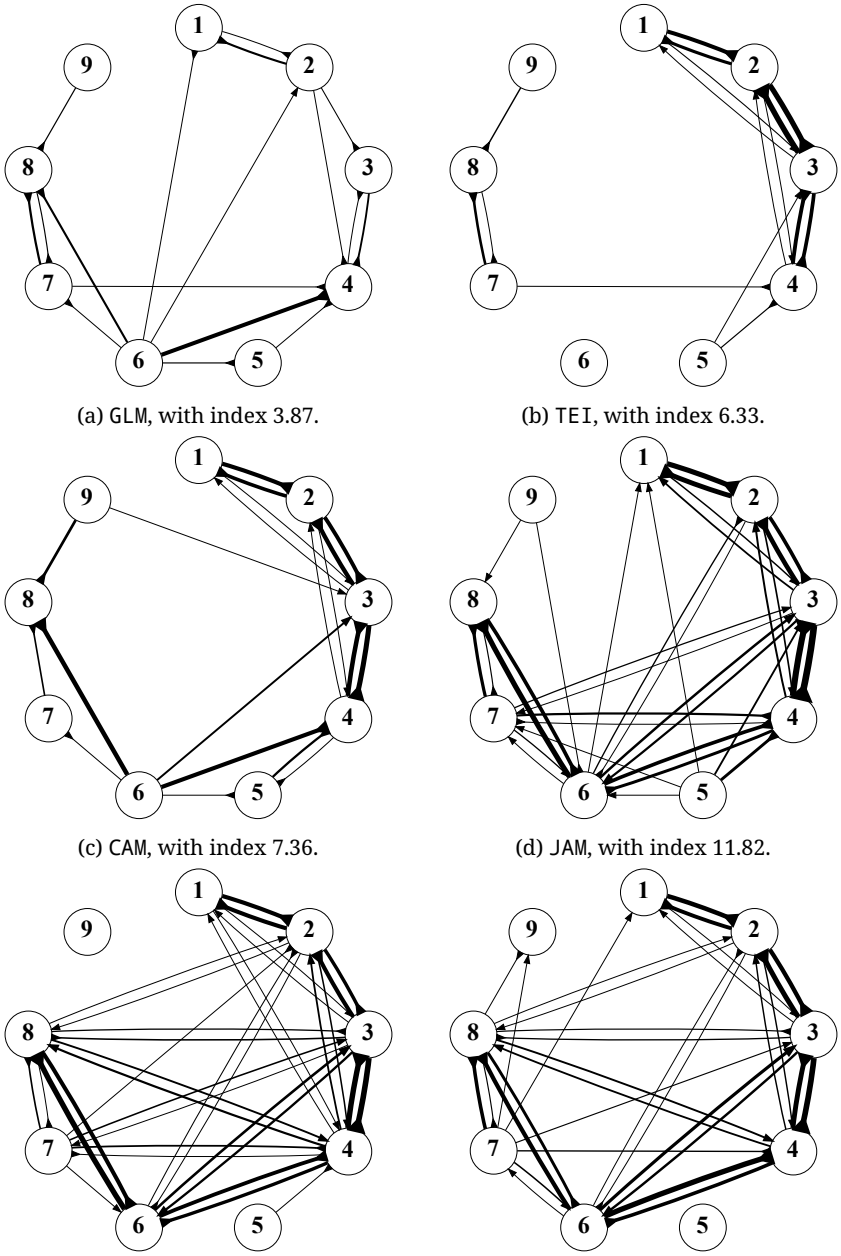
4.2 Verse Types

Readers of heroic and Sapphic verses can usually identify the two forms. Not many verses truly allow for both heroic and Sapphic readings and when both interpretations are permitted, readers may interpret these verses either way. Since this article only considers rhythmic possibilities, verses are classified into the traditional types according to certain criteria.

A verse is said to be *purely heroic* if it can be stressed on the sixth syllable but cannot be stressed on the fourth syllable. A purely heroic verse, thus, does not allow for a Sapphic reading and must be read as a heroic verse.

A verse is said to be *purely Sapphic* if it can be stressed on the fourth syllable but cannot be stressed on the sixth syllable. A purely Sapphic verse, thus, does not allow for a heroic reading. A further constraint is that a purely Sapphic verse must either not be stressed on the seventh syllable or, if such a stress is allowed, then the eighth syllable must be able to carry a stress too. In that case, when the verse is read aloud, the stress clash is resolved in favor of the eighth syllable.

An *ambiguous* verse is one that can be either heroic or Sapphic because it can be stressed on the fourth and the sixth syllables. Such verses are only ambiguous in the strict sense that they theoretically allow for either a heroic or a Sapphic reading; in practice, however, readers treat most if not all of these verses as either heroic or Sapphic.



(a) GLM, with index 3.87.

(b) TEI, with index 6.33.

(c) CAM, with index 7.36.

(d) JAM, with index 11.82.

(e) SIL, with index 11.95.

(f) GAM, with index 13.62.

Figure 3: Influence of a possible stress on one syllable on other stress placements

Code	Heroic	Sapphic	Ambiguous	Provençal	Other	Eighth
NUN	58.53	11.86	29.33	(0.11)	(0.18)	94.0
FAG	57.65	12.50	29.81	(0.00)	(0.05)	99.3
GON	64.74	5.18	30.03	(0.00)	(0.05)	**100.0**
MNC	51.83	8.57	32.54	**(3.58)**	**(3.49)**	30.4
GAM	46.04	20.92	32.75	(0.00)	(0.29)	96.5
MAG	55.45	11.33	32.98	(0.05)	(0.18)	99.8
TEI	**65.16**	(0.73)	34.00	(0.00)	(0.11)	(69.1)
RAB	54.09	10.22	35.20	(0.18)	(0.31)	96.5
ANJ	43.44	20.54	35.58	(0.21)	(0.24)	88.0
DUR	48.49	14.22	37.22	(0.03)	(0.04)	98.7
BOC	48.46	12.92	37.69	(0.48)	(0.45)	98.8
EMM	43.12	16.61	39.26	(0.25)	(0.76)	87.4
MAT	52.16	6.31	40.29	(0.43)	(0.81)	92.4
MEN	48.30	11.04	40.42	(0.05)	(0.19)	97.3
DIA	37.72	20.91	40.77	(0.15)	(0.45)	99.0
NTA	44.07	13.36	41.76	(0.48)	(0.34)	98.6
GLM	56.78	0.79	42.24	(0.05)	(0.14)	82.9
COS	49.04	7.83	43.08	(0.00)	(0.05)	**100.0**
SIL	31.06	**24.83**	43.77	(0.19)	(0.14)	99.2
TIG	51.26	3.49	44.98	(0.03)	(0.23)	91.6
COL	51.63	(2.35)	45.01	(0.85)	(0.16)	(60.4)
JAM	30.91	22.46	45.97	(0.37)	(0.29)	99.4
PIN	40.25	11.53	47.72	(0.17)	(0.33)	97.8
CAM	48.38	2.16	**49.25**	(0.09)	(0.12)	94.7

Table 4: Frequency of verse types (%) for all poets, sorted based on their use of ambiguous verses. The *Eighth* column indicates the frequency (%) of a potentially stressed eighth syllable in purely Sapphic verses.

A *Provençal* verse is one that can be stressed on the fourth and seventh syllables, but cannot be stressed on either the sixth or the eighth syllables. In other words, a basic 4-7-10 rhythmic pattern is the only possible reading of such a verse.

Table 4 shows the percentage of each type of verse for all works in the corpus. Works are sorted based on the presence of so-called ambiguous verses.

An old question in the realm of Portuguese versification is whether the eighth syllable must be stressed in Sapphic verses. The last column of Table 4 shows the percentage of purely Sapphic verses whose eighth syllable may be stressed. For most works, the rate is more than 90% and among those with at least 100 purely Sapphic verses, only one work had a result below 80% (MNC, with 30.4%). It should be noted that in purely Sapphic verses, as defined in this study, there can be no stress on the fifth, sixth, seventh and ninth syllables. A consequence is that if the eighth syllable is not stressed, there will be a total of five unstressed syllables in a row, which is highly uncommon in poetry. We may therefore expect the frequency of a potentially stressed eighth syllable to

be high. As such, a rate of less than 95% probably indicates that in the poet's style, the eighth syllable tends to bear a rather artificial stress since it is not generally derived from lexical stress.

The last column of Table 4 shows that for two poets, (GON and COS), up to a rounding error of one decimal place, 100% of their purely Sapphic verses allow for a stressed eighth syllable. This syllable will most likely be stressed when read since the last stressed syllable was the distant fourth one. Another five poets (FAG, MAG, DIA, SIL and JAM) have a rate of more than 99%.

On the other hand, there are poets such as ANJ, whose purely Sapphic verses may have a stressed eighth syllable only 88.0% of the time. We can therefore find verses such as these in his works:

> Da Fantasia nos itinerários
> Eu, de saudades me despedaçando,
> Outras cabeças aparecerão
> Hoje, porém, que se desmoronou

We would fully expect many, if not most, human readers to stress the eighth syllable in such verses despite the absence of a primary lexical stress there. The sequence without any stress from the fourth to the 10th syllables simply continues for too long for speakers not to add intermediate stresses. In the case of A. dos Anjos', this artificial stress on the eighth syllable in his purely Sapphic verses goes hand in hand with his highly artificial synaloephas.

4.3 Stress Clashes

A *stress clash* takes place when two adjacent syllables could potentially be stressed. When the verse is read aloud, such clashes are resolved by the reader since the two adjacent syllables cannot both be stressed. Consider the second line of Camões' *Os Lusíadas*, for example:

> Que da ocidental praia Lusitana,

Here, in principle, either the last syllable of "o-ci-den-**tal**" or the first one of "**prai**-a" could be stressed by the reader, but a stress on both syllables is not possible. Experienced readers will identify in advance that in this type of verse, a fairly regular heroic *decassílabo*, the fifth syllable is not stressed and so resolve the clash by stressing the sixth syllable, that is, the first syllable of "**prai**-a".

Although these latent stress clashes are not evident in an actual reading, their analysis can reveal stylistic tendencies since not all poets make equal use of them. A *stress clash index* can be calculated as follows. Stress clashes are represented in positive tile plots by the cells adjacent to the main diagonal; however, those percentages are relative to the number of verses in each row and, if we were to average them all, the resulting index could be skewed by the high percentages in rows with few verses. The index is instead calculated based on the total number of clashes, which can be obtained by adding together all the stress clash tiles after multiplying them by the total number of verses in the relevant row. Furthermore, because stress clashes are symmetrical (a stress clash between the first and second syllables also counts as a clash between the

Code	Index	3+4	4+5	5+6	6+7	Code	Index	3+4	4+5	5+6	6+7
MNC	0.23	3.61	11.34	6.22	10.22	ANJ	0.43	6.83	13.50	4.96	13.91
NUN	0.30	4.76	7.64	4.70	9.08	SIL	0.47	8.17	6.23	6.84	14.93
FAG	0.32	3.77	3.40	3.46	15.05	NTA	0.48	6.50	13.72	8.75	14.33
COL	0.34	(4.40)	(2.20)	3.96	14.83	DIA	0.51	3.82	8.11	7.28	**21.83**
JAM	0.35	5.94	1.18	6.03	17.34	RAB	0.56	9.57	3.91	8.71	15.45
CAM	0.35	10.53	11.05	9.26	5.30	BOC	0.57	9.51	3.48	5.82	19.55
GAM	0.35	3.47	4.51	5.68	11.51	MAT	0.58	6.44	10.23	10.94	18.27
TEI	0.37	(4.41)	**(26.47)**	6.37	9.90	COS	0.59	8.02	10.69	10.13	18.88
GON	0.39	5.56	0.93	5.59	5.89	PIN	0.61	**10.55**	8.90	12.50	18.53
DUR	0.41	8.85	1.16	7.02	13.54	EMM	0.63	10.10	7.58	**14.20**	18.29
MEN	0.41	9.20	4.60	9.38	13.45	MAG	0.64	9.79	4.66	8.52	17.29
TIG	0.42	10.08	11.76	8.30	11.68	GLM	**0.73**	7.66	15.12	8.79	17.20

Table 5: Stress clash index and stress clash frequency between syllable pairs. Poets are sorted by the relevant index. The minimum value in each column is underlined; the maximum value is in bold.

second and first), only numbers above the main diagonal are included. This total number is then divided by the total number of verses in the tile. In other words, the stress clash index is the expected number of stress clashes per verse.

Table 5 shows the stress clash index for all works in the corpus, with values ranging from 0.23 (MNC) to 0.73 (GLM) clashes per verse. In general, we may expect a lower index to correspond with stricter adherence to the traditional clash avoidance rule. The high rate of stress clashes in the verses of GLM may relate to the single-mindedness with which he wrote verses with a 2-6-10 pattern. As the reader already knows where the stresses will be, the poet can allow more clashes into his poetry and perhaps use them to his own stylistic advantage. On the other hand, the low index score of MNC may be explained by the infrequency of synaloephas in his work.[5] Synaloephas usually join unstressed syllables and so bring stressed syllables closer together. A low rate of synaloephas therefore translates to fewer stress clashes.

Of particular interest are stress clashes that involve a verse's main ictus. Since there is less room for interpretation in this case, the poet can produce interesting effects by forcing the reader to stress a syllable that would not normally be favored. This is because when there is a stress clash between, say, the first and the second syllables, the reader may resolve it either way and, if the poet specifically intends for one syllable to be stressed, the reader may not notice. In contrast, when the clash occurs between the sixth and the seventh syllables, the reader knows that given tradition and the innate stanzaic tendency towards regularity, it is the former syllable that must bear the stress, whatever word happens to be there. Table 5 therefore also shows the frequency (%) of verses with a clash between the fourth and sixth syllables and their adjacent syllables. In the former case, only purely Sapphic verses are counted; in the latter, we consider only purely heroic ones.

[5] In experiments yet to be published, we show that synaloephas are eight times more common in the works of CAM than in those of MNC.

The following examples illustrate the effect achieved through a clash between the sixth and seventh syllables of purely heroic verses. CAM, who uses such clashes in only 5.30% of cases, provides this model:

> A sazão e o lugar, fazem cruezas
> Assi foi do Saber, alto e profundo,

The effect is remarkable because words that one would expect to be stressed ("**fa**-zem", "**al**-to") are instead robbed of their rhythmic strength. The following two lines come from DIA, who used stress clashes in 21.83% of cases:

> Tornou-lhe Jurucei: "Paz aos Gamelas,
> Ministros de Tupã, núncios da glória?"

5 Conclusion

The methods proposed by this article are a form of distant reading (Moretti 2013). Some of the works in our corpus are not well known (MNC, COL and TEI) and would normally receive little to no attention based on close reading approaches. However, because Aoidos (Mittmann, Wangenheim, et al. 2016) is used to scan all verses automatically, we can include such works in our analyses. We believe it is important not to exclude works solely because of their lack of popularity; we would like to include as many works as possible in our future corpora and so let the verses speak for themselves.

The methods introduced in this article are capable of distinguishing poets who might otherwise appear rhythmically very similar. This is possible because in examining raw stress patterns, we look beyond the actual resolution of a line's rhythm: instead we pay attention to stress clashes before they are resolved. The poets CAM and MAT, whose work is rhythmically similar (Mittmann, Pergher, et al. 2019) are, thus, shown to have different tolerance of stress clashes. On the other hand, the texts of MEN and DUR, though rhythmically different, show a very similar tolerance of these clashes. The analysis of such clashes is therefore another tool for telling poets apart. Our previous work has shown how poet B. Tigre was able to skillfully mimic rhythmic patterns in his parody of L. de Camões' *Os Lusíadas*. The present article shows that Tigre allows stress clashes between the sixth and the seventh syllables of purely heroic verses with twice the frequency of Camões.

The main conclusion of this article is that although we may speak about the general form of the *decassílabo* or its development over time, individual poets can and do use the *decassílabo* differently and so effectively create their own types of verse. To take one example, although the structure of heroic verse is generally quite flexible, GLM imposes the rule that the second syllable must be stressed in addition to the sixth. In the case of Sapphic verses, for some poets (for example, GON, FAG and COS), a stress on the eighth syllable is a strict requirement. Others clearly do not see the need for such a stress (or at least not one derivable from lexical stress) on the eighth syllable.

Acknowledgments

The authors would like to thank Paulo Pergher and Samanta Maia for their assistance with converting and proofreading many of the works included in the corpus. Samanta, in particular, was responsible for updating the spelling of `GLM`, which was a major undertaking.

References

Bandeira, Manuel (1997). *Seleta de prosa*. Rio de Janeiro: Nova Fronteira.

Mittmann, Adiel, Paulo Henrique Pergher, and Alckmar Luiz dos Santos (2019). "What Rhythmic Signature Says About Poetic Corpora". In: Quantitative Approaches to Versification (2019).

Mittmann, Adiel, Aldo von Wangenheim, and Alckmar Luiz dos Santos (2016). "Aoidos: A System for the Automatic Scansion of Poetry Written in Portuguese". In: 17th International Conference on Intelligent Text Processing and Computational Linguistics (2016).

Moretti, Franco (2013). *Distant reading*. London: Verso.

Proença, Manuel Cavalcanti (1955). *Ritmo e poesia*. Rio de Janeiro: Simões.

Said Ali, Manuel (1999). *Versificação Portuguesa*. São Paulo: Editora da Universidade de São Paulo.

Spina, Segismundo (2003). *Manual de versificação românica medieval*. São Paulo: Ateliê Editorial.

https://doi.org/10.51305/ICL.CZ.9788076580336.07
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

Rhythm and Vocabulary of Greek Hexameter: From Formula to Topolexis

Juan Sebastián Páramo Rueda
Universidad Nacional de Colombia
sebastianparamo@gmail.com
0000-0002-9994-6837

Anastasia Belousova
Universidad Nacional de Colombia / Moscow State University, Russia
metroyritmo@gmail.com
0000-0002-0771-4271

Paula Ruiz Charris
Universidad Nacional de Colombia
kromyon@gmail.com
0000-0002-1792-6286

Abstract

The article discusses the results of our application of a computer program created for the automatic analysis of lexical distribution based on rhythmic position in Greek hexameter.

For this purpose, we introduce the concept of the topolexis (in Greek τοπολέξις: from τόπος "place" and λέξις "expression, word"), which describes each word based on its position in the given line and is expressed as the word in combination with two sets of numerals. The topolexis "52Ἀχιλῆος62", for example, indicates that the word Ἀχιλῆος begins at the second syllable of the fifth foot (52) and ends at the second syllable of the sixth foot (62).

We investigate the behavior of topolexes in a corpus that includes Homer's *The Iliad* and *The Odyssey* and Apollonius Rhodius' *The Argonautica*. We find that the distribution of topolexes of different frequencies varies among these texts. While *The Argonautica* contains a greater number of unique topolexes, higher-frequency topolexes are more common in Homer's poems. The "formulaicity ratio", which we define as the ratio of distinct topolexes in a text to its overall topolexis count, is higher for Homer. In addition, we obtain and analyze data about Hesiod's *The Theogony*. Although *The Theogony* is only 1,023 lines long, it exhibits the same tendencies as Homeric hexameter. We are, thus, able to clearly and accurately compare the behavior of topolexes in epic hexameter in the formulaic style and in its literary imitation by Apollonius.

Lastly, we run a test to compare the performances of the topolexes and the most frequent words (MFW) as stylometric indicators for determining text authorship. We find that while topolexes enable us to correctly cluster fragments by their author, they do not outperform the MFW in this respect.

1 Introduction

The present study is devoted to the relationship between rhythm and vocabulary in verse, or more specifically, the organization of linguistic material within the poetic line.[1] Our methodology is inspired by two research streams, of which one stems from work on verse syntax, which has been advancing in Russian verse studies since the 1920s, while the other is based on the oral-formulaic theory that originated with Milman Parry and Albert Lord's work on Homeric and other epic traditions.

The first of these two inspirations refers particularly to Osip Brik's work. In his pioneering study, "Ritm i sintaksis" (Rhythm and Syntax) (1927/2012), Brik pointed out that poetic speech reflects a dynamic interaction of rhythm, syntax, and semantics. He saw the most conspicuous display of this interaction in what he termed "rhythmic-syntactic clichés (formulae)": "A poetic cliché is precisely the result of a complete fusion of the rhythmic, syntactic, and semantic aspects of the poetic word [whereby] conventional word combinations are formed in [terms of] which the poet himself thinks" (Brik 2012, p. 520). Brik's ideas have been fruitfully developed by Russian verse scholars over the last century. Mikhail Gasparov's contribution is especially significant in this respect: drawing on Brik's and Boris Yarkho's ideas, Gasparov succeeded in creating a linguistically substantiated method, which he used to analyze a wealth of material.[2] In particular, he claimed that "(a) a *rhythmic cliché* is a fixed sequence of rhythmic words (which in verse studies is referred to as *word boundary line variation* [словораздельная вариация] [...]; (b) a *syntactic cliché* is a recurring sequence of the same phrase/sentence constituents [...]; (c) a *rhythmic-syntactic cliché* is a combination of both [(a) and (b)] [...]; (d) a *rhythmic-syntactic formula* is a rhythmic-syntactic cliché involving the exact recurrence of one or more words" (Gasparov 1986, p. 189).

Gasparov's associates and followers have enhanced his methods and applied them to different poetic material in recent decades. In this way, they have also developed his theories about the relationship between rhythm and syntax (Akimova 2017a; Belousova et al. 2019; Kruglova et al. 2019; Tarlinskaja 2015; Tverianovich 2019).

The second methodological tradition that we draw on arose from the study of what are known as Homeric formulae. Parry and Lord set out to demonstrate that *The Iliad* and *The Odyssey* were, in fact, instances of oral poetry and that these formulae played a special role in the two texts. In this context, Parry defined a formula as "*une expression qui est régulièrement employée, dans les mêmes conditions métriques, pour exprimer une certaine idée essentielle*"[3] (Parry 1928, p. 16; Lord 1960; Parry 1971).[4] Curiously, Parry, like Brik, linked vocabulary, rhythm, and semantics together. In the years since, classical philologists and specialists in other types of epic have made copious and often extremely

[1] The authors wish to thank the editors and anonymous reviewers of this volume for their insightful comments and suggestions.

[2] Concerning Brik, Yarkho, Gasparov, and the history of the study of verse syntax, see Akimova 2012, 2015, 2017b, including an extensive bibliography; see also Tver' yanovich 2008, pp. 110–114.

[3] "an expression regularly used under the same metrical conditions to express an essential idea".

[4] For a discussion of and amendments to this definition and related amendments, see Nagy 1990.

productive efforts to apply the oral-formulaic theory to texts from various poetic traditions (see Foley 1985).

Predictably the advent of computer tools has also prompted an ever-increasing number of statistical analyses of the formulaic style. In the second edition of his influential book on *The Iliad*, Martin Mueller, thus, added an entire chapter called "Homeric Repetitions" (2009, pp. 135–172) which used statistics to describe the workings of recurring elements in the Homeric epic. More recently, Sklaviadis et al. (2019) have completed a similar study that uses n-gram analysis and independently confirms Mueller's findings. This work also examines recurrences in relation to their linear location by identifying the positions within hexameter verse where n-grams recur most often (2019, p. 243 Fig. 4).

The current study continues this search for applications of computer-aided analysis of the formulaic style. While the new methodology that we propose does not directly evolve from any of the approaches described above, it is our hope that it will produce more in-depth analysis and more accurate descriptions of recurring lexical-rhythmic elements in Greek hexameter and verse in general.

2 Method

We developed a Python program called `Hexametron` that automatically scans Greek hexameter. This program is also able to process lines that present challenges because, for example, they contain the false diphthongs εω and εα, omit a consonant whose presence in a word must be assumed (e.g. ἔδFεισε instead of ἔδεισε), or include a caesura in the third foot that must be assumed to produce an elongation (e.g. χείλει ἐφεσταότες · ἀπὸ γὰρ δειδίσσετο τάφρος, Il. XII, 52). Of the 15,682 total lines in The Iliad, `Hexametron` correctly scanned 15,328, that is, 97.7%.

Using `Hexametron`'s output, we developed a second program that associates each word with its position within the hexameter. We named this association a *topolexis* (τοπολέξις: from the Greek τόπος, "place" and λέξις, "expression, word"). Using this approach, the first line of *The Iliad* (Μῆ-νιν ἄ-|-ει-δε θε-|-ὰ Πη-|-λη-ϊ-ά-|-δεω Ἀ-χι-|-λῆ-ος) was transformed, for example, into 11Μῆνιν12 13ἄειδε22 23θεα31 32Πηληϊάδεω51 52Ἀχιλῆος62. Here, the word Ἀχιλῆος begins at the second syllable of the fifth foot (52) and ends at the second syllable of the sixth foot (62) while the word θεά begins at the third syllable of the second foot (23) and ends at the first syllable of the third foot (31). As can be seen, in these double-digit numbers that precede or follow the word, the first digit refers to the foot where the word begins or ends and the second to the syllable in that foot.

To explore the behavior of topolexes, we ran a series of tests on a textual corpus consisting of portions of *The Iliad* (the first 5,837 lines) and *The Odyssey* (the first 5,837 lines) and the entire text of Apollonius Rhodius' *The Argonautica* (5,837 lines). The size of the text sample was, of course, determined by the length of the shortest text (*Argonautica*). For each text, we established all of

The Iliad, 1–5837		*The Odyssey*, 1–5837		*The Argonautica*	
Topolexis	Rec.	Topolexis	Rec.	Topolexis	Rec.
12δ’12	356	12δ’12	349	21δ’21	221
21δ’21	247	11ὥς11	239	12δ’12	184
33καί33	203	21δ’21	229	32τε32	167
12δέ12	199	33καί33	215	11καί11	142
32καί32	193	11καί11	184	33καί33	136
42καί42	169	22δ’22	183	22δ’22	125
22δ’22	165	11ἀλλ’11	178	32δέ32	121
11καί11	160	32καί32	177	52δέ52	121
11ἀλλ’11	158	42καί42	162	12δέ12	113
11ὥς11	150	32τε32	145	22καί22	112

Table 1: The most common topolexes and their recurrence

the topolexes that it contained as well as the number of times that each one appeared.

3 Findings and Data Analysis

Table 1 presents the 10 most common topolexes along with the number of times they occur in each text. As the table makes clear, each topolexis contains a coordinating conjunction that occurs in various positions in the hexameter (the sole exception is ὥς, which is an adverb, in Homeric texts). On their own, these data shed some light on the functioning of hexametric lines and their syntactic tendencies.[5]

We were, however, primarily interested in the recurrence of content words and their distribution. For this reason, we excluded all function words from our lists by applying a stop-word list (this included conjunctions, articles, demonstrative adverbs, and pronouns). Having, thus, filtered our data, we obtained the results given in Table 2 (for the full data, see footnote 5).

Figure 1 is a line-plot which represents the 20 most common topolexes. These are sorted in descending order by frequency (x-axis) and number of recurrences (y-axis) and plotted on a logarithmic scale. As can be seen, *The Argonautica* differs strikingly from Homer’s works: there are fewer recurrences of the most common topolexes in the former than in the latter.

Table 3 shows the number of topolexes that occur between 1 and 20 times in each sample (for the full data, see footnote 5). Here the first row gives the number of unique topolexes in each text while subsequent entries list the number of topolexes that recur, for example, 3, 10, 15, and 17 times. We can see that in *The Iliad*, for example, there are 14 topolexes that recur 16 times while there are 12 such topolexes in *The Odyssey* and 3 in *The Argonautica*.

[5] The full data are available in CSV format at https://cutt.ly/ohzJfjY.

The Iliad, 1–5837		*The Odyssey*, 1–5837		*The Argonautica*	
Topolexis	Rec.	Topolexis	Rec.	Topolexis	Rec.
53Ἀχαιῶν62	104	53Ὀδυσσεύς62	72	42ἀλός43	30
53Ἀχαιοί62	62	53Ἀθήνη62	60	51Αἰήταο62	27
52Ἀγαμέμνων62	61	61θυμῷ62	44	12φάτο13	25
53Ἀχαιούς62	46	61εἶναι62	40	13ἔφη21	24
33Ἀχαιῶν42	44	12ἔφατ’13	38	11νῆα12	23
32προσέφη41	44	53θαλάσσης62	36	53Ἰήσων62	23
53μάχεσθαι62	39	32προσέφη41	34	53ἰδέσθαι62	22
53Ἀθήνη62	39	51δῖος52	33	53ἔειπεν62	22
33Διός41	35	33θεοί41	33	33Διός41	20
61ἵππους62	35	33θεά41	32	52προσέειπεν62	20

Table 2: The most common topolexes and their recurrence (content words only)

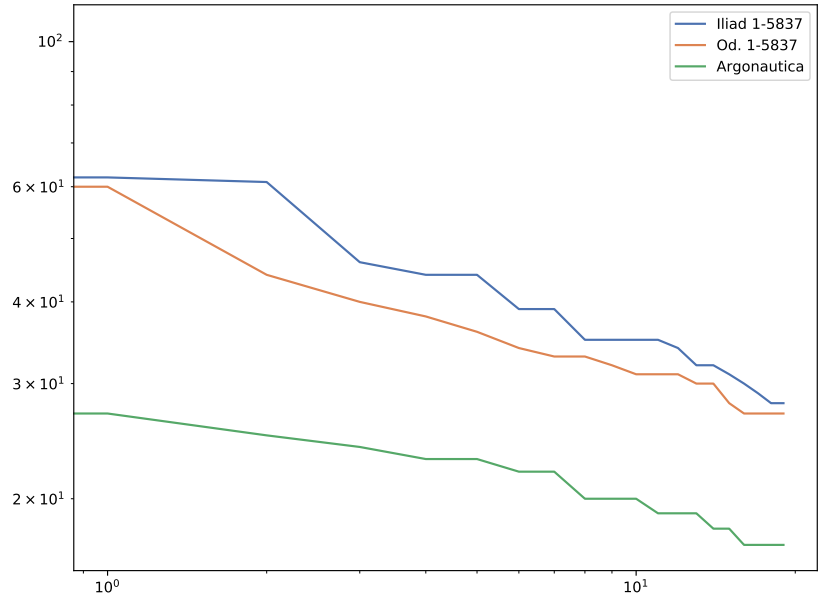


Figure 1: The 20 most common topolexes (content words only) and their recurrence plotted on a logarithmic scale

The number of unique topolexes in *The Argonautica* also turns out to be 20% higher than the figure for *The Iliad* and 26% higher than the result for *The Odyssey*. Notably, the results for topolexes that recur 2 to 10 times are somewhat similar across all three texts. However a trend emerges when we consider topolexes that recur more than 10 times: there are fewer of these higher-frequency topolexes in *The Argonautica* than in the two Homeric texts. Figure 2 plots the data in Table 3 on a logarithmic scale. We can see a clear differ-

Frequency	*The Iliad*, 1–5837	*The Odyssey*, 1–5837	*The Argonautica*
1	9599	9120	11517
2	2004	2018	1968
3	750	790	708
4	361	374	296
5	212	221	164
6	132	146	120
7	82	86	75
8	61	55	54
9	39	47	35
10	31	23	21
11	29	22	15
12	26	18	6
13	14	23	6
14	9	17	6
15	12	7	10
16	14	12	3
17	7	8	4
18	5	10	2
19	5	6	3
20	4	6	3

Table 3: Total number of topolexes of each frequency (content words, from 1 to 20)

ence between *The Iliad* and *The Odyssey* on the one hand and *The Argonautica* on the other: high-frequency topolexes occur at a higher rate in the Homeric works.

Based on the same dataset, if we increase our range to the 50 most common topolexes and normalize the values, we obtain the results in Figure 3 (violin plot). This graph is probably most representative. It clearly shows that the distribution of topolexes of various frequencies is nearly identical in the two Homeric poems and very different in Apollonius' text. In addition, we see a thickening of *The Argonautica* graph that reflects its greater number of unique topolexes compared to the Homeric samples. In contrast, the thickening in the top part of the graphs for *The Iliad* and *The Odyssey* corresponds to their higher count of high-frequency topolexes.

Finally, we wished to investigate the correlation between the total number of topolexes and the number of recurring topolexes in our texts. To do this, we used the following formula:

$$FR = \frac{total\ number\ of\ topolexes}{number\ of\ distinct\ topolexes} \tag{1}$$

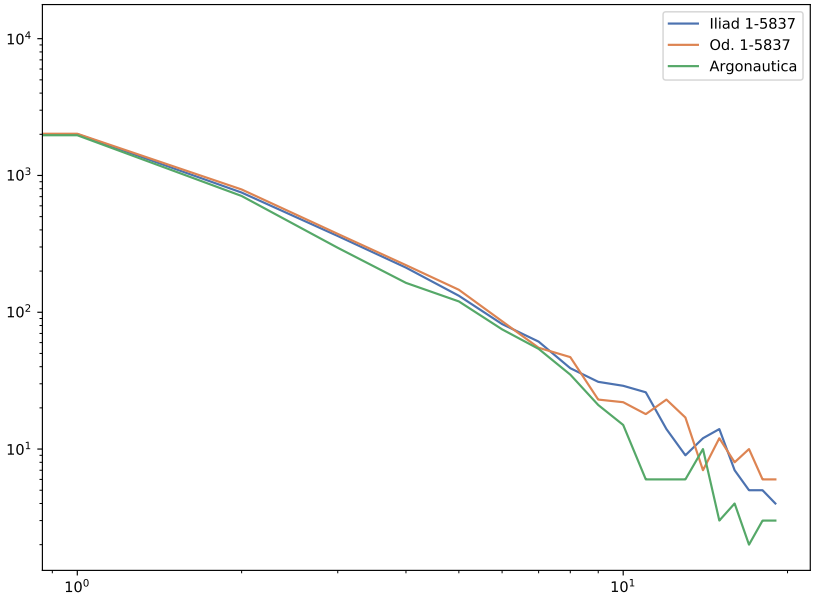


Figure 2: Total number of topolexes of each frequency (content words, from 1 to 20) plotted on a logarithmic scale

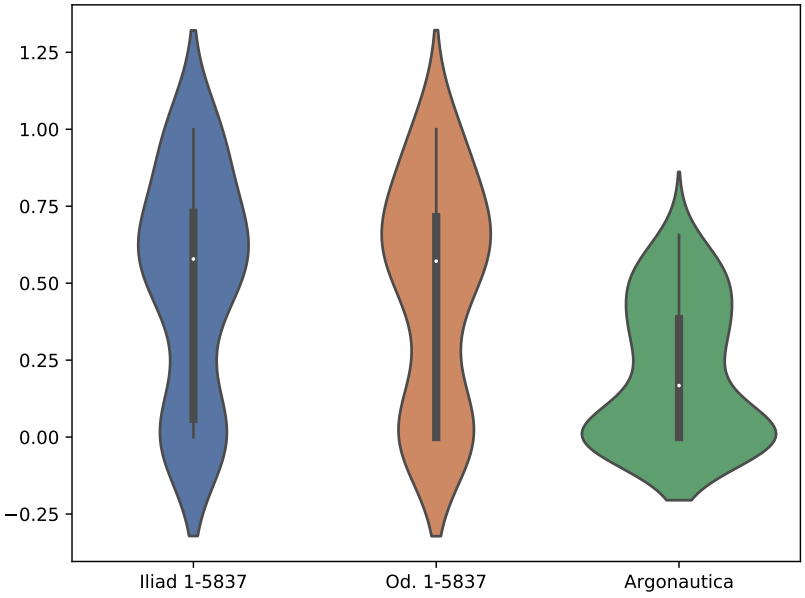


Figure 3: Total number of topolexes of each frequency (content words, from 1 to 50); normalized data

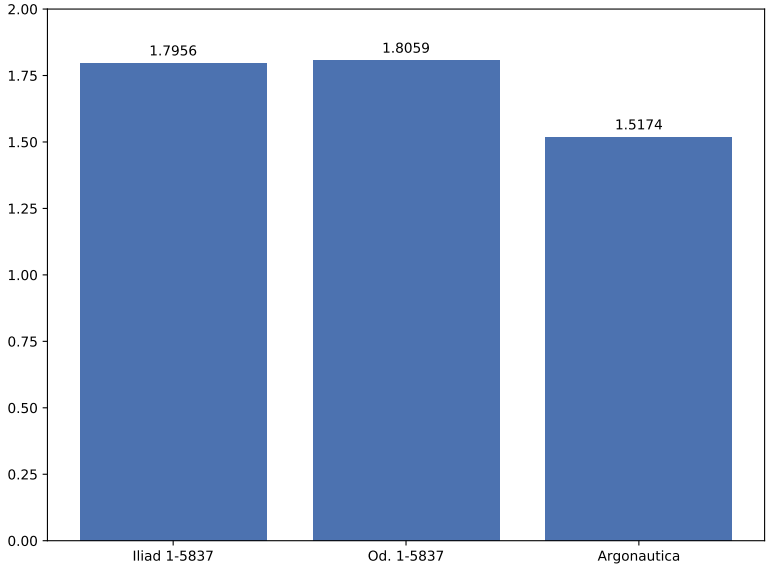


Figure 4: Formulaicity ratios

We called the resulting number the "formulaicity ratio".[6] Figure 4 shows our results for the three samples (*The Iliad*: 1.80; *The Odyssey*: 1.81; *The Argonautica*: 1.52).

The formulaicity ratio was, thus, higher for *The Iliad* and *The Odyssey* than it was for *The Argonautica*. This suggests that the Homeric texts contained a lower number of unique topolexes, that is, they showed greater "formulaicity".

4 Analysis of an Additional Corpus Including Hesiod's Hexameter

We wished to apply the analysis above to a corpus that included an additional sample of Ancient Greek hexameter: Hesiod's *The Theogony*. There are only 1,023 lines in this text so its inclusion in the main corpus would have excessively reduced the size of the other sample texts. After selecting fragments of the same size from *The Iliad*, *The Odyssey*, and *The Argonautica*, we therefore compiled a new corpus. The line plot in Figure 5 depicts the 20 most common topolexes in each text, as sorted in descending order by frequency (x-axis) and number of recurrences (y-axis) and plotted on a logarithmic scale (cf. Figure 1). As the graph illustrates, although *The Theogony* is not identical to the Homeric works, it is closer to them than it is to Apollonius's text.

[6] As briefly noted in our introduction, the definition of a "formula" remains controversial. Based on our data and for strictly practical reasons, we therefore limited the term "formula" to topolexes that occurred more than once in the same rhythmic position in the hexameter. This was also the sense in which we used the term "formulaicity".

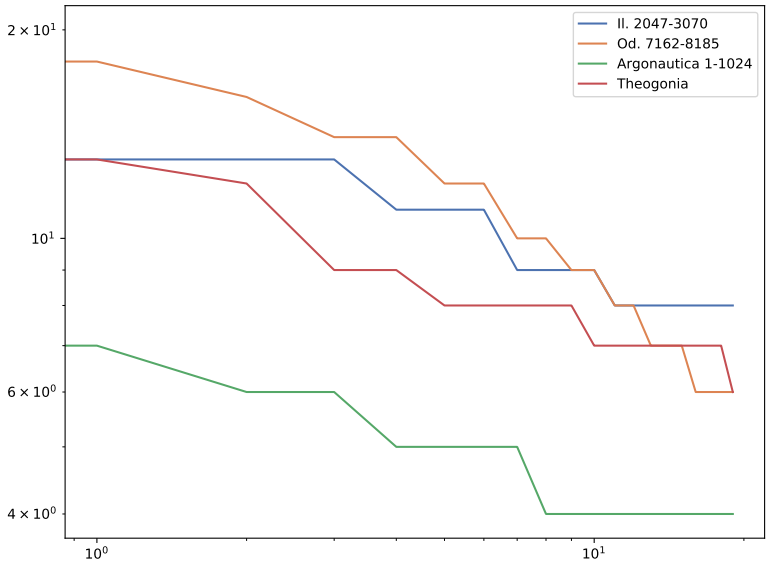


Figure 5: The 20 most common topolexes and their recurrence (content words only) plotted on a logarithmic scale

Figure 6 is analogous to the violin plot model in Figure 6 with the addition of *The Theogony*. Based on even this small sample, we may conclude that Hesiod demonstrates the same behavior regarding topolexes as Homer. This behavior is also characteristic of the formulaic style of poetry.

The formulaicity ratios for this corpus are as follows:

- *The Iliad*, 2047–3070: 1.30
- *The Odyssey*, 7162–8185: 1.32
- *The Argonautica*, 1–1024: 1.15
- *The Theogony*: 1.32

5 A Comparative Stylometric Experiment

One popular indicator in contemporary stylometry is the relative frequency of the most frequent words (MFW) (see Plecháč et al. 2018, including its bibliography). This indicator is obtained as follows: we count the number of times a given word occurs in a text and then divide this figure by the total word count. Once these values are obtained, we can choose a given number of most frequently used words. Based on these MFW, we can then perform a multivariate analysis.

To calculate the stylometric proximity between texts or sets of texts, the "distances" are measured between them. Just as we measure the distance between two points on a Cartesian plane, we calculate the "distance" between

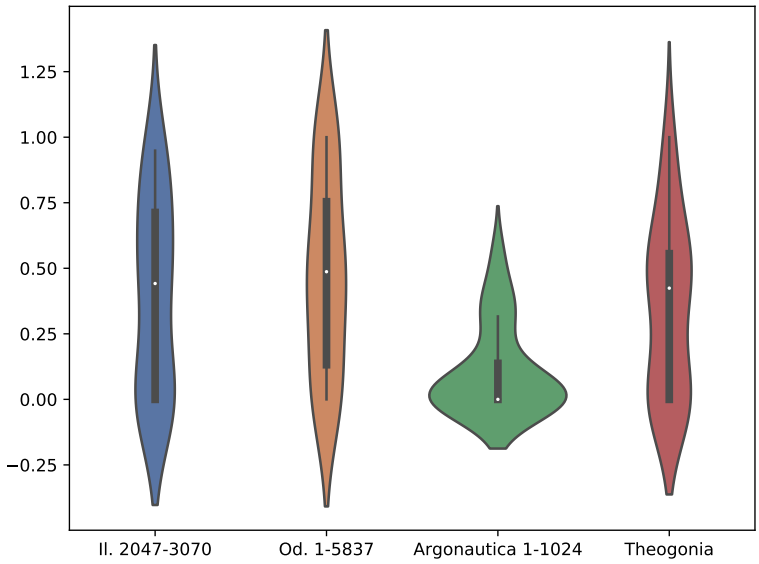


Figure 6: Total number of topolexes of each frequency (content words, from 1 to 50); normalized data

two texts (or sets of texts) based on the relative frequency of their most frequent words. Each text (or set of texts) is, thus, represented by a point in a multi-dimensional space where each coordinate is the value of the corresponding relative frequency, and the number of dimensions is the total number of most frequent words chosen for analysis. The distances thus obtained are then subjected to hierarchical clustering, and the outcome is often visualized as a tree diagram.

Generally speaking, the tree diagram that results from this cluster analysis is a graph with the following format: its y-axis shows the distances between the two texts while its x-axis shows the blocks established from the binary nodes based on those distances.

To assess the stylometric potential of topolexes, we compared the results of applying the clustering method to 1) the traditional indicator of MFW relative frequency and 2) the relative frequency of topolexes. Since we sought only to compare these two methods and were not interested in comparing the works with each other, we created a corpus consisting of Homer's *The Iliad* and Apollonius Rhodius' *The Argonautica*. We then ran the following test:

1. First of all, we divided *The Iliad* and *The Argonautica* into blocks of 750 lines each (this yielded slightly more than 5000 words per block, which is believed to be the minimum word count needed for this type of analysis to be efficient). We then calculated the distances between the texts and produced two tree diagrams: the first was based on the top 50 MFW while the second represented the 50 most common topolexes.

As can be seen in Figures 7 and 8, when applied to blocks of this length, both indicators yielded reliable results. The blocks from *The Iliad* and *The Argonautica* are, thus, respectively grouped together and form two separate clusters.

2. We repeated the experiment with blocks containing 350 lines each. As Figures 9 and 10 show,[7] both stylometric indicators again yielded good results.

3. We also achieved reasonably good results for both stylometric indicators when using smaller blocks of 100 lines each. To make this work, however, we had to extend the ranges to the 350 most common words and the 900 most common topolexes, as illustrated in Figures 11 and 12 (see footnote 7).

4. Finally, when using blocks smaller than 50 lines, we found that no matter how much we extended the frequency ranges (and even when we went as high as the 10,000 most common words/topolexes), we failed to obtain satisfactory results for either the MFW or the topolexis indicator (the best results can be seen in Figures 13 and 14; see footnote 7).

Our initial hypothesis was that topolexes might be a better stylometric indicator for verse texts. The results above did not, however, confirm our thesis: while topolexes allow us to correctly cluster fragments by author, they do not outperform the MFW in this respect.

6 Conclusion

We investigated the behavior of topolexes in a corpus of texts which included Homer's *The Iliad* and *The Odyssey* and Apollonius Rhodius' *The Argonautica*.

Our findings showed that Apollonius' text has a higher number of unique topolexes while Homer's works contain more high-frequency topolexes. The "formulaicity ratio", i.e. the ratio of the number of different topolexes in a text to its total topolexis count, is higher for Homer than it is for Apollonius. Our indicator, thus, accurately reflected the greater formulaicity of Homer's texts.

Our analysis of Hesiod's *The Theogony* by the same method demonstrated its similarity to Homer's poems in terms of topolexis behavior.

Finally, we ran a test to compare the performances of the topolexis and the MFW as stylometric indicators for determining authorship. We found that while topolexes enable the correct clustering of fragments by author, they do not outperform the MFW in this regard.

We have limited the current article to the presentation of the main quantitative results of our study. However a qualitative analysis of issues such as poetic syntax, style, and topolexis distribution by line has yet to be conducted.

[7] Tree diagrams corresponding to Figures 9 to 14 can be found at: https://cutt.ly/ohzJfjY. That site also includes all of the texts scanned and tokenized by topolexis (in CSV format) as well as tables with full topolexis lists for each text sorted by frequency.

Figure 7: MFW

Figure 8: Topolexes

As a next step for future research, we envisage a study of typical topolexis sequences within the hemistich and the line as well as the application of our method to other verse forms. In the meantime, we hope that we have added another useful instrument to the toolkit for stylometric analysis.

Acknowledgments

This paper is part of a research project based at Lomonosov Moscow State University and supported by Russian Science Foundation Grant № 19-78-10132. We thank Mikhail Oslon for his help with preparing the English version of this paper.

References

Akimova, Marina Vyacheslavovna (2012). “Osip Maksimovich Brik. Ritm i sintaksis (Materialy k izucheniyu stixotvornoj rechi). Vstupitel'naya zametka, podgotovka teksta i primechaniya M. V. Akimovoj”. In: *Slavyanskij stix, vol. 9*. Moskva: Rukopisnye pamyatniki Drevnej Rusi, pp. 501–550. URL: http://ruformalism.feb-web.ru/docusr/Brik-Akimova.pdf.

Akimova, Marina Vyacheslavovna (2015). “Tradicii izucheniya russkogo stixotvornogo sintaksisa: O. M. Brik i M. L. Gasparov”. In: *Antropologiya kul'tury, vol. 5*. Ed. by Marina Vyacheslavovna Akimova and Dmitrij Vadimovich Val' kov. Moskva: Novoe izdatel'stvo, pp. 351–356. URL: https://www.academia.edu/15279097/.

Akimova, Marina Vyacheslavovna (2017a). “Citata ili klishe v poėticheskom tekste: popytka razgranicheniya”. In: *M. L. Gasparovu-stixovedu: In memoriam*. Ed. by Marina Vyacheslavovna Akimova and Marina Grigor' evna Tarlinskaya. Moskva: Yazyki slavyanskoj kul'tury, pp. 221–254. URL: https://www.academia.edu/34475225/.

Akimova, Marina Vyacheslavovna (2017b). “Tradicii izucheniya russkogo stixotvornogo sintaksisa: B. I. Yarxo i M. L. Gasparov”. In: *Trudy Instituta russkogo yazyka im. V.V. Vinogradova* 14, pp. 89–112. URL: http://ruslang.ru/doc/trudy/vol14/05-Akimova.pdf.

Belousova, Anastasia and Juan Sebastián Páramo Rueda (2019). “Macroanalysis of the strophic syntax and the history of the italian ottava rima”. In: *Quantitative Approaches to Versification*. Ed. by Petr Plecháč, Barry Scherr, Tatyana Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. Prague: The Institute of Czech Literature of the Czech Academy of Sciences, pp. 23–30. URL: http://versologie.cz/conference2019/proceedings/belousova-paramo-rueda.pdf.

Brik, Osip Maksimovich (2012). “Ritm i sintaksis (Materialy k izucheniyu stixotvornoj rechi). Vstupitel'naya zametka, podgotovka teksta i primechaniya M. V. Akimovoj”. In: *Slavyanskij stix, vol. 9*. Moskva: Rukopisnye pamyatniki Drevnej Rusi, pp. 501–550. URL: http://ruformalism.feb-web.ru/docusr/Brik-Akimova.pdf.

Foley, John Miles (1985). *Oral-Formulaic Theory and Research: An Introduction and Annotated Bibliography*. New York: Garland Publishing, Inc.

Gasparov, Mixail Leonovich (1986). "Ritmiko-sintaksicheskaya formul'nost' v russkom 4-stopnom yambe". In: *Problemy strukturnoj lingvistiki. 1983*. Moskva: Nauka, pp. 181–199.

Kruglova, Anastasia, Olga Smirnova, and Tatyana Skulacheva (2019). "Syntax and pauses in a verse line: Statistical analysis". In: *Quantitative Approaches to Versification*. Ed. by Petr Plecháč, Barry Scherr, Tatyana Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. Prague: The Institute of Czech Literature of the Czech Academy of Sciences, pp. 113–124. URL: http://versologie.cz/conference2019/proceedings/kruglova-smirnova-skulacheva.pdf.

Lord, Albert (1960). *The Singer of Tales*. Cambridge, Mass.: Harvard University Press.

Mueller, Martin (2009). *The Iliad*. London: Bristol Classical Press.

Nagy, Gregory (1990). "Greek Mythology and Poetics". In: Ithaca: Cornell University Press. Chap. Formula and Meter: the Oral Poetics of Homer, pp. 18–35. URL: http://nrs.harvard.edu/urn-3:hul.ebook:CHS_Nagy.Greek_Mythology_and_Poetics.1990.

Parry, Milman (1928). *L'épithète traditionnelle dans Homère*. Paris: Les Belles Lettres.

Parry, Milman (1971). *The Making of Homeric Verse: The Collected Papers of Milman Parry*. Oxford: Clarendon Press.

Plecháč, Petr, Klemens Bobenhausen, and Benjamin Hammerich (2018). "Versification and authorship attribution: A pilot study on Czech, German, Spanish, and English poetry". In: *Studia Metrica et Poetica* 5.2, pp. 29–54. DOI: 10.12697/smp.2018.5.2.02.

Sklaviadis, Sophia and James K. Tauber (2019). "Homeric Formulas and Meter". In: *Quantitative Approaches to Versification*. Ed. by Petr Plecháč, Barry Scherr, Tatyana Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. Prague: The Institute of Czech Literature of the Czech Academy of Sciences, pp. 237–244. URL: https://versologie.cz/conference2019/proceedings/sklaviadis-tauber.pdf.

Tarlinskaja, Marina (2015). "Ants Oras: Did He Know Russian "Formalists"?" In: *Studia Metrica et Poetica* 2.2, pp. 10–24. DOI: 10.12697/smp.2015.2.2.02.

Tver' yanovich, Kseniya Yur' evna (2008). *Poėtika Benedikta Livshica: Sistema stixa*. S.-Peterburg: SYMPOSIUM. URL: https://www.academia.edu/29763494/.

Tverianovich, Kseniia (2019). "Rhythm and Syntax in Aleksandr Sumarokov's Odes". In: *Quantitative Approaches to Versification*. Ed. by Petr Plecháč, Barry Scherr, Tatyana Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. Prague: The Institute of Czech Literature of the Czech Academy of Sciences, pp. 255–262. URL: http://versologie.cz/conference2019/proceedings/tverianovich.pdf.

https://doi.org/10.51305/ICL.CZ.9788076580336.08
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

Petrarch's Poetic Style from a Computational Perspective: A Digital Quantitative Approach to Italian Petrarchism

Jan Rohden

German Research Foundation (DFG), Germany
janrohden@gmail.com
0000-0001-7998-8629

Abstract

Few authors have shaped the history of European poetry as much as Petrarch (1304–1374). Based on its remarkable poetic style, Petrarch's most important Italian text, a collection of love poems called *Canzoniere* not only had an enormous impact on the poetry of his time but also became a model for centuries to come.

Scholars usually use the term "Petrarchism" to refer to Petrarch's influence on the literary landscape. Yet despite this common notion, there are still many competing approaches to defining Petrarchism. One reason for this may be the reliance of most studies of Petrarchism on a fairly small corpus of texts. While many scholars give a detailed account of Petrarch's influence on a single work or poet, only a few analyses of Petrarchism are based on a larger corpus. This may also help explain why there is still no comprehensive inventory of the stylistic or semantic elements that distinguish Petrarchism.

The goal of this essay is to take a first step towards creating such an inventory. To this end, digital methods, in particular stylometric and co-occurrence analyses, are applied to a corpus of 55 Italian poetry collections in order to determine the characteristic features of Italian Petrarchism.

1 Petrarch, the Founder of Petrarchism

Few authors have shaped the history of European poetry as much as Petrarch (1304–74). The reason for this lies in his varied literary works, which include a famous poetry collection in Italian entitled *Canzoniere*.[1] *Canzoniere* was widely read and became an important model for love poetry in Europe over ensuing

[1] Petrarch worked on this collection for many years. During that time, he changed its structure and title several times (Santagata 1992).

centuries. Over this time, many authors also tried to imitate Petrarch's poetic style in their own texts.[2]

One of the factors in *Canzoniere*'s success was the distinctive style of the collection, which consists of 366 poems about the unrequited love of the lyric's speaker ("I") for a married woman named Laura. This love, which does not end even after the woman's death, puts the speaker in highly contradictory states of mind that range from euphoria to deep sadness. This causes him suffering and creates a painful level of desire. The intermingling of suffering and desire results in a conflicted form of love for which literary researchers have proposed the term *dolendi voluptas* (Friedrich 1964, pp. 217–219). *Dolendi voluptas* is a good example of the dualities that characterise the form and content of Petrarch's collection. The poems, which take up motifs from Latin, Provençal and Italian literature, are arranged in a sequence that not only gives *Canzoniere* a temporal order, supported by references to specific dates and periods of the ecclesiastical year (Fornasiero 2001, pp. 59–89), but also has a narrative dimension. Petrarch's poetry collection becomes a memoir of its speaker's changing feelings for Laura, from the moment he falls in love with her at first sight to the rejection of his love described in the final poem of *Canzoniere* (Geyer 2009).

Petrarch's influence on his literary imitators is so far-reaching that a corresponding term has been established in the literary research: Petrarchism.

2 Towards a Definition of Petrarchism

Across multiple research contributions, literary scholars have identified elements they consider characteristic of Petrarch's poetic style in the texts of many European authors. In this way, they have shown his influence on European literature (Chines et al. 2006; Bernsen et al. 2011). Petrarch was particularly well received by Italian authors of the 16th century (Baldacci 1957), a trend not least illustrated by a series of poetry collections whose title, *Canzoniere* suggested their parallels with Petrarch's work.[3]

The wide-ranging impact of Petrarch's poetic approach on European literature was noted long ago by Leonard W. Forster (1963), who called Petrarchism "training in poetic diction". More recently, Klaus W. Hempfer (1987) and Gerhard Regn (1993) have applied a systems theory perspective to describe Petrarchism as a literary system. In contrast, Rainer Warning (1987) uses Mixail Baxtin's concept of dialogicity to interpret Petrarchism as a literary appropriation of competing poetic discourses. Meanwhile Michael Bernsen understands Petrarchism as a European *lingua franca* that was standardised and grammaticalised in the 16th century and became a means of communication at European level (Bernsen et al. 2011, pp. 15–30).

[2] Petrarch wrote another Italian-language text that had a large readership. This was *Triofi*, an allegorical poem in tercets. However, while *Triofi* deals with love, it is only one of the themes of this work, whose style was inspired by the genre of visionary literature. As such, most of the research on Petrarch's love poetry has focused on *Canzoniere*. The present article keeps with this tradition.

[3] Collections of poems entitled *Canzoniere* were published by Giusto de' Conti, Angelo Galli, Lorenzo de' Medici, Niccolò de' Rossi and Luigi Tansillo, among others.

These different approaches to defining Petrarchism illuminate not only the general reception of Petrarch's poetic style but also the different ways in which it has been adapted by other authors. Literary scholars have pointed out recurring elements of European literature that they believe reflect characteristics of Petrarch's poetry. Apart from *dolendi voluptas*, such elements include:

- Themes such as love as a form of war (Hoffmeister 1973, p. 25);
- Stylistic devices that express contrast (Friedrich 1964, p. 217);
- Glorification of the beloved in ethical and aesthetic terms (Regn 2013).

Despite these important findings, there is still no definitive typology of the distinctive elements of Petrarchan poetry. One cause of this may be the reliance of many studies of Petrarchism on a fairly small corpus of texts (see, for example, Pyritz 1963; Warning 1987; Morales Saravia 1998; Schiffer 2000). A list of these distinctive elements would, then, be valuable for the study of Petrarchism for at least two reasons: firstly, it would enable us to trace the development of specific Petrarchan features across the boundaries of individual works. Secondly, it could help us determine whether or not a text or author may be considered Petrarchan. An inventory of Petrarchan motifs is, thus, an important supplement to existing approaches to defining Petrarchism.

In order to establish such a list, we require a method that can extend beyond individual texts and be used to analyse a large corpus of poetry collections written by different authors. Quantitative digital approaches, including the contrastive analyses used in the context of stylometry, are well suited for such purposes.

3 Contrastive Stylometric Analyses of Poetry

Contrastive textual analysis is based on a principle of comparison: first, the textual corpus to be analysed is divided into two parts, a target and a comparison. The target is then compared with the comparison part to identify the target's overrepresented and therefore distinctive elements. To measure the distinctiveness of an element in one part, we may apply different methods or what are sometimes called distance measures.[4] Within the digital humanities, two tools are often used for contrastive textual analysis: `stylo` for R (Eder et al. 2016) and `pyzeta` for Python (Schöch 2017).

In recent years, the distance measure Zeta, originally introduced by John Burrows (2007) and mathematically described by Christof Schöch (2018), has been adopted widely in the context of contrastive analysis and proven helpful for various research contributions. Based on the original measure developed by Burrows, different Zeta variants have been presented. Using an evaluation procedure, the E-Humanities junior research group Computational Literary Genre Stylistics (CLiGS) showed that different variants of Zeta can produce quite divergent results in contrastive analyses of prose texts and drama (Schöch

[4] For a good introduction to contrastive text analysis, see Schöch 2018.

et al. 2018). Regarding poetry, however, only a few contrastive analyses and no studies on the effects of different Zeta variants exist to date.[5]

Nevertheless in the case of Petrarchan poetry, I would suggest that contrastive stylistic analysis can certainly contribute to the scholarship. In particular, it can help us to create a comprehensive typology of distinctive motifs.

4 Approach and Tools

If Petrarchism refers, as its name implies, to the particular impact of Petrarch's *Canzoniere* on the works of other authors, then there must be elements in Petrarch's text and in the writings of his poetic followers that distinguish them from non-Petrarchan works. In order to identify these distinctive elements in Italian poetry, it may be useful to complete a contrastive analysis in which a target group of Petrarchan poems is compared to a comparison group of poems that are undoubtedly non-Petrarchan. Texts that can be considered to be definitely non-Petrarchan would pre-date the appearance of Petrarch's *Canzoniere*.[6]

Given the lack of research on the impact of different Zeta variants on contrastive analyses of poetry, four contrastive text analyses were carried out. Each one used a different variant of Zeta.[7] The four resulting word lists were then compared[8] in order to create a list of overlapping distinctive elements. After a general review, the most striking elements were examined in more detail within their respective contexts by using collocation analyses. To this end, I used the tool `TXM` (Heiden et al. 2010).

5 Corpus

The corpus consisted of 55 different Italian collections of poems that dated mainly from the 14th, 15th or 16th centuries. It contained two parts: the target included 51 poetry collections (10,212,741 characters in total) in which researchers had been able to show Petrarch's influence.[9] The comparison part, on the other hand, included nearly all of the Italian love poetry that appeared before Petrarch's *Canzoniere* (four collections of poetry or 680,707 characters in total). Although the size of each part, thus, differed, the corpus reflected the actual literary-historical state of affairs: the number of collections of love

[5] Several works do, however, deal with the digital analysis of poetry (see, for example, Hoover 2008; Navarro-Colorado 2018a; Navarro-Colorado 2018b).

[6] Even before the appearance of Petrarch's *Canzoniere*, Italian love poetry was astonishingly varied. For an overview, see Friedrich 1964, pp. 1–156.

[7] These variants were "Craig's Zeta" and "Eder's Zeta" (both included in `stylo`) and "dd2" and "sd2". The latter two variants were developed by CLiGS and had achieved remarkable results in tests on prose and drama corpora (Schöch et al. 2018). To perform the contrastive analyses, "dd2" and "sd2" were also implemented with `stylo`. The code is available at http://dx.doi.org/10.20375/0000-000E-8B18-B.

[8] For the comparison, I used the tool `Multiple List Comparator` (MLC 2020).

[9] These collections were identified with the assistance of the research literature on Petrarchism, in particular Hempfer et al. 2005, pp. 24–76.

Part	Type of poetry	# of collections	# of characters
Target part	Petrarchan	51	10,212,741
Comparison part	Pre-Petrarchan	4	680,707
Total		*55*	*10,893,448*

Table 1: Overview of the corpus

poetry published since Petrarch's *Canzoniere* greatly exceeds the number of such collections published earlier. All but one of the texts in the corpus[10] were based on scholarly editions produced by a reputable editor and publisher. The digital versions of these editions are available in the *Biblioteca Italiana* (BI 2020) digital library. In all of these texts, the following items were removed: punctuation marks, page and line numbers, footnotes, endnotes, headings, title pages, forewords, epilogues and any other comments. Moreover, passages explicitly identified as prose texts (for example, dedications, comments and authors' introductions) were deleted, and all majuscules were converted to minuscules. Otherwise, the corpus was not harmonised any further at a linguistic level. Each collection of poems was saved as a plain text file with UTF-8 encoding.[11] This led to the following corpus (see Table 1).

6 Contrastive Analysis of Petrarchan and Pre-Petrarchan Poetry

Four contrastive analyses—i.e. one analysis for each of the four Zeta variants—were conducted on the Petrarchan target and the pre-Petrarchan comparison part. This resulted in four word lists.

The comparison of these four lists revealed 802 words that were preferred in the Petrarchan target.[12] Among these distinctive words, some semantic fields stood out, as can be seen in Table 2.

The semantic fields in the Petrarchan target part suggested essential aspects of Petrarchism.

6.1 The Motif of Sweetness as an Allusion to *Dolce Stil Novo*

The occurrence of words from the semantic field of sweetness among the preferred terms refers back to the pre-Petrarchan poetry of *Dolce stil novo*. The latter is also reflected in co-occurring expressions from various other semantic fields, for example, visual perception (see Section 6.7.1). Aside from expressions related to sweetness, other elements of *Dolce stil novo* can be identified in the Petrarchan target; this relates especially to the idealisation of the beloved. In the poetry of *Dolce stil novo*, the beloved is idealised based on a concept of

[10] The collection in question is Zaccagnini et al. 1915.

[11] A detailed list of all the poetry collections included in the corpus is available at http://dx.doi.org/10.20375/0000-000E-8B17-C.

[12] For the complete list of words, see http://dx.doi.org/10.20375/0000-000E-8B1A-9.

Semantic Field	Words
Bitterness	amaro
Character traits	
Cruelty	crudel
Grace	pietade, pietoso
Gracefulness	grazia
Gravity	grave, gravi
Honour	onor, onora, onorata, valor
Humility	umil
Ingenuity	ingegno
Misery	miser, misera, miseri, misero
Nobility	nobil
Piety	pio
Purity	casto, pure, puro
Value	virtute
Coldness	freddo, gelo, ghiaccio
Divinity and heavenliness	ciel, cielo, divin, divina, divino, santa, santi, santo
Fate and glory	corona, destino, eterna, fama, fortuna, lauro, orte
Feelings and emotional states	
Anger and hatred	ira, irato, odio
Desire	brama, bramo
Fear	teme, temer, timore
Indignation	sdegno
Grief and sorrow	affanni, fame, lacrime, lagrime, lamenti, piante, pianti, piant, tristi, tristo
Happiness	beata, beati, beato, lieto, lieta, serena, sereno
Hope	speme, sperar
Longing	desio, desir, desire, disiri
Love and affection	affetto, amori, amorose
Pain	dolor, duol, duole, duolo
Pride	fiera, fiere, fiero
Revenge	vendetta
Solitude	sol, sola, soli
Tiredness	stanca, stanco
Fire	arder, ardor, ardore, faville, fiamma, fiamme, infiamma

continued on the next page

Semantic Field	Words
Nature	acqua, acque, bosco, colli, erba, erbe, fiori, fiorito, fiume, fiumi, flora, fonte, fonti, legno, luna, mar, monte, monti, nido, nubi, notte, pietra, pioggia, rio, sasso, scoglio, selva, seme, venti, vento
Physicality and humanity	
Blood	sangue
Body parts and Physicality	braccio, chioma, corpo, fronte, grembo, man, mano, morso, occhio, passo, passi, petti, petto, piede, piedi, sen, seno, volto
Humanity	uman, umano, uom, uomini
Poetry and language	
Language	lingua, lingue
Meaning	senso
Muse	musa
Rhymes	rime
Speaking	parlar
Style	stil
Verses	versi
Words	parole
Reign	mercede, regni, regno, servi
Sensory perception	
Colours	bianco, bianche, biondo, bionde, colori, oro, verde
Darkness	tenebre
Light	luce, lucenti, lume, raggi, raggio, riluce, splende, splendor
Odour	odor
Sign	segni, segno
Sound	suon, suona, udite
taste	amaro, crud
Vision	mirar, mirando, vedo, vedrai, vedrá, veggia, veggio, visto
Voice	voce, voci
Soul	alma, anima, animo, aura, spirti, spirto
Sweetness	dolcezza, dolci, soave, soavi
Thoughts and memory	memoria, oblio, pensier, pensieri, pensiero
War	guerra, pace, preda, spada

Table 2: Notable semantic fields in the Petrarchan target

gentilezza (spiritual nobility), a particular aptitude for love that is based not on the speaker's origins or social status but on his soul alone. This idealisation based on *gentilezza* leads the lyric's speaker to ascribe angelic qualities to the beloved, as is illustrated in Section 6.5.[13]

6.2 Love Expressed as War

Expressions from the semantic field of war are also present among the Petrarchan target's distinctive words. This suggests that in Petrarchan poetry, love can be expressed as a form of war, as has previously been pointed out in the literary research (Hoffmeister 1973, p. 25).

6.3 Duality as a Stylistic Device

The co-occurrence of expressions from sharply contrasting semantic fields attests to the high number of dualities often found in Petrarchan poetry. The contrasts of coldness/fire and bitterness/sweetness are a clear indication of this; a co-occurrence analysis of words from the semantic field of coldness highlights the first example further. Among the words with the highest co-occurrence scores[14] are a considerable number from the semantic field of fire. The 18 words with the highest co-occurrence scores for "*fredd*.*"[15] (cold) are listed in Table 3.

6.4 *Dolendi Voluptas*

The presence of opposing forces is also clear in the semantic fields of love and pain, which occur together quite often. This points to the role of *dolendi voluptas*, a concept that scholars consider typical of Petrarchan poetry. This can be seen in the terms co-occurring with "*dolo*.*" (pain) in Table 4. These words, which include "*core*" (heart) and "*gioia*" (joy), at once capture the opposition between pain and pleasure and refer to love, which is symbolised by the heart.

6.5 Idealisation of the Beloved as a Superhuman Being

The idealisation of the beloved, another typical element of *Dolce stil novo*, is illustrated by the co-occurrence of words from the semantic fields of divinity and heavenliness on the one hand and humanity on the other. Table 5 shows words co-occurring with "*divin*.*" (the divine). Here divinity is often detected among humanity through the stylisation of the beloved.

[13] For an overview of *Dolce stil novo*, see Pirovano 2014.

[14] The co-occurrence score indicates the degree of specificity of the co-occurrence of two words. See Lafon 1980 for the mathematical background.

[15] This is a search query for words that start with "fredd" and can have any number of characters. The dot (".") stands for any character (except the space character); the asterisk ("*") indicates that the directly preceding character can occur any number of times (or even not at all). For a clear introduction to the functions and search syntax of TXM, see Schöch 2020.

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	caldo	warm	392	42	38	2.9
2	neve	snow	400	38	32	3.1
3	ghiaccio	ice	296	29	25	2.5
4	pietra	stone	253	20	16	1
5	calda	warm	62	12	14	3.2
6	verno	winter	320	18	12	1.8
7	gelo	frost	338	18	11	2.9
8	smalto	enamel	64	10	11	1.5
9	ardente	ardent	577	22	11	3.7
10	core	heart	2162	42	10	3
11	e	and	86639	635	10	4
12	sasso	stone	371	17	10	2.1
13	foco	fire	1705	35	9	3.2
14	marmo	marble	213	13	9	1.5
15	borea	Bores[a]	84	9	8	2.3
16	nevi	snows	59	8	8	3.9
17	duro	hard	597	19	8	2.6
18	austro	Austro[b]	52	7	7	2.9

Table 3: The 18 terms with the highest co-occurrence scores for "fredd.*". Expressions from the semantic field of fire are marked in red; those from the semantic field of coldness are marked in blue

[a] The god of the north wind and bringer of winter in Greek mythology.
[b] God of southern winds and/or of the South in Roman mythology.

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	mio	my	10666	566	104	2.4
2	pianto	crying	1492	123	41	3.6
3	il	the	32222	829	20	3.6
4	core	heart	2162	110	19	5.4
5	morte	death	3924	151	15	4.6
6	lagrime	tears	318	34	15	4.1
7	sento	I hear	753	52	14	3.5
8	gioia	joy	522	38	11	3.7
9	pietà	pity	1352	66	11	4.4
10	tanto	much	3914	137	11	3.7
11	l'aspro	the harsh	75	15	11	0.5
12	pena	penalty	969	50	9	3.7
13	lassa	leave	270	24	9	4.5
14	mi	me	11226	300	9	4.2
15	magior	greater	198	20	8	2.7
16	gran	big	3671	122	8	2.1
17	pene	penalties	491	32	8	4.3
18	pianti	crying	344	26	8	2.6

Table 4: The 18 terms with the highest co-occurrence scores for "*dolo.**". Expressions from the semantic field of pain are marked in red; those from the semantic field of happiness are marked in blue

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	spirto	spirit	817	47	16	2
2	bellezze	beauties	235	23	13	2.8
3	alte	high	250	23	12	1.5
4	virtù	virtue	1507	55	10	2.9
5	luce	light	1501	53	10	3.2
6	beltà	beauty	756	35	9	1.1
7	luci	lights	441	26	9	1.8
8	mortal	mortal	903	37	8	3.7
9	contemplando	contemplating	18	7	8	2.7
10	celeste	heavenly	738	32	8	4.8
11	del	of	10388	200	7	3.9
12	angelica	angelic	96	11	7	2.1
13	umane	human	81	10	7	3.3
14	nel	in	5090	110	6	3.7
15	concetti	concepts	51	8	6	3.6
16	l'eterno	the eternal	119	11	6	4
17	verbo	verb	25	6	6	0.5
18	lume	light	1257	39	6	2.2

Table 5: The 18 terms with the highest co-occurrence scores for "*divin.**". Expressions from the semantic field of humanity are marked in red; those from the semantic field of divinity and heavenliness are marked in blue

Among these co-occurring terms, "*angelica*" (angelic) is particularly interesting because it evokes the concept of "*donna angelicata*" (angelic woman). The latter is relevant not only to *Dolce stil novo* poetry but also to Dante's *Vita Nova*.

6.6 Subjectivity as a Key Theme in Petrarchan Poetry

While the observations above broadly confirm issues already noted in the literary research, contrastive analyses reveal other elements of Petrarchism that have received less attention. Of particular significance are the four semantic fields with the highest numbers of distinctive words: character traits; feelings and emotional states; nature; and sensory perception. The first two fields have a highly subjective focus: the speaker's preoccupation with the beloved and her character traits and features that has ultimately led him to declare his feelings. This leads to a detailed characterisation of the beloved and, at the same time, reveals the great range of emotional states of the lyric's "I". The encounter between the "I" and the beloved highlights one of the characteristic dichotomies of Petrarchan poetry. This is expressed in different ways, as the large number of expressions from the semantic fields "character traits" and "feelings and emotional states" demonstrates. As the contrastive analysis shows, the detailed exploration of the soul of the "I" and his various emotional states which is achieved in Petrarch's *Canzoniere* (Geyer 2009) proves to be characteristic of Italian Petrarchan poetry more generally.

6.7 Sensory Perception as a Key Theme in Petrarchan Poetry

In the corpus studied, the first-person speaker's exploration of his own subjectivity takes place especially in the realm of sensory perception. In this context, visual perception has an important role, as is particularly striking in two semantic fields: the perception of nature and the perception of the human body. Although Italian poetry employed visual perception as a means of approaching the beloved even before Petrarch (Zeiner 2006), both the importance and complexity of this perception reach a new level in Petrarchan poetry.

6.7.1 Visual Perception as the Preferred Approach to the Beloved

In this regard, the eyes have special significance as both a symbol of visual perception and a characteristic feature of the beloved. This can be seen from the many expressions in which composites of "*occh*.*" (eye) are among the most common co-occurring terms. These patterns characterise visual experience, and thus, sensory perception in general as an important aspect of love. They are exemplified in the 18 terms with the highest co-occurrence scores for "*amor*.*" (love) (Table 6).

The eyes are also related to beauty and female beauty in particular. This is made clear by the most common co-occurring terms for "*bell*.*" (beauty) in Table 7.

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	mi	me	11226	1334	75	3.8
2	core	heart	2162	317	33	4.9
3	strali	arrows	236	72	26	3.4
4	d	of	640	125	24	2.6
5	cor	heart	5018	544	22	4.9
6	pargoletti	children	34	25	21	0
7	strale	arrow	210	57	18	3.4
8	l'arco	the bow	243	60	17	3.4
9	m'ha	(he/she/it) has me	447	85	16	2.2
10	et	and	8790	826	15	4.4
11	sforza	strain	229	55	15	3.8
12	m	me	433	81	14	3.6
13	me	me	8531	790	13	4.2
14	begli	beautiful	739	111	13	4.7
15	occhi	eyes	3684	380	12	4.6
16	ove	where	2262	254	12	3.7
17	mio	my	10666	946	12	4.6
18	dolce	sweet	3926	397	12	4.5

Table 6: The 18 terms with the highest co-occurrence scores for "*amor*.*". Expressions alluding to the eyes are marked in red; those from the semantic field of love are marked in blue

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	donna	woman	2133	365	96	2.5
2	più	more	14784	1262	82	2.7
3	la	the	27102	1866	51	3.2
4	sì	that	9034	727	39	3
5	vaga	vague	491	89	26	3.2
6	e	and	86639	4812	24	3.8
7	ninfa	nymph	177	49	23	1.9
8	men	less	1567	170	21	2.4
9	man	hand	1586	169	20	2.6
10	casta	chaste	126	38	19	1.9
11	l'altre	the others	411	70	19	2.9
12	saggia	wise	119	34	16	2.5
13	imago	image	120	34	16	1.8
14	honesta	honest	83	28	16	1.5
15	le	the	13962	884	15	3.9
16	onesta	honest	81	27	15	1.7
17	occhi	eyes	3684	288	14	3.5
18	aurora	aurora	117	29	12	18

Table 7: The 18 terms with the highest co-occurrence scores for "*bell.**". Expressions alluding to the eyes are marked in red; those referring to femininity are marked in blue

The eyes are connected with beauty in two ways: they are sensory organs for perceiving beauty, and they are body parts which themselves contribute to the beauty of the beloved.

Moreover, the eyes frequently appear together with expressions from the semantic field of sweetness (for example "*dolc*.*" or "*soav*.*"), which can be understood as a reference to *Dolce stil novo* (see Section 6.1).

6.7.2 Visual Perception as the Basis for the Idealisation of the Beloved

Furthermore, the eyes appear in the context of different parts of the body, including the hair ("*chiom*.*"), forehead ("*front*.*"), feet ("*pied*.*"), breasts ("*seno*") and face ("*volt*.*").[16] As such, the eyes are not only relevant as organs that are physical objects but they are also the preferred way of perceiving the body of the beloved. This visual perception of the physical beauty of the beloved is the starting point for her idealisation, which is expressed through different semantic fields. The latter include, on the one hand, many terms from the semantic field of light, for example, "*luc*.*" (light), "*lum*.*" (light), "*ragg*.*" (ray), "*riluc*.*" (shining), "*splend*.*" (splendid).[17] On the other, they extend to expressions from the semantic field of fire, for example, "*ard*.*" (burn) and "*fiamm*.*" (flame).[18] Based on this idealisation, the beloved attains the divine qualities typically celebrated in the poetry of *Dolce stil novo*. This can be seen from the co-occurring terms for "*uman*.*" (human) listed in Table 8.

This visually perceived semblance ("*sembiante*", "*forma*") reveals the beloved to be a creature who is equally human ("*petti*", "*corpo*") and divine ("*angelo*", "*divin*").[19]

The analysed corpus, thus, highlights three essential aspects of the depiction of love in Italian Petrarchism: the prevalence of visual perception in general; the use of idealising perception in the style of *gentilezza* in *Dolce stil novo*; and the resulting conception of the woman as both a human and divine creature. The common basis of all three aspects is vision, as evidenced by the co-occurring terms for "*visio*.*" (vision) listed in Table 9. These terms cover expressions from a range of semantic fields that include the following: divinity and heavenliness, humanity and sweetness.

[16] The associated tables can be accessed at https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD3-8, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD2-9, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD7-4, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8ADA-1, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8ADC-F.

[17] The associated tables can be accessed at https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD5-6, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD6-5, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD8-3, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD9-2, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8ADB-0.

[18] The associated tables can be accessed at https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD1-A, https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000E-8AD4-7.

[19] Similarly, the co-occurring terms for "*divin*.*" also refer to the divine and human dimensions of the beloved (see Section 6.5.)

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	sembiante	semblance	238	21	15	0.2
2	petti	chests/breasts	118	14	12	1
3	menti	minds	140	13	9	1.2
4	ingegno	ingenuity	542	23	9	0.1
5	angelo	angel	18	6	8	2.2
6	divin	divine	376	17	7	4
7	forma	form	378	17	7	1.2
8	seme	seed	221	12	6	0.8
9	ingegni	ingenuities	144	10	6	0.4
10	corpo	body	845	23	6	1
11	velo	veil	557	18	6	2.4
12	spezie	type	10	4	6	0
13	d'ogni	of every	1167	27	5	2.1
14	potere	power	46	6	5	4.3
15	ogni	every	3912	60	5	2.4
16	questo	this	3166	51	5	3.6
17	in	in	25194	265	5	4.2
18	l'occhio	the eye	187	10	5	2.1

Table 8: The 18 terms with the highest co-occurrence scores for "*uman.**". Expressions from the semantic field of humanity are marked in red; those from the semantic field of divinity and heavenliness are marked in blue

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	in	in	25194	503	20	3
2	dolce	sweet	3926	115	14	2.5
3	sua	his/her	3182	94	12	2.1
4	humana	human	82	14	11	1.5
5	l'ho	I have him/her/it	102	13	9	2
6	ho	have	1526	51	8	0.7
7	contrista	sadden	18	7	8	4
8	angelica	angelic	96	12	8	0.2
9	trista	sad	228	17	8	4.5
10	altera	haughty	320	19	7	0.7
11	occhi	eyes	3684	86	6	5.1
12	umana	human	141	12	6	1.2
13	altrove	elsewhere	271	16	6	2.9
14	mirar	gaze at	309	17	6	4.8
15	turbata	troubled	28	6	5	0.8
16	bella	beautiful	2148	54	5	3.4
17	amata	loved	99	9	5	0
18	lieta	happy	620	23	5	3.3

Table 9: The 18 terms with the highest co-occurrence scores for "*visio.**". Expressions from the semantic field of divinity and heavenliness are marked in blue; those from the semantic field of humanity are in red; those from the semantic field of sweetness are in green; those from the semantic field of vision are in purple

6.7.3 Visual Perception as the Basis for Petrarchan Poetry

The significance of visual perception for Petrarchism is also evident from the expressions co-occurring with one of poetry's most characteristic formal elements: verse ("*vers*.*"). These co-occurrences suggest important motifs in Petrarchan love poetry, as Table 10 demonstrates.

The 18 terms with the highest co-occurrence scores reflect essential elements of Petrarchan poetry, namely subjectivity ("*miei*", "*mei*"), emotionality ("*lagrime*", "*pianto*"), physicality ("*sangue*", "*vena*"), visual perception ("*occhi*"), allusions to *Dolce stil novo* ("*dolci*") and nature ("*fiume*", "*fiori*"). The reference to prose ("*prosa*", "*prose*") is also noteworthy. On the one hand, it highlights the prose dimension while on the other, it creates a contrast with the language of verse ("*versi*"). In this way, it also illustrates the stylistic device of duality (see Section 6.3).

6.7.4 Nature as a Key Theme in Petrarchan Poetry

Significantly, individual expressions from the semantic field of nature do not usually occur in isolation in the corpus studied. Rather they are accompanied by other expressions from the same semantic field. This suggests that Petrarchan poetry does not use nature as a mere point of reference but instead sketches natural panoramas by combining different expressions from this semantic field. This can be seen, for example, from the terms co-occurring with "*bosc*.*" (forest) listed in Table 11.

6.8 Unassignable Co-Occurrences

In addition to clearly assignable co-occurrences, Tables 3 to 11 also include expressions that cannot be easily attributed to a specific semantic field. A comparison of these words reveals two patterns. Firstly, among the expressions that cannot be clearly assigned are words that lack any lexical meaning and instead have a grammatical function. These include articles,[20] conjunctions,[21] demonstrative pronouns[22] and prepositions,[23] that is, basically function words whose significance for stylometric analyses has already been demonstrated (Argamon et al. 2005; Kestemont 2014). Secondly, the tables feature quite a large number of object pronouns and possessive pronouns in the first-person singular.[24] These pronouns reflect at a grammatical level the intense involvement of the lyrical "I" with the beloved, which indirectly leads to a preoccupation with his own emotional state. As such, they highlight the highly subjective dimension of Petrarchan love poetry, as outlined in Section 6.6. Notwithstanding these

[20] These are "il" in Table 4, "l'"("eterno") in Table 5, "la" and "le" in Table 7 and "i" in Tables 10 and 11.
[21] These are "e" in Tables 3, 7 and 11 and "et" in Table 6.
[22] These are "questo" in Table 8 and "questi" in Table 11.
[23] These are as "del" and "nel" in Table 5, "d" in Table 6, "d'"("ogni") in Table 8 and "in" in Tables 8 and 9.
[24] These are "mio" and "mi" in Tables 4 and 6, "m'"("ha"), "m" and "me" in Table 6, "l'"("ho") in Table 9 and "miei" and "mei" in Table 10.

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	rime	rhymes	436	50	33	2.2
2	miei	my	2057	80	20	1.7
3	i	the	11351	237	19	2.9
4	lagrime	tears	318	29	17	2.5
5	prosa	prose	27	12	16	2.3
6	pianto	crying	1492	57	14	3.3
7	lacrime	tears	233	21	12	2.2
8	sangue	blood	780	37	12	3.3
9	versi	verses	468	27	11	5.2
10	prose	proses	37	10	11	1.8
11	tersi	clear	19	8	10	4.8
12	mei	my	572	28	9	1.3
13	udite	hear	61	10	8	4.1
14	fiume	river	458	22	7	4
15	occhi	eyes	3684	76	6	3.7
16	fiori	flowers	551	22	6	3.9
17	vena	vein	141	11	6	3.3
18	dolci	sweet	1121	33	6	2.1

Table 10: The 18 terms with the highest co-occurrence scores for “*vers.**”

Number	Co-occurring term	Translation	Frequency	Co-frequency	Score	Mean distance
1	i	the	11351	185	41	2.6
2	campagne	countryside	131	28	36	2.2
3	selve	forests	259	27	26	2.3
4	monti	mountains	419	30	24	3
5	folto	dense	37	14	22	0.1
6	folti	dense	28	13	22	0.3
7	colli	hills	323	26	22	2.7
8	prati	lawn	102	18	21	2.2
9	ombrosi	shady	49	14	20	0.6
10	questi	these	1076	39	20	2.3
11	verdi	green	284	21	17	3.1
12	ombroso	shady	50	12	16	1.5
13	sassi	stones	203	18	16	3.5
14	fiumi	rivers	315	20	15	3.7
15	fonti	fountains	203	17	15	3.4
16	e	and	86639	602	14	3.9
17	dumi	thorns	33	9	13	3
18	fronde	leaves	390	19	12	3

Table 11: The 18 terms with the highest co-occurrence scores for “*bosc*.*”. Expressions from the semantic field of nature are marked in blue

function words, the majority of the co-occurring terms listed in Tables 3 to 11 have a lexical meaning. In other words, at least the items examined with the highest co-occurrence scores are comparatively highly interpretable. From a literary studies perspective, this is encouraging as it suggests that a co-occurrence analysis may be a useful addition to contrastive analyses.[25]

6.9 Conclusion

The aim of the present study was to identify distinctive motifs in the poetry of Italian Petrarchism through a combination of contrastive stylometric analyses and co-occurrence analyses of a corpus of Italian poetry. In this way, I sought to contribute to a comprehensive typology of Petrarchan motifs. The chosen methodology enabled me to uncover distinctive elements in the Petrarchan target on whose basis characteristic semantic fields could be identified.

Based on these semantic fields and their respective co-occurring terms, I found evidence of not only some aspects of Petrarchism already observed in the literary research but also elements that had hitherto been less prominent in that discourse. Of particular note was the prevalence of visual perception, which proved to be fundamental to three themes: the conception of love by the lyric's first-person speaker; the depth of his subjectivity; and the notion of poetry itself. Vision may, thus, be considered an important element in a considerable number of Italian Petrarchan poetry collections.

Nevertheless, the present study can only serve as a first step in creating a comprehensive inventory of Petrarchan motifs. To gain a more complete picture of Petrarchism, it may be helpful to conduct similar analyses of Petrarchan corpora in other languages such as German, English and Spanish. After all, if there is one point of consensus in the literary research on Petrarchism, it is Petrarch's far-ranging international reception. In exploring this topic, digital methods will certainly be of great use.

References

Argamon, Shlomo and Shlomo Levitan (2005). "Measuring the usefulness of function words for authorship attribution". In: *Proceedings of the 2005 ACH/ALLC Conference*. Victoria: University of Victoria, pp. 33–34. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.6935&rep=rep1&type=pdf (visited on 11/15/2020).

Baldacci, Luigi (1957). *Il Petrarchismo Italiano Nel Cinquecento*. Milano, Napoli: Ricciardi.

Bernsen, Michael and Bernhard Huss, eds. (2011). *Der Petrarkismus – ein europäischer Gründungsmythos*. Göttingen: V&R. URL: http://hdl.handle.net/20.500.11811/544.

BI (2020). *Biblioteca Italiana*. URL: http://www.bibliotecaitaliana.it/ (visited on 11/15/2020).

[25] Schöch et al. (2018) illustrate the comparatively high interpretability of word lists generated on the basis of Zeta.

Burrows, John (2007). “All the Way Through: Testing for Authorship in Different Frequency Strata”. In: *Literary and Linguistic Computing* 22.1, pp. 27–47. DOI: 10.1093/llc/fqi067.

Chines, Loredana, Floriana Calitti, and Roberto Gigliucci, eds. (2006). *Il petrarchismo: Un modello di poesia per l’Europa*. Roma: Bulzoni.

Eder, Maciej, Jan Rybicki, and Mike Kestemont (2016). “Stylometry with R: A Package for Computational Text Analysis”. In: *The R Journal* 8.1, pp. 107–121. DOI: 10.32614/RJ-2016-007.

Fornasiero, Serena (2001). *Petrarca: guida al Canzoniere*. Roma: Carocci.

Forster, Leonard W. (1963). “European Petrarchism as Training in Poetic Diction”. In: *Italian Studies* 18.1, pp. 19–32. DOI: 10.1179/its.1963.18.1.19.

Friedrich, Hugo (1964). *Epochen der italienischen Lyrik*. Frankfurt am Main: Klostermann.

Geyer, Paul (2009). “Petrarcas Canzoniere als Bewusstseinsroman”. In: *Petrarca und die Herausbildung des modernen Subjekts*. Ed. by Paul Geyer and Kerstin Thorwarth. Göttingen: V&R, pp. 109–156.

Heiden, Serge, Jean-Philippe Magué, and Bénédicte Pincemin (2010). “TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement”. In: *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*. Vol. 2. Milano: LED, pp. 1021–1032. URL: https://halshs.archives-ouvertes.fr/halshs-00549779/fr/ (visited on 11/15/2020).

Hempfer, Klaus W. (1987). “Probleme der Bestimmung des Petrarkismus. Überlegungen zum Forschungsstand”. In: *Die Pluralität der Welten: Aspekte der Renaissance in der Romania*. Ed. by Wolf-Dieter Stempel and Karlheinz Stierle. München: Fink, pp. 253–277.

Hempfer, Klaus W., Gerhard Regn, and Sunita Scheffel, eds. (2005). *Petrarkismus-Bibliographie, 1972-2000*. Stuttgart: Steiner.

Hoffmeister, Gerhart (1973). *Petrarkistische Lyrik*. Stuttgart: Metzler.

Hoover, David L. (2008). “Searching for style in modern American poetry”. In: *Directions in Empirical Literary Studies. Linguistic Approaches to Literature, 5*. Ed. by Sonia Zyngier, Marisa Bortolussi, Anna Chesnokova, and Jan Auracher. Amsterdam: John Benjamins Publishing Company, pp. 211–227. DOI: 10.1075/lal.5.18hoo.

Kestemont, Mike (2014). “Function Words in Authorship Attribution. From Black Magic to Theory?” In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg: Association for Computational Linguistics, pp. 59–66. DOI: 10.3115/v1/W14-0908.

Lafon, Pierre (1980). “Sur la variabilité de la fréquence des formes dans un corpus”. In: *Mots* 1, pp. 127–165. DOI: 10.3406/mots.1980.1008.

MLC (2020). *Multiple List Comparator*. URL: http://www.molbiotools.com/listcompare.html (visited on 11/15/2020).

Morales Saravia, José (1998). “Vanitas y petrarquismo en el soneto XXIII de Garcilaso de la Vega”. In: *Iberoromania* 47, pp. 47–71. DOI: 10.1515/iber.1998.1998.47.47.

Navarro-Colorado, Borja (2018a). “A metrical scansion system for fixed-metre Spanish poetry”. In: *Digital Scholarship in the Humanities* 33.1, pp. 112–127. DOI: 10.1093/llc/fqx009.

Navarro-Colorado, Borja (2018b). "On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry". In: *Frontiers in Digital Humanities* 5, pp. 1–12. DOI: 10.3389/fdigh.2018.00015.

Pirovano, Donato (2014). *Il Dolce stil novo*. Roma: Salerno Editrice.

Pyritz, Hans (1963). *Paul Flemings Liebeslyrik: Zur Geschichte des Petrarkismus*. Göttingen: V&R.

Regn, Gerhard (1993). "Systemunterminierung und Systemtransgression. Zur Petrarkismus-Problematik in Marinos Rime amorose (1602)". In: *Der Petrarkistische Diskurs: Spielräume und Grenzen*. Ed. by Klaus W. Hempfer and Regn Gerhard. Stuttgart: Steiner, pp. 255–281.

Regn, Gerhard (2013). "Petrarkismus". In: *Historisches Wörterbuch Der Rhetorik Online*. Ed. by Gert Ueding. Berlin: De Gruyter. DOI: 10.1515/hwro.

Santagata, Marco (1992). *I frammenti dell'anima storia e racconto nel Canzoniere di Petrarca*. Bologna: Il Mulino.

Schiffer, James (2000). "Shakespeare's Petrarchism". In: *Shakespeare's sonnets: Critical essays*. Ed. by James Schiffer. New York: Garland, pp. 163–183.

Schöch, Christof (2017). *Pyzeta. Python implementation of the Zeta score for contrastive text analysis*. Version 0.3.0. URL: https://github.com/cligs/pyzeta (visited on 11/15/2020).

Schöch, Christof (2018). "Zeta für die kontrastive Analyse literarischer Texte". In: *Quantitative Ansätze in den Literatur- und Geisteswissenschaften*. Ed. by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht. Berlin, Boston: De Gruyter, pp. 77–94. DOI: 10.1515/9783110523300-004.

Schöch, Christof (2020). *TXM-Tutorial. Figurenbeschreibungsprojekt, Zoom, Juli 2020*. URL: https://christofs.github.io/txm-tutorial/#/ (visited on 11/15/2020).

Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho (2018). "Burrows' Zeta: Exploring and Evaluating Variants and Parameter." In: *ADHO 2018. Book of Abstracts*. URL: https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/ (visited on 11/15/2020).

Warning, Rainer (1987). "Petrarkistische Dialogizität am Beispiel Ronsards". In: *Die Pluralität der Welten: Aspekte der Renaissance in der Romania*. Ed. by Wolf-Dieter Stempel and Karlheinz Stierle. München: Fink, pp. 327–358.

Zaccagnini, Guido and Amos Parducci (1915). *Rimatori siculo-toscani del dugento*. Bari: Laterza. URL: https://www.liberliber.it/online/autori/autori-r/rimatori-siculo-toscani-del-dugento/rimatori-siculo-toscani-del-dugento/ (visited on 11/15/2020).

Zeiner, Monika (2006). *Der Blick der Liebenden und das Auge des Geistes. Die Bedeutung der Melancholie für den Diskurswandel in der Scuola Siciliana und im Dolce Stil Nuovo*. Heidelberg: Winter.

https://doi.org/10.51305/ICL.CZ.9788076580336.09
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

Metric Variation in the Finnic Runosong Tradition: A Rough Computational Analysis of the Multilingual Corpus

Mari Sarv
Estonian Literary Museum, Estonia
mari@folklore.ee
0000-0001-5309-2357

Kati Kallio
Finnish Literature Society, Finland
kati.kallio@finlit.fi
0000-0002-3673-1409

Maciej Janicki
HELDIG, University of Helsinki, Finland
maciej.janicki@helsinki.fi
0000-0003-3981-8021

Eetu Mäkelä
HELDIG, University of Helsinki, Finland
eetu.makela@helsinki.fi
0000-0002-8366-8414

Abstract

This article represents a first step in the corpus-based study of metric variation in Finnic runosong, a poetic tradition shared by several Finnic peoples and documented extensively in the 19th and 20th centuries. Runosong metre has generally been assumed to be a syllabic tetrametric trochee with specific rules about the placement of stressed syllables according to their quantity: long stressed syllables occupy the strong positions in the trochaic schema while short stressed syllables appear in the weak positions. Recent studies by Mari Sarv (2008, 2015, 2019) of Estonian runosong metre have shown, however, that due to linguistic changes, it has gradually lost its quantitative properties and acquired the features of accentual metre.

Using computational methods, this study aims to give a preliminary overview of the extent of metric variation on the quantitative-accentual scale across the entire Finnic runosong area. After an approximate syllabification, we apply two separate indirect methods for estimating variation. These appear to generate coherent results: quantitative runosong metre dominates in the north-east and has gradually been replaced by accentual runosong metre towards the south-west. Subsequent studies should verify these results through more precise and detailed investigations.

1 Introduction

The present paper forms the first part of a larger study which aims to describe the metric features and variations in the corpus of a poetic tradition that is documented across almost the whole Finnic area. The entire corpus comprises

over 240,000 runosong texts in Estonian, Finnish, Karelian, Ingrian (Izhorian) and Votic that were mainly recorded during the 19th and 20th centuries. We seek to answer such questions as, What are the characteristics of runosong metre? What features appear across the corpus? And how do they relate to poetic genres and regional singing cultures? Statistical analysis of the corpus may allow us to uncover trends and connections not detected in previous research, which was limited to selected regions, sub-genres and singers. The current article offers a preliminary estimation of the extent of metric variation within the corpus based on an approximate syllabification.

Oral poetry only exists in fleeting performances and memory structures. The texts that we have are more or less accurate transcriptions of individual performances. Neither the poems nor any individual verses have a single original or authoritative form; rather they developed from performance to performance and from singer to singer. Typically singers of these works have a number of options for varying the performance of their poems and lyrics. Albert Lord (1960) notes that in many oral traditions, singers tend not to use poems learned verbatim by heart; instead they create a poem during its performance based on their knowledge of traditional storylines, formulae (metrically and poetically motivated collocations), poetic language and the needs of each situation.

The creation and transmission process peculiar to folklore inevitably produces variation across different aspects of folkloric expression. This also applies to the corpus under investigation and needs to be kept in mind with respect to this research. The properties of the texts vary based on the regional peculiarities of their content, performance, usage, poetics and language. Their metre also varies both as a standalone poetic feature and as a result of other layers of variation. Regional and genre-based variations in rhythmic melodic structure, for example, place different restrictions on metric structure. Given all of its linguistic and poetic variation, the current corpus, thus, poses serious challenges for a large-scale computational analysis. In addition, there are biases in the geographical and typological distribution of this material that stem from regional traditions and the collection history. As such, any quantitative results require careful interpretation and contextualisation.

2 The Finnic Runosong Tradition and Its Metric Properties

Most Finnic languages (see Figure 1) share a specific poetic-musical tradition that is characterised by stichic diction (no stanzas or rhymes), interdependent use of alliteration and parallelism and a syllabic metre with a trochaic core.[1] In the English-language scholarship, this tradition has various names of which the most common are Finnic oral poetry, Kalevalaic poetry and runosong although researchers have also noted the shortcomings of these various terms (Kallio et al. 2017). What is distinctive about this poetic form is its use across a wide variety of sung genres (e.g., epic, lyric and ritual songs, occasional songs, charms

[1] This tradition is, however, unknown among Livonians and Vepsians. Audio recordings of traditional runosong together with original texts and English translations can be accessed online; see, for example, the samples in Tampere et al. (2016) and Oras (2015).

Figure 1: The main dialects of the Finnic languages at the beginning of the 20th century (Grünthal forthcoming)

and improvisations) and also to some degree in colloquial speech and spoken word genres such as proverbs, sayings and riddles (Krikmann 1997; Kuusi 1994; Tampere et al. 2017). Alongside this prevalent poetic system, other less common poetic structures in various genres (e.g., laments and children's songs) also evolved in this region. Sometimes, as in the case of bagpipe songs, they were influenced by other cultures. Eventually, rhymed stanzaic songs with Russian, German and Swedish influences prevailed in the popular poetic expression of the second half of 19th century. Today these traditions continue in various forms.

The earliest preserved examples of runosongs were transcribed in the 16th and 17th centuries, and the majority of related texts were recorded during the 19th and early 20th centuries. The beginning of this folklore collection was tied to the awakening of national consciousness in Finland and Estonia, then both part of the Russian empire – and the need for both these small states to create a national culture and history (see, e.g., Anttonen 2005; Tarkka et al. 2018). The nationalist call for collection generated extremely rich and voluminous

collections. From 1904 onwards, songs were also recorded. The first recordings were made on phonographs but later other audio recording techniques evolved and were used (see, e.g., Oras 2008; Huttu-Hiltunen 2008).

The metre of Finnic runosongs varies by region but is generally thought to derive from a common original form: this is a syllabic metre with a trochaic core that has four stresses (and feet) and follows specific rules about the quantity of stressed syllables (in Finnic languages, the stress falls, as a rule, on the first syllable of a word). This metre is believed to have been established during the Proto-Finnic linguistic phase when the prosodic system of the language proved itself suitable for the evolution of syllabic metre (Korhonen 1994). This was in the first millennium BCE, a period when according to contemporary archaeogenetic research, the Finno-Ugric genomic component arrived in the area around the Baltic Sea (Saag et al. 2019).[2] The poetic system of this tradition is thought to have developed from (1) earlier poetic forms;[3]; (2) changes in the prosodic structure of the language; and/or (3) contact between different cultures in this era when Proto-Finnic had not yet divided into many separate languages (Frog 2019; Korhonen 1994; Sarv 2019).

Runosong metre was first described vaguely by 17th-century scholars (Sarajas 1956). In his work *De poesi Fennica*, Henrik Gabriel Porthan, the most notable early researcher of runosong, mentioned the specific role of trisyllabic words in the form's rhythm. Ultimately, however, he echoed his predecessors' view that syllabic quantity did not play a role in runosong metre (Porthan 1983, p. 41). The quantitative basis of this metre was discovered by Arvid Genetz, who in 1881 formulated the "quantity rules" of *Kalevala* metre, relying mainly on Karelian runosongs (Genetz 1884 [1881]).[4] These rules describe the placement of stressed syllables according to their quantity: except in the case of the first two line positions, long stressed syllables occupy the strong positions in the trochaic schema while short stressed syllables fill the weak positions. These rules, thus, describe and explain the very essence of the peculiar rhythmic alterations of runosongs, something that researchers had been pondering for decades.

Below we provide an example. This is a wedding song from northern Estonia in quantitative runosong metre:

Pei-u-ke-ne, **poi**-si-ke-ne, (44)	Bridegroom, dear boy,
ei **mi**-na **ra**-ha-ta **lau**-la, (1232)	I'm not singing without payment,
kul-la-ta keelt en **ku**-lu-ta, (3113)	not using my tongue without gold,
keelt en **pek**-sa **pen**-nin-gi-ta! (1124)	not beating my tongue without a penny!
Mi-na **lõ(p)**-pe-tan **lu**-gu-da, (233)	I'm going to make up my song,
si-na **kae**-va **kar**-man-tu-da, (224)	you'd better dig into your pocket,
ra-ha **tas**-ku-da **ta**-o-ta! (233)	reach for your wallet!

[2] Other scholars have made similar assumptions about the general period of runosong's genesis based on different information; see, for example, Leino (1986, p. 140), Rüütel (1998) and Frog (2019).

[3] Speculation about the common Uralic or Finno-Ugric poetic and metric forms that might have preceded runosong falls outside the scope of this paper. For more information on this topic, however, see, for example, Korhonen (1994), Helimski (1998) and Frog (2019).

[4] Elias Lönnrot had also used Karelian runosong when composing his epic, *Kalevala*.

These lyrics have been syllabified while we have marked the stressed syllable/s in the polysyllabic words in metrically relevant positions in bold. After each line of verse, its syllabic structure is also given with numbers that indicate the syllabic composition of the line's wording. The line structure 233, for example, consists of a disyllabic and two successive trisyllabic words.

> E, StK 8, 86 (2) < Lüganuse – Mihkel Jürna, stud. < auntie of Karja, 85 years old (1921).

Since Genetz's discovery, these quantity rules have been absorbed into the general descriptions of runosong metre in Finnish and Estonian scholarly as well as educational materials. They have also been assumed to apply to runosong more generally without any further verification of this claim. Subsequent scholarly research on this metre has either (1) considered deviations to be inconsistencies peculiar to folk metre (Anderson 1935), (2) been based on parts of the *Kalevala* itself or on songs from the most *Kalevala*-metric areas, genres or singers (see, e.g., Saarinen 2018; Sadeniemi 1951) or (3) discussed the origins and features of *Kalevala* metre as a model without analysing specific texts (see, e.g., Korhonen 1994; Kuusi and Tedre 1979).

At the same time, systematic deviations of folksong metre from the *Kalevala* model have sometimes been noted. Estonian scholars have, thus, sometimes introduced the rules of *Kalevala* metre with the caveat that they are "statistical", i.e. they apply to the majority of verse lines, without further specifying the percentage of lines or texts that this affects (e.g. Viidalepp 1959, p. 121). On the one hand, regional differences have been noted, especially in southern Estonia, while on the other, these deviations have been attributed to foreign influences and the disintegration and extinction of the tradition (Viidalepp 1959, p. 126). The only researcher to question the relevance of *Kalevala* metre to Estonian runosong was the literary scholar Jaak Põldmäe, who raised the issue in his *Estonian Metrics* (1978). His approach, however, proceeded from a tradition of literary analysis and assumed consistent rules throughout a song. As such, it neglected the consequences of the transmission process and the variations characteristic of folklore.

With some exceptions, most Finnish research has concentrated on Karelian poems, i.e. on the most regular and *Kalevala*-metric part of the tradition. Pentti Leino (2002 [1975]) compared the metre of three groups (19th-century folk poets, earlier scholarly poets who used versions of traditional metre and some 20th-century Karelian singers) to *Kalevala* metre, noting interesting variations both in the syntax and metre of these poems. He also observed that in southwest Finland, the quantity rules about short stressed syllables were not as strict (see Laitinen 2006, p. 38; Leino 1994, p. 71). Matti Kuusi (1983, pp. 184–187) analysed the repertoire of one Ingrian Finnish singer, Maria Luukka from western Ingria and pointed out that researchers he needed to standardise some of her verses and that in the case of the shorter verses, it was impossible to know how they were performed (and thus, how singers had traditionally interpreted these verses in relation to the metrical system; see Oras 2010; Kallio 2021). Similarly, Petri Lauerma (2001, 2004) examined the metre and language used by the famous Izhorian singer Larin Paraske from northern Ingria. In this quite exceptional case, recorders had transcribed both the dictated and sung versions

of the same poems and also noted Paraske's explanations. Lauerma concluded that the words in the dictated lines, and especially the line endings, often took shorter forms that came closer to the contemporary spoken language, while the sung performance used more regular lines, i.e. with all the verse positions filled.

The question of "metric dialects" of runosong was first raised by Kuusi (1983, pp. 188–190, 1994, p. 55), who also expressed a wish to carry out further research. As part of this project, he proposed that the Finnic area could be divided metrically into two regions: (1) the *Kalevala*-metric zone, including Karelia, eastern Finland and Ingria and (2) the late-*Kalevala*-metric zone of western Finland, southern Karelia, eastern Ladoga Karelia and Estonia where, according to Kuusi, the tendency to adopt accentual metre had developed in the 16th century (Kuusi and Tedre 1979, p. 70). This suggestion was not elaborated further, however, since verification would have required an immense amount of repetitive and time-consuming analysis in the pre-computer era.

In fact, the thorough study of the variability of runosong metre only became feasible with the aid of computing technology (e.g. Sarv 2008, 2015, 2019). In the case of the Estonian material, such study has demonstrated that (1) there are clear and significant regional differences in Estonian runosong metre and (2) Estonian runosong metre represents a transitional stage between *Kalevala* metre, where the quantity of stressed syllables is a distinctive feature, and an accentual metre that relies on the placement of stressed syllables without any additional quantitative rules. The main difference between the two metrical variants lies in the placement of short stressed syllables: in quantitative runosong metre, these occupy the weak positions in the trochaic schema, while in accentual runosong metre, there is no difference quantitatively, i.e. just like their long counterparts, the short stressed syllables may be placed in strong positions (see Table 1). Throughout Estonia, both variations were used simultaneously although the proportion of lines following each metre varied significantly by region.

We still, however, lack even a general picture of the variability of runosong metre in the other Finnic regions. We have neither the slightest idea of the extent of metric variation, nor any sense of whether or how this might relate to linguistic borders. The current study is the first enquiry in the research project "Formulaic Intertextuality, Thematic Networks and Poetic Variation across Regional Cultures of Finnic Oral Poetry FILTER" (funded by the Academy of Finland), which aims to understand the metric properties of this vast corpus of material. The main goal of this first step is to obtain a rough idea of the extent of geographic variation of runosong metre using the data at our disposal and automatic analysis only. We are, of course, aware that such an approach is not equipped to handle details or special cases. It may, however, offer ideas and hints about how to plan a more substantial enquiry. At the same time, it may reveal how biases in the data are reflected in the automatic analysis.

Trochaic rhythm – intermittent **strong (S)** and **weak (W)** verse positions: SW/SW/SW/SW	
QUANTITATIVE RUNOSONG METRE	ACCENTUAL RUNOSONG METRE
Long stressed syllables of polysyllabic words are placed in **strong** line positions in both metrical variants: näin mie / **un**-ta / **moa**-tes-/sa-ni (verse structure 1124) “I dreamt a dream while sleeping”	
Short stressed syllables are placed in **weak** line positions in **quantitative** runosong metre: jo **tu-**/li **sy-**/vä **sy-**/ky-sy (1223) “a deep autumn set in” joo-tis / lau-li-/ku **ka-**/bas-ta (233) “let me drink from a singer’s cup”	**Short** stressed syllables are placed in **strong** line positions in **accentual** runosong metre: **o**-leks / **mi**-nu / **o**-le-mi-ne (224) “if it were up to me” Kad-ri / var-vad / **va**-lu-/ta-vad (224) “Kadri’s toes are aching”

Table 1: Principles for the placement of stressed syllables in two concurrent runosong metres: (1) quantitative runosong metre (*Kalevala* metre) and (2) accentual runosong metre. In the examples, the stressed syllables relevant for detecting the metrical variant (i.e. quantitative or accentual) are highlighted in bold.

3 Research Material

The material in the current study consists of runosong texts and metadata from two major runosong databases: (1) the Estonian runosong database (Oras et al. 2003–2021) and (2) the SKVR database of Karelian, Ingrian and Finnish runosongs (SKVR 2021).[5] These databases were developed in parallel and have broadly the same structure in which texts and original metadata are formatted in extended markup language (XML) and are combined with tables about the unified metadata as well as other classification information. The databases have been united in a joint system for research purposes within the FILTER project.

Despite the general coherence of the two databases, there are some slight differences in their composition principles and history that should be mentioned. The Estonian runosong database was formed directly from the archival collections at the Estonian Folklore Archives, which consist mainly of Estonian and Seto material and to some extent also Ingrian and Votic songs. The key features of this database are as follows: (1) it is not complete but contains approximately two thirds of the runosongs in the Estonian Folklore Archives; these are the oldest collections and the database is continuously being supplemented with additional texts from the newer collections (the data for the current study con-

[5] Concerning the background and development of these databases, see Sarv 2020; Harvilahti 2013 and Klemettinen 2006.

sisted of the 100,034 texts that had been added to the database by June 2020); (2) the texts in the database have not been fully classified; they are provided with preliminary classification data from the archival indexes, and the revision of the system is ongoing; (3) although the main body of the database consists of runosongs, it also contains some material from other poetic folklore genres and hybrid poetic forms; in the current stage of classification, these cannot be fully distinguished from runosongs based on the metadata; and (4) the database contains an orthographically unified version of the song texts. The SKVR database, in contrast, arose from an academic edition of runosong texts, *The Ancient Poems of Finnish People (1908–1997)*, which drew on the folklore collections of the Finnish Literary Society. It contains a total of 89,247 texts of Karelian, Finnish, Ingrian, Votic and Ingrian Finnish runosongs. For our analyses, the following attributes of the SKVR database should be kept in mind: (1) it includes all of the poems listed in the above publication but not all runosong texts in the Finnish archives; (2) it contains texts that are fully classified based on years of effort by several researchers; (3) it consists mainly of runosong texts; and (4) it is orthographically varied and, depending on the language or dialect in question, uses special marks to reflect pronunciation (for search purposes, the SKVR database also offers an automatically standardised or simplified version of texts where these special marks are removed and specific letters are replaced; see skvr.fi/ohje).

The corpus is uneven in many ways. First, there are considerable differences in the distribution of subgenres of runosong. In Estonia, for example, traditional wedding ceremonies were occasions accompanied and structured by runosongs, and wedding songs probably form the majority of Estonian runosongs. In contrast, in Finland, wedding songs are almost unknown. In Finland, Karelia and Ingria, long and narrative incantations represent a considerable part of the runosong collection, while in Estonia, incantations in runosong form comprise only a few types and they are considerably shorter in nature. Although collectors recorded all kinds of poems in these archives, they preferred metrically and poetically coherent epic or mythological works. Some kinds of songs tended to be considered unsuitable for recording, particularly if they were more improvisational or covered sexual topics. Regions such as Viena Karelia and Setomaa, where the traditions were more active and versatile during the collection period, were prioritised and surveyed most closely. The extent of the runosong collection, thus, differs substantially across regions.

The runosong texts in the corpus were transcribed from several Finnic languages that are not in themselves homogenous but have dialects and subdialects. As well as exhibiting purely linguistic variation, runosongs often conserved archaic and usually longer word forms. In some regions, runosong language is a specific archaic idiom that is structurally different from the spoken language based on its prosodic and morphological features. The peculiarities of texts in dialects that lack a consistent literary standard have sometimes motivated collectors to invent their own writing systems to record phenomena and sounds not present in literary Finnish or Estonian. Some words in the research corpus may have hundreds of different forms due to dialectal variation, archaic runosong idioms, orthographic inconsistencies and morphologic variants. There are no

dictionaries or automatic tools for lemmatisation or grammatical analysis of this variability; even the simplest syllabification rules may apply differently across spoken languages as well as across runosong forms.

4 Methodological Considerations

For the computational analysis of a syllabic metre with subsequent quantity regulation, it would seem natural to use line syllabification as a starting point. This would involve detecting the stress and quantity of the syllables to see how they map onto the octapositional verse lines. However, such an automatic analysis of our corpus is complicated for several reasons:

1. Although runosong metre is syllabic in nature, regionally different replacement rules apply; for example, one position is often filled by two light syllables while in some regions, a long or overlong syllable may be stretched over two line positions.

2. Syllables with long vowels, and especially diphthongs that historically derive from two syllables, behave flexibly: they may either fill one position in the verse schema or be divided between two positions. In addition, there are no strict syllabification rules for this kind of long vowel that could serve as a basis for automatic syllabification.

3. In the case of compounds, the stress on the first syllable of the second component is also relevant to the metre but there is no quick way to divide compounds automatically.

4. Depending on the region, the eighth line position sometimes remains unfilled or the last two line positions may be filled by one long syllable, which in historical linguistic practice would have been two syllables.

Syllabification (and performance) practices vary by genre, region and individual. Singers may have interpreted verse rhythms based on different linguistic layers or made choices about lyrics based on different interpretations of word structures. A combination of these factors complicates the automatic syllabification process. This affects the detection of which syllables are stressed and whether they should be considered short or long as well as the final mapping of the syllables onto positions in the verse schema.

In addition, there are deviations that affect only minor parts of the corpus:

1. Although most works in the runosong tradition used a performance style in which a structural melodic note corresponded to a verse position (generally a syllable), specific melodies and regional performance styles may have required repetition or additional structural syllables. This may, in turn, have affected a song text and occasionally its metre. Performance factors may also have influenced how singers created songs and verse. The performance style could, for instance, affect the number of syllables in words and lines by causing singers to make verses fit into rhythmic song structures in various ways.

2. There are different linguistic and orthographic variants and peculiarities across the corpus that may also influence the syllabification rules in some cases (in addition, some collectors tended to "translate" songs in local dialects into standard Finnish or Estonian).

3. Recorders sometimes used abbreviations and incomplete wording; singers also used different and less metrical word forms when they dictated a song as opposed to when they sang it (see, e.g., Lauerma 2004); some spoken poetic genres like charms and songs for children tended to take less regular verse forms.

In sum, all of these circumstances pose a substantial obstacle for anyone attempting a quick but sound automatic analysis of the metre throughout the corpus. This led us to look for an indirect approach that might provide a preliminary idea of the extent of metric variation in the area associated with runosong.

In proposing a method to indirectly estimate metrical variation across the corpus, we relied on the results of an analysis of Estonian runosong metre by Mari Sarv (2008) that was based on a sample of around 50,000 verse lines (equal samples of 500 lines from each of the Estonian parishes). This study had revealed that the regional variability of metre clearly overrides any metrical differences among the main song genres (lyric songs, lyroepic songs and ritual songs), variability due to recording notation methods or differences among individual singers. The metre of Estonian runosongs turned out to follow a pattern of clear geographical variation. Second, this study managed to demonstrate that the features of verse metre depended on dialectal prosody (Sarv 2008, 2019). As many past researchers have noted, the metres widely used in a language need to conform with its prosodic system, especially in the case of indigenous folk metres that develop over extended periods together with the language (see, e.g., Ross et al. 2001, p. 3; Lotman 1998, pp. 1853, 1858–1859; Jakobson 1979, pp. 148–150).

In the case of Estonian runosong, where metrical changes reflected substantial changes in the language, the clearest and most straightforward linguistic correlate with metre was the average length of words in song texts: in parishes where the majority of lines reflected typical runosong quantitative verse, the average word length in syllables tended to be longer. In contrast, in parishes where most lines conformed with typical runosong accentual metre, the average word length was shorter. The correlation between the percentage of typical quantitative lines and word length across 104 parishes was 0.83. The reasons for this correlation can be traced to a set of major historical changes in the Estonian language that systematically decreased the number of syllables in words. To adapt to these changes, the metre took on a new form better suited to the prosodic system of the revised language. These linguistic changes did not take place all at once or across all of Estonia but rather gradually and with different effects on Estonian dialects, and thus, also on runosong metre.

Concerning biases, Sarv's study showed that on average, more runosongs had been recorded from parishes with a higher percentage of runosong lines typical of quantitative metre (the correlation between the percentage of lines typical of quantitative runosong metre in a parish and the number of songs from

the same parish in the Estonian runosong database was $r = 0.49$). On the one hand, this may point to a connection between metrical changes and the fading of tradition (the active tradition in some parishes kept the metre from changing abruptly). On the other, we may assume that the collection process relied on the material available at a time when the tradition was gradually vanishing: it would have been hard to find singers in places where the tradition was about to disappear and easier to collect songs in those with an active practice. Collectors sometimes also preferred to record songs from regions where songs were longer and followed quantitative metre because these were seen as the characteristics of "true" runosong.

Based on Sarv's findings, we hypothesised that the average word length in the runosongs recorded in a parish might reveal the extent of runosong metric variation on the quantitative-accentual scale. We did not, however, have any direct means to automatically test this hypothesis and therefore needed to take an alternative approach: since the most common line structures typical of quantitative runosong metre (233, 323, 332) have no regular match in accentual runosong metre, we used their occurrence as a complementary indicator of metric variation.

5 Analysis and Results

The two aims of this preliminary research were (1) to calculate the average number of syllables in words in the current corpus by location and region and (2) to determine the percentage of line structures typical of quantitative runosong metre. In order to achieve these goals, we first needed, however, to syllabify the corpus. For the approximate syllabification of the whole corpus, we applied Finnish syllabification rules irrespective of differences in language, dialect or orthography. We were aware that this rough method might result in occasional mis-syllabification, in particular concerning: (1) diphthongs and long vowels and (2) compounds. Treating compounds as one word rather than two or three words systematically increases the average number of syllables and, to some degree, distorts line structure percentages. We assumed, however, that these occasional and systemic differences would affect regional texts fairly evenly and so would not seriously modify general trends.

In order to test our assumptions, we first compared the average word lengths obtained from 1) the rough syllabification of the whole Estonian part of the corpus and 2) the results of Sarv's 2008 study based on samples of 500 lines per parish. In the latter, the compounds had been divided and the automatic syllabification of diphthongs was re-checked manually. It must be reiterated that the corpus of the current study, which includes far more varied material, was somewhat different from Sarv's research sample where certain formal criteria served to restrict the material to only typical runosongs.

As expected, the results of this test demonstrated a clear positive correlation ($r = 0.80$) between the average word length in syllables based on the current rough syllabification of the corpus, and the average word length found in the previous study (Figure 2). Closer examination of the divergence of the results

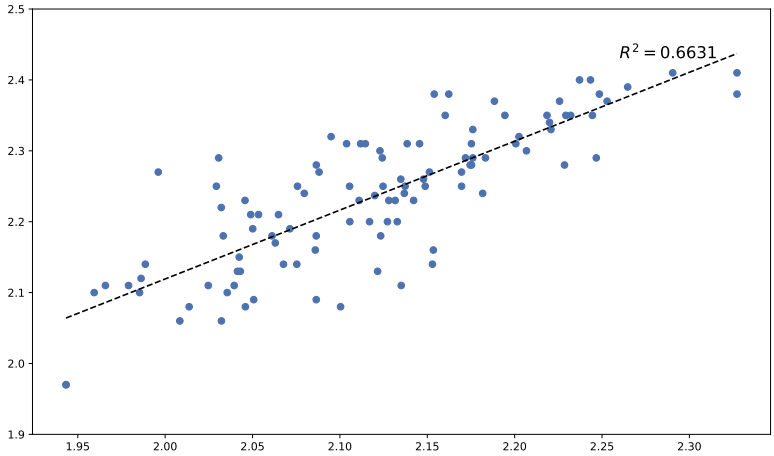


Figure 2: A comparison of the average word length in syllables based on 1) the rough syllabification of the Estonian part of the corpus of the current study (y-axis) and 2) Mari Sarv's 2008 study of Estonian runosongs (x-axis) with a linear regression line. Each dot represents a parish in Estonia.

from the linear regression line (or the predicted results) revealed that in general the results matched the trendline better in parishes where more material was available; greater differences were recorded in parishes where there was less material (fewer words) in the corpus (Figure 3). There appeared to be no clear regional difference in this distribution. The correlation between the percentage of lines typical of quantitative metre in the previous study and the average length of words in syllables was almost the same for the rough corpus-based syllabification ($r = 0.84$) and the re-checked syllabification from the previous study ($r = 0.83$). These results, thus, suggested that our rough syllabification was a feasible method at least for detecting the average word length in syllables, which could then be used to assess the regional variation of runosong metre on the quantitative-accentual scale.

The maps below present the results of two independent surveys in our study and give a preliminary estimate of the extent and spread of metric variation across the whole runosong-associated region.

The average word length in syllables for the entire corpus is depicted in Figure 4 (per parish) and Figure 5 (per county or dialect area).[6] In Figure 4, the results are shown in greater geographical detail. The less granular view in Figure 5 relies on the regional division in relevant databases. This view counterbalances the extreme results that tend to occur in cases where only a few recordings are available from a parish.

Figures 6 and 7 show the total percentage in the corpus of the most common line structures typical of quantitative runosong metre (233, 323, 332).

[6] In the maps on the following pages, parishes/regions are divided into four classes of approximately equal total area. The colours in these maps represent different intervals for the given variable.

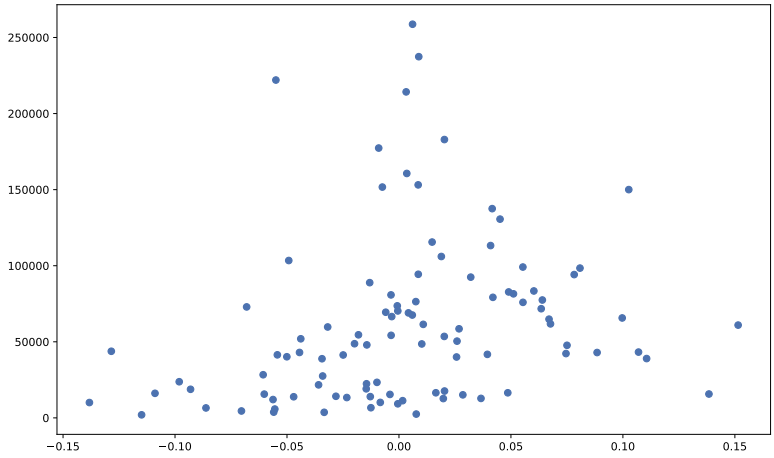


Figure 3: The deviation of the results of our corpus-based syllabification from the predicted average word length in syllables (trendline from Figure 2) is shown on the x-axis. The number of words in the corpus from the given parish appears on the y-axis. Each dot represents a parish in Estonia.

The results of the two surveys are generally congruent with one another and reflect a clear geographical logic: the indicators of quantitative runosong metre, i.e. the average word length and percentage of specific line structures, decrease gradually from the north-east (Karelia) to the south-west (Estonia). The somewhat scattered picture in the results at parish level (Maps 4 and 6), especially in Finland, is probably a result of scant collection data, which tends to represent a specific selection rather than the "average runosong" of the parish. As we have seen (Figure 3), where less material is available, the outcome is less balanced, and this leads to more diffuse results.

The correlation between the parish-level results for the two surveys is 0.61. However if we exclude from this calculation the 109 parishes where fewer than 100 verse lines were collected, the correlation rises to 0.76. Across the counties/dialect areas, the correlation between the two surveys is 0.79. The correlation between these two features suggests that longer words are required to form the lines typical of quantitative runosong metre. In this respect, the deviations from the expected values based on a linear regression line can be seen on Figure 8. We see that in Karelia, eastern Finland and Ingria, the percentage of lines typical of quantitative runosong metre is higher than expected based on the average word length in syllables. On the other hand, there are regions where the traditional use of lines typical of quantitative metre has declined despite the length of the words. In these regions, the Finnish population has mixed with either the Swedish or Sámi population and this may have had an impact on the runosong tradition. (Sarv (2011) describes a comparable situation in the Estonian borderlands.) The deviations in either direction from expected values probably relate to the tension between the endurance and disappearance of the runosong tradition in these areas.

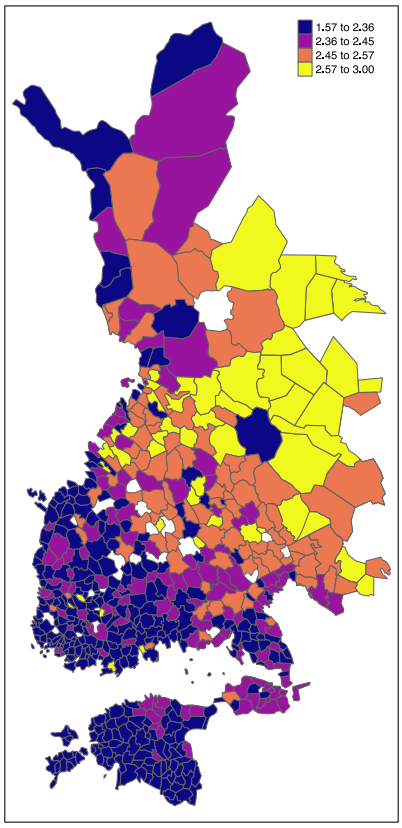


Figure 4: Average word length in syllables in the runosong texts for each parish in the corpus

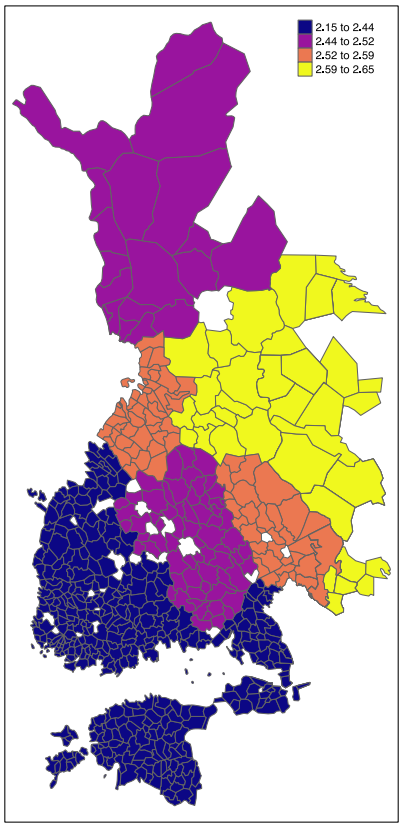


Figure 5: Average word length in syllables in the runosong texts for each region (county or dialect area) in the corpus

Based on the results of this study, we may assume the following:

1. the metre of runosongs is generally more quantitative in Karelia and eastern Finland and less quantitative in Ingria, western Finland and Estonia (this generally conforms with the earlier proposal by Matti Kuusi except that Ingria belongs to the less quantitative region in the current results) and

2. changes in metric features depend on changes in linguistic/dialectal prosody (as is most clearly shown in the shortening of words in our study). Such changes do not necessarily follow language or dialect borders (cf. Sarv 2008, 2019).

The previous Estonian study (Sarv 2008) showed that lines typical of quantitative runosong metre dominated the tradition across half of Estonia, while in the other half, accentual metre prevailed. The more quantitative metre-defined

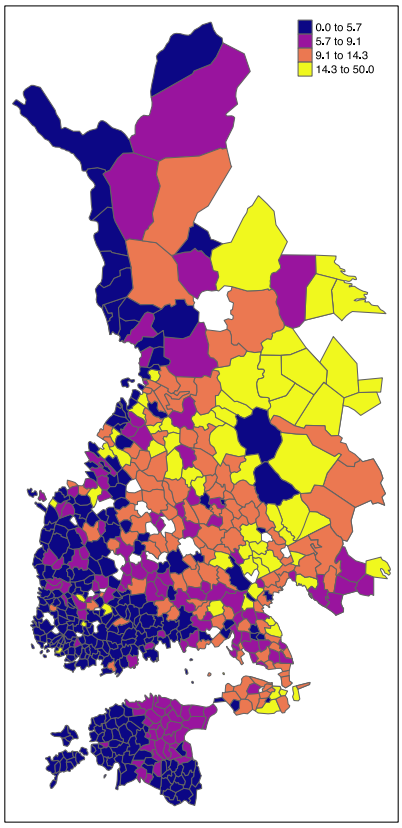


Figure 6: Total percentage of the most common line structures typical of quantitative runosong metre (233, 323, 332) in the runosong texts for each parish in the corpus

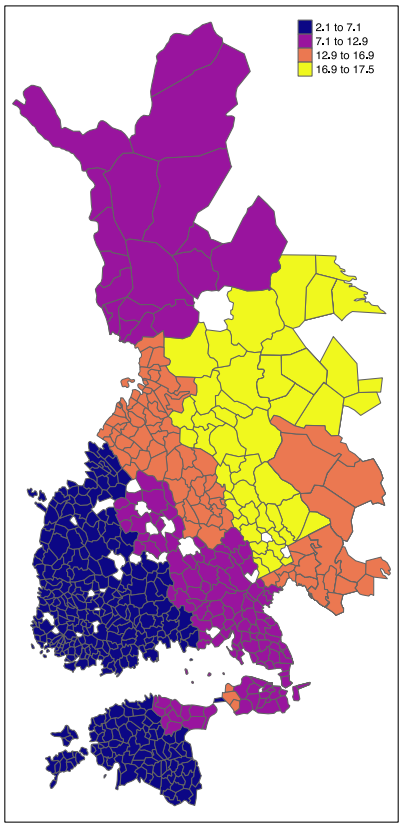


Figure 7: Total percentage of the most common line structures typical of quantitative runosong metre (233, 323, 332) in the runosong texts for each region (county or dialect area) in the corpus

area largely overlaps with the violet area on Figure 4, while the dark blue shading corresponds with the area more associated with accentual metre in Sarv's study. On this basis, we may propose a working hypothesis: in one quarter of the Finnic runosong area (the dark blue areas), accentual metre prevails, while in the other three quarters of this space, features typical of quantitative metre predominate.

These assumptions and hypotheses call for verification through more detailed analysis of the actual metric qualities of runosongs at the level of parish, region, genre, singer, collector and related variables.

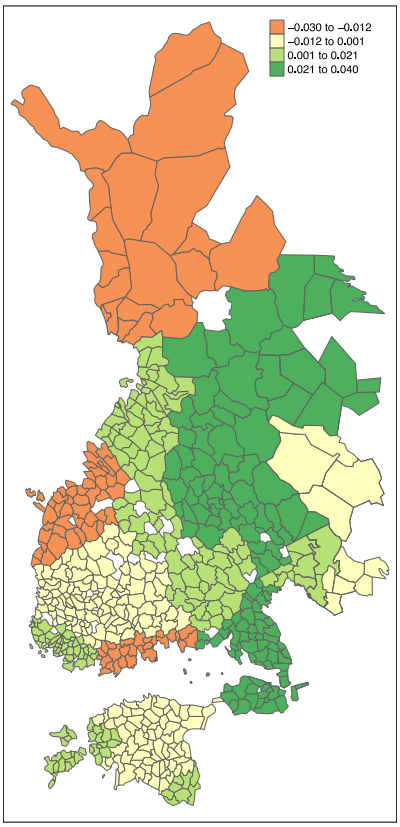


Figure 8: Deviations from expected values (based on a linear regression line) for the percentage of lines of typical quantitative runosong metre with the given average word length

6 Conclusion

An immense amount of work will be needed to complete a careful and sound metric analysis of thousands of texts of oral poetry that have been transcribed in different but related languages and dialects, in different orthographies and with different degrees of accuracy. Even if we use computing tools for the metrical analysis, collecting and considering background information about linguistic and folkloric variation, performance styles and the like remain enormous tasks. However, the current study has shown that a large-scale approximation performed by computational analysis can deliver meaningful results in more general terms if we agree to overlook the details.

Finnic runosong metre is generally thought to have originated from a model of quantitative metre (so-called *Kalevala* metre) with specific rules about the placement of stressed syllables based on their quantity. Along with linguistic changes, especially the shortening of words, and under various cultural influences in many regions in the area, the metre gradually lost its quantitative

features and shifted towards an accentual metre in which quantity rules were loosened or disappeared altogether. There is no previous data-based research of this metrical variation across this entire Finnic runosong area. The current study has estimated the extent and geographic nature of variation in the Finnic runosong tradition on the quantitative-accentual scale based on an approximate syllabification of a corpus of approximately 200,000 song texts. To this end, we have used two different indicators: (1) the average word length in syllables by parish/region, and (2) the percentage of the most common line structures typical of quantitative runosong metre by parish/region. The results of both surveys appear to be generally coherent, thus providing the very first large-scale, data-based overview of metric variation in the Finnic runosong tradition. The data and code used in the current study are available at the Zenodo repository (Sarv et al. 2021).

Acknowledgments

This study was supported by the Finnish Academy (project no. 333138 Formulaic Intertextuality, Thematic Networks and Poetic Variation across Regional Cultures of Finnic Oral Poetry). Further support was received from the Estonian Ministry of Education and Research (PRG1288 A Corpus-based Approach to Folkloric Variation: Regional Styles, Thematic Networks, and Communicative Modes in Runosong Tradition, and EKKD65 Source Documents in the Cultural Process: Estonian Materials in the Collections and Databases of the Estonian Literary Museum) and by the European Regional Development Fund (Centre of Excellence in Estonian Studies).

References

Anderson, Walter (1935). *Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder. Eesti Rahvaluule Arhiivi Toimetused 2./Acta et Commentationes Universitatis Tartuensis B XXXIV1*. Tartu.

Anttonen, Pertti (2005). *Tradition through Modernity. Postmodernism and the Nation-State in Folklore Scholarship*. Helsinki: Finnish Literature Society. DOI: 10.21435/sff.15.

Frog (2019). "The Finnic Tetrameter – A Creolization of Poetic Form?" In: *Studia Metrica et Poetica* 6.1, pp. 20–78. DOI: 10.12697/smp.2019.6.1.02.

Genetz, Arvid (1884 [1881]). *Suomen kielioppi. Äänne-, Muoto- ja Runous-oppi. Oppikouluja varten. Toinen painos*. Helsinki: K. E. Holm'in kustantama.

Grünthal, Riho (forthcoming). "The Finnic Languages and Main Dialects". Revised version from Grünthal & Sarhimaa 2004 Itämerensuomalaiset kielet ja niiden päämurteet. URL: https://www.sgr.fi/muutjulkaisut/ItamerensuomalaisetKieletMurteet2012.pdf.

Harvilahti, Lauri (2013). "The SKVR Database of Ancient Poems of the Finnish People in Kalevala Meter and the Semantic Kalevala". In: *Oral Tradition* 28.2, pp. 223–232.

Helimski, Eugen (1998). *Samojedit ja šamanismi: Viisi luentoa samojedeista, šamanismista ja uralilaisesta kulttuurista*. Ed. by Timo Leisiö Larisa; Leisiö. Tampere: Tampereen Yliopisto, Kansanperinteen Laitos.

Huttu-Hiltunen, Pekka (2008). *Länsivienalainen runolaulu 1900-luvulla. Kuuden runolaulajan laulutyylin kulttuurisensitiivinen musiikkianalyysi*. Kuhmo: Juminkeko.

Jakobson, Roman (1979). "Metrics". In: *Selected Writings V. On Verse, Its Masters and Explorers*. The Hague, Paris, New York, pp. 147–159.

Kallio, Kati (2021). "Performance, music and meter in Finnic Kalevala-metric poetry". In: *Versification: Metrics in Practice*. Ed. by Frog, Satu Grünthal, Jarkko Niemi, and Kati Kallio. Helsinki: SKS, pp. 59–78. DOI: 10.21435/sflit.12.

Kallio, Kati, Frog, and Mari Sarv (2017). "What to Call the Poetic Form: Kalevala-Meter, Regivärss, Runosong, Alliterative Finnic Tetrameter, or Something Else?" In: *RMN Newsletter* 12–13, pp. 139–161.

Klemettinen, Pasi (2006). *"Ei se synny synnyttämättä" Selvitys digitointiprojektin vaiheista ja työprosesseista*. Suomalaisen Kirjallisuuden Seura. URL: https://www.finlit.fi/sites/default/files/mediafiles/tutkimus/elias_loppuraportti.pdf.

Korhonen, Mikko (1994). "The Early History of the Kalevala Metre". In: *Songs Beyond the Kalevala. Transformations of Oral Poetry. Studia Fennica Folkloristica 2*. Ed. by Anna-Leena Siikala and Sinikka Vakimo. Helsinki: SKS, pp. 75–87.

Krikmann, Arvo (1997). *Sissevaateid folkloori lühivormidesse I*. Tartu: Tartu Ülikooli kirjastus. URL: http://haldjas.folklore.ee/~kriku/LEX/KATUS.HTM.

Kuusi, Matti (1983). *Maria Luukan laulut ja loitsut*. Suomalaisen Kirjallisuuden Seuran toimituksia 379. Mikkeli.

Kuusi, Matti (1994). "Questions of the Kalevala Metre". In: *Songs Beyond the Kalevala. Transformations of Oral Poetry. Studia Fennica Folkloristica 2*. Ed. by Anna-Leena Siikala and Sinikka Vakimo. Helsinki: SKS, pp. 41–55.

Kuusi, Matti and Ülo Tedre (1979). "Regivärsilise ja kalevalamõõdulise laulutraditsiooni vahekorrast. Dialoog üle lahe". In: *Keel ja Kirjandus* 2, pp. 70–79.

Laitinen, Heikki (2006). "Runolaulu". In: *Suomen musiikin historia 8: Kansanmusiikki*. Ed. by Anneli Asplund et al. Helsinki: WSOY, pp. 14–79.

Lauerma, Petri (2001). "Larin Parasken metriikasta". In: *Virittäjä* 1, pp. 44–58.

Lauerma, Petri (2004). *Larin Parasken epiikan kielellisestä variaatiosta*. Helsinki: SKS.

Leino, Pentti (1986). *Language and Metre: Metrics and the Metrical System of Finnish. Studia Fennica 31*. Helsinki: SKS.

Leino, Pentti (1994). "The Kalevala Metre and its Development". In: *Songs beyond the Kalevala. Transformations of Oral Poetry. Studia Fennica Folkloristica 2*. Ed. by Anna-Leena Siikala and Sinikka Vakimo. Helsinki: SKS, pp. 56–74.

Leino, Pentti (2002 [1975]). "Äidinkieli ja vieras kieli: rahvaanrunouden metriikkaa". In: *Mittoja, muotoja, merkityksiä*. Helsinki: SKS, pp. 207–230.

Lord, Albert B. (1960). *The Singer of Tales*. Repr. 2000. Cambridge: Harvard University Press.

Lotman, Mihhail (1998). "Värsisüsteemidest. Peamiselt eesti ja vene värsi näitel". In: *Akadeemia* 9–10, pp. 1846–1874, 2058–2078.

Oras, Janika (2008). *Viie 20. sajandi naise regilaulumaailm. Arhiivitekstid, kogemused ja mälestused. Eesti Rahvaluule Arhiivi toimetused 27*. Tartu: Eesti Kirjandusmuuseumi Teaduskirjastus.
Oras, Janika (2010). "Musical manifestations of textual patterning in Estonian regilaul". In: *Journal of Ethnology and Folkloristics* 4.2, pp. 55–68.
Oras, Janika (2015). *Vadja ja isuri rahvalaulud*. Tartu: MTÜ Rahvalauluselts Hellero, Estonian Folklore Archives of the Estonian Literary Museum. URL: http://www.folklore.ee/pubte/eraamat/vadjaisuri/en.
Oras, Janika, Liina Saarlo, and Mari Sarv, eds. (2003–2021). *Eesti regilaulude andmebaas*. Tartu: Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiiv. DOI: 10.15155/9-00-0000-0000-0000-0008FL. URL: http://www.folklore.ee/regilaul/.
Põldmäe, Jaak (1978). *Eesti värsiõpetus*. Tallinn: Eesti Raamat.
Porthan, Henrik Gabriel (1983). *Suomalaisesta runoudesta 1766–1778 [De poësi Fennica]*. Käänt. Iiro Kajanto. Helsinki: SKS.
Ross, Jaan and Ilse Lehiste (2001). *The Temporal Structure of Estonian Runic Songs. Phonology and Phonetics 1*. Ed. by Aditi Lahiri. Berlin, New York: Mouton de Gruyter.
Rüütel, Ingrid (1998). "Estonian Folk Music Layers in The Context of Ethnic Relations". In: *Folklore* 6. URL: http://haldjas.folklore.ee/folklore/vol6/ruutel.htm.
Saag, Lehti et al. (2019). "The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East". In: *Current Biology* 29.10, 1701–1711.e16. DOI: https://doi.org/10.1016/j.cub.2019.04.026.
Saarinen, Jukka (2018). *Runolaulun poetiikka. Säe, syntaksi ja parallelismi Arhippa. Perttusen runoissa. Akateeminen väitöskirja*. Helsinki: Helsingin yliopisto. URL: http://urn.fi/URN:ISBN:ISBN%20978-951-51-3919-1.
Sadeniemi, Matti (1951). *Die Metrik des Kalevala-Verses. FF Communications 139*. Helsinki: Suomalainen Tiedeakatemia.
Sarajas, Annamari (1956). *Suomen kansanrunouden tuntemus 1500–1700-lukujen kirjallisuudessa*. Porvoo: WSOY.
Sarv, Mari (2008). *Loomiseks loodud: regivärsimõõt traditsiooniprotsessis*. Tartu: EKM Teaduskirjastus.
Sarv, Mari (2011). "Possible foreign influences on the Estonian regilaul metre: language or culture?" In: *Frontiers in Comparative Prosody*. Ed. by Mihhail Lotman and Maria-Kristiina Lotman. Linguistic Insights; 113. Peter Lang Verlag, pp. 207–226.
Sarv, Mari (2015). "Regional Variation in Folkloric Meter: The Case of Estonian Runosong". In: *RMN Newsletter* 9, pp. 6–17.
Sarv, Mari (2019). "Poetic metre as a function of language: linguistic grounds for metrical variation in Estonian runosongs". In: *Studia Metrica et Poetica* 6.2, pp. 102–148. DOI: 10.12697/smp.2019.6.2.04.
Sarv, Mari (2020). "From tradition to data: The case of Estonian runosong". In: *Arv. Nordic Yearbook of Folklore* 76, pp. 105–117.
Sarv, Mari, Kati Kallio, Maciej Janicki, and Eetu Mäkelä (2021). *Metrical variation in the Finnic runosong tradition: challenges and possibilities for a computational approach (code + dataset) (Version 1.0)*. [Dataset]. Zenodo. DOI: 10.5281/zenodo.4537606.

SKVR (2021). *SKVR-tietokanta – kalevalaisten runojen tietokanta*. Suomalaisen Kirjallisuuden Seura. URL: http://skvr.fi (visited on 07/14/2021).

Tampere, Herbert, Erna Tampere, and Ottilie Kõiva (2016). "Eesti rahvamuusika antoloogia. Veebiväljaanne". In: *Helisalvestusi Eesti Rahvaluule Arhiivist 3*. Ed. by Janika Oras and Kadi Sarv. Tartu: Eesti Kirjandusmuuseumi Teaduskirjastus. URL: http://www.folklore.ee/pubte/eraamat/rahvamuusika/en/.

Tampere, Herbert, Erna Tampere, and Ottilie Kõiva (2017). "The Field of Song and the Four-Legged Horse: On the Dialogue of Genres in Kalevala-Meter Poetry". In: *Classics@: Singers and Tales in the 21st Century; The Legacies of Milman Parry and Albert Lord*. Harvard: Center for Hellenic Studies, Harvard University. URL: https://chs.harvard.edu/CHS/article/display/6617.

Tarkka, Lotte, Eila Stepanova, and Heidi Haapoja-Mäkelä (2018). "The Kalevala's Languages: Receptions, Myths, and Ideologies". In: *Journal of Finnish Studies* 21.1–2, pp. 15–45. URL: http://hdl.handle.net/10138/301432.

Viidalepp, Richard (1959). "Eesti rahvalaulude poeetika ja keel". In: *Eesti rahvaluule ülevaade*. Ed. by Richard Viidalepp. Eesti Riiklik Kirjastus: Tallinn, pp. 116–177.

https://doi.org/10.51305/ICL.CZ.9788076580336.10
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

The Influence of Verse on Cognitive Processes: A Psycholinguistic Experiment

Tatyana Skulacheva
Institute of Russian Language of the Russian Academy of Sciences, Russia
skulacheva@yandex.ru
0000-0002-0679-6084

Natalia Slioussar
Higher School of Economics / St. Petersburg State University, Russia
slioussar@gmail.com
0000-0003-1706-6439

Alexander Kostyuk
Institute of Russian Language of the Russian Academy of Sciences, Russia
kostyuk.ae@gmail.com
0000-0002-8104-4245

Anna Lipina
Russian University for Humanities, Russia
anna.lipina.94@mail.ru

Emil Latypov
Russian University for Humanities, Russia
latypow.emil@gmail.com

Varvara Koroleva
Russian University for Humanities, Russia
varyak1999@gmail.com

Abstract

Modern psycho- and neurolinguistics use standards of precision typical of the natural sciences. As verse scholarship also bases its standards on those of the natural sciences, it can be combined fruitfully with the natural sciences, including neuroscience. This may ultimately allow us to answer the fundamental question of how verse and prose are processed in the brain. In this paper, we present the preliminary results of our project that aims to uncover how verse's effects on cognitive processes compare to those of prose. We conducted 3 experiments with 110 informants who were native speakers of Russian between 18 and 55 years old. These experiments had the same design but involved different stimulus texts and groups of informants (40+40+30). Informants are known to slow down their reading considerably if they detect a textual error. Our aim was to compare the reading times for different verse and prose fragments when they contained errors and when they were error-free. We found that errors in verse remain undetected while the same errors are easily perceived in a corresponding prose text. The observation of this phenomenon in all three experiments is important proof of its validity. We suggest that prose and verse differently activate two ways of processing information in the brain: the first way is logical and relies on critical thinking including error detection, while the second is associative and depends on mental imagery rather than sequential logic.

1 Introduction

Modern psycho- and neurolinguistics use standards of precision typical of the natural sciences. As the standards of verse studies also reflect those of the natural sciences (Yarxo 2006), it can be paired productively with natural science disciplines, including neuroscience. This may eventually allow us to answer the key question of how verse and prose are each processed in the brain. In this paper, we present the preliminary results of our project that aims to uncover how verse's effects on cognitive processes compare with those of prose.

In our previous studies, we collected a vast amount of data on the linguistic structure of verse (Skulacheva 1996, 2004, 2014, 2016; Skulacheva and Buyakova 2010; Skulacheva and Kostyuk 2018, 2019). The main point of these investigations was to identify the most stable regularities in verse irrespective of language, period, versification system, literary trend or individual style. In other words, we were interested in the properties that presumably characterize verse as such as opposed to prose. At the time of writing, we have found a number of such regularities. However these do not refer to meter, rhyme, stanzas, alliteration, enjambment or any other parameter readily associated with verse form. Regular meter, rhyme and stanzaic structure are often absent from irregular types of verse. Alliteration and enjambment are never regular even in modern classical verse. Nevertheless it turns out that several aspects of verse structure are present in all verse types across different languages, periods and literary trends, including the most vaguely organized free verse.

In this regard, parataxis significantly prevails over hypotaxis in verse works. This has been shown for Russian, French, English, German and Spanish verse and prose (Skulacheva and Buyakova 2010; Skulacheva and Kostyuk 2019, 2020). Secondly, monotonous intonation with no distinct descent at the end of the sentence is characteristic of verse—the diapason is twice as narrow in verse as it is in prose. These generalizations have been made regarding Russian verse and prose (Kostyuk 2017; Krivnova et al. 2020; Skulacheva and Kostyuk 2020), and we have recently obtained comparable data for English. A similar phenomenon has been observed by Yanko (2010, 2015) with regard to Russian and Arabic prayer. Thirdly, the context of verse often precludes the choice of a single meaning of a polysemous word while the prose context facilitates the choice of a particular meaning. A number of other regularities have also been found. These work at different linguistic levels and seem to be aimed at hampering comprehension. But why compose a text and, at the same time, make it difficult to understand? Our hypothesis is that all of the linguistic mechanisms mentioned above serve to impede sequential logical thinking. As a result, other processing strategies (and specifically those that are parallel, associative and rely on mental imagery) become more prominent.

We came across many examples where logical and critical thinking had apparently been suppressed by the structure of verse, and as a result, readers consistently missed even serious logical paradoxes and ambiguities. These included a famous poem by A. Blok "V goluboj, dalekoj spalenke…" ("In a distant blue room…"). The poem describes a room in which we find a young woman and a child. The child is in bed, and the verb used to refer to him is *opochil*, which in

Russian may mean both “fell asleep” and “died”. Normally, the context shows unambiguously which of the two meanings should be chosen. But Blok’s context is so subtly suggested that it is hard to be sure which of the two meanings is correct. After discussing this poem with different audiences, we found that they tended to divide into two almost equal groups: those who were sure that the author was depicting a cozy room that held a woman he loved and a sleeping child, and those who believed that the child was dead and the poem described a tragedy. It was only when the text was retold in prose that both audiences realized that a different interpretation was possible and they had not fully understood the text.

We wished, however, to proceed beyond examples to testable hypotheses. How could we prove that information in verse is processed so differently that mistakes and ambiguities obvious in prose remain undetected? To do this, we designed a series of psycholinguistic experiments. The current paper presents the first statistically valid evidence that mistakes of different kinds are not noticed in verse whereas they are easily detected in prose. They prove that mistakes of different kinds are not noticed in verse whereas they are easily detected in prose.

2 Methods

Many experimental studies in linguistics rely on measuring reading time. One basic observation recorded since at least the early 1980s (Just and Carpenter 1980; Just, Carpenter, and Wooley 1982) is that various types of errors cause the reader to slow down. To estimate how much the reader slows down and to determine whether the effect is statistically significant, the average reading time for a textual fragment is compared when it contains a particular error versus when it is error-free. More recent studies have shown that the extent of this effect depends on how easily the error is detected: errors which are intuited as more serious and noticed more easily (as measured independently via error detection rates) lead to an extension of reading time. In contrast, errors that tend to go unnoticed barely slow down the reader. Studies of subject-verb agreement errors may serve as an example (e.g. Clifton et al. 1999; Dillon et al. 2013; Pearlmutter et al. 1999; Wagers et al. 2009).

In our study, we relied on these observations. Our aim was to compare reading times for different verse and prose fragments when they did and did not contain errors. We conducted 3 experiments in which a total of 110 informant volunteers took part. These experiments had the same design but involved different stimulus texts and different groups of informants (40+40+30) — the goal was to check whether the results would be replicated despite these differences. The informants were Russian native speakers from 18 to 55 years old who were not aware of the real focus of the experiments.

Since we hypothesized that the failure to detect mistakes in verse was due to the suppression of logical thinking, it was important not to activate logical thinking through the design of our experiments. A direct question like “Did you notice a mistake in this text?” might neutralize a suppressive effect by drawing

the reader's attention to the text's logic. The informants were therefore told that we were studying the moods in which people preferred to read different poems and prose fragments. At the beginning of the experiment, they were also asked to complete a questionnaire. In addition to regular questions about their age, gender and education, they were asked about their current mood. The mood classification for this purpose was taken from a previous study (Proxorov 2011). After they read each piece of verse or prose, the informant was asked if the fragment had resonated with their mood.

At the same time, the reading time data that we were interested in were obtained as follows: the experiments were conducted online on the `Ibexfarm` platform (Drummond et al. 2016). Each informant read a verse or prose fragment and then pressed a button to proceed to the questions about the fragment. At this point, the program recorded their reading time. Later we compared the reading times for fragments that did and did not contain a mistake for both verse and prose. The informant was instructed to read the text fragments and answer questions about them. In each case, the experiment started with two "filler" fragments (see below) so that the informant could get used to the procedure.

The stimulus texts (i.e. the texts for the experiments) were prepared as follows: we chose four-line fragments from different types of Russian verse—these ranged from classical meters to dolnik to verse libre. We also prepared prose counterparts of the fragments. The prose variants were kept as close to the original texts as possible in lexis and syntax. Only obvious meter, rhyme and word order characteristics typical of verse were changed. One error was then introduced in each text, preferably in the second or third line of a four-line verse fragment, and the same error was used in the prose variant. The errors were of the following types:

1. Lexical-semantic
   (a) *three* instead of *two* for *the count and the countess – three of them* instead of *two of them*
   (b) *up* instead of *down* (where "underneath" is the intended meaning) and vice versa – *climb up to the gulley/cellar* instead of *walk down to the gulley/cellar* and *step down to the very top* instead of *step up to the very top*
2. Syntactic-semantic
   (a) *trees of branches* instead of *branches of trees*
   (b) *the beard pours from the water* instead of *the water pours from the beard* (of a drowned man in A. Pushkin's poem "Utoplennik").

Below we provide some examples that show 1) the original verse text; 2) its prose counterpart; 3) the verse text with an error; and 4) the prose counterpart with an error.

1.1 My tajnobrachnye cvety...
Nikto ne znal, chto my lyubili,

Chto aromat lyubovnoj pyli
Vdoxnuli *dvoe — ya i ty*!

1.2 My — tajnobrachnye cvety. Nikto ne znal o nashix chuvstvax, o tom, chto *dvoe — ya i ty* — vdoxnuli lyubvi aromat.

("We are like secret flowers. Nobody knew that we loved one another, that the *two of us—you and I*—inhaled the perfume of love."—Original verse fragment and its prose counterpart)

1.3 My tajnobrachnye cvety...
Nikto ne znal, chto my lyubili,
Chto aromat lyubovnoj pyli
Vdoxnuli *troe — ya i ty*!

1.4 My — tajnobrachnye cvety. Nikto ne znal o nashix chuvstvax, o tom, chto *dvoe — ya i ty* — vdoxnuli lyubvi aromat.

("We are like secret flowers. Nobody knew that we loved one another, that the *three of us—you and I*—inhaled the perfume of love."—Verse and prose fragments with an error introduced: *three* instead of *two*)

2.1 O, tol'ko by domoj dojti! Pyatnadcat'
Minut xod'by. Pyat' ulic minovat'.
Po lestnice na samyj verx podnyat'sya
I v dver' uslovnym stukom postuchat'...

2.2 O, dojti by tol'ko domoj! Pyatnadcat' minut xod'by. Minovat' pyat' ulic. *Podnyat'sya na samyj verx po lestnice* i postuchat' uslovnym stukom v dver'.

("Oh, if I could only get home! It is a 15-minute walk. Cross five streets. *Climb the stairs up to the very top* and knock on the door with a secret knock."—Original verse fragment and its prose counterpart)

2.3 O, tol'ko by domoj dojti! Pyatnadcat'
Minut xod'by. Pyat' ulic minovat'.
Po lestnice na samyj verx spustit'sya
I v dver' uslovnym stukom postuchat'...

2.4 O, dojti by tol'ko domoj! Pyatnadcat' minut xod'by. Minovat' pyat' ulic. *Spustit'sya na samyj verx po lestnice* i postuchat' uslovnym stukom v dver'.

("Oh, if I could only get home. It is a 15-minute walk. Cross five streets. *Climb down the stairs to the very top* and knock on the door with a secret knock."—Verse and prose fragments with an error introduced: *climb down* instead of *climb up*)

3.1 Pod oknom shumyat i mechutsya
Vetki klenov i berez...
Bez ulybok my vstrechalisya
I rasstanemsya bez slez.

3.2 Pod oknom mechutsya i shumyat *vetki klenov i berez*...My bez ulybok vstrechalis'i bez slez rasstanemsya.

("Under the window the *branches of the maples and birches* move and rustle. We met without smiling and will part without tears."—Original verse fragment and its prose counterpart)

3.3 Pod oknom shumyat i mechutsya
Kleny vetok i berez...
Bez ulybok my vstrechalisya
I rasstanemsya bez slez.

3.4 Pod oknom mechutsya i shumyat *kleny vetok i berez*... My bez ulybok vstrechalis' i bez slez rasstanemsya.

("Under the window the *maples of the branches and birches* move and rustle. We met without smiling and will part without tears."—Verse and prose fragments with an error introduced: *maples of the branches* instead of *branches of the maples*)

4.1 Iz-za tuch luna katitsya —
Chto zhe? Golyj pered nim:
S borody voda struitsya,
Vzor otkryt i nedvizhim.

4.2 Luna katitsya iz-za tuch — chto zhe? Pered nim golyj: *s borody struitsya voda*, otkryt i nedvizhim vzglyad.

("The moon appears from behind the clouds. What is this? Someone naked is before him, *water pours from the beard*, the eyes are open and fixed."—Original verse fragment and its prose counterpart)

4.3 Iz-za tuch luna katitsya —
Chto zhe? Golyj pered nim:
Boroda s vody struitsya,
Vzor otkryt i nedvizhim.

4.4 Luna katitsya iz-za tuch — chto zhe? Pered nim golyj: *boroda s vody struitsya*, otkryt i nedvizhim vzglyad.

("The moon appears from behind the clouds. What is this? Someone naked is before him, *the beard pours from the water*, the eyes are open and fixed."—Verse and prose fragments with an error introduced: *the beard pours from the water* instead of *the water pours from the beard*)

We also prepared a number of "filler" texts, i.e. four-line verse fragments and their prose versions. These fragments contained no errors. Each experiment involved a total of sixteen (16) fragments: eight (8) stimulus fragments of which four (4) contained errors (the experimental condition) and four (4) were error-free (the controlled condition; see below) as well as eight (8) filler fragments. As a result, the share of fragments with errors was relatively small. In this way, we also avoided drawing too much attention to the errors.

Text type	Condition	Average reading time (sec)
prose	control[1]	6.6 (1.5)[2]
	experimental[1]	7.9 (1.9)
	filler	6.6 (2.0)
poetry	control	7.8 (2.0)
	experimental	7.7 (1.7)
	filler	7.3 (1.8)

Table 1: Average reading times for prose and verse fragments with and without an error (Experiment I, 40 informants)

We used the "Latin square" principle, which is often applied in experiment design. This involved comparing texts with and without an error on the one hand, and verse and prose texts on the other. The group of texts in each experiment served as a control for the other two. In each of the three experiments, the informants were divided into four groups. Each group received its own set of texts so that a text presented to one group with an error was shared error-free with another group.

1. 4 verse texts from set I with errors + 4 verse texts from set II without errors + 8 verse fillers
2. 4 verse texts from set II with errors + 4 verse texts from set I without errors + 8 verse fillers
3. 4 prose texts from set I with errors + 4 prose texts from set II without errors + 8 prose fillers
4. 4 prose texts from set II with errors + 4 prose texts from set I without errors + 8 prose fillers

Groups of informants received either verse or prose fragments. If verse did suppress logical thinking, we did not want to turn this mechanism on and off by mixing verse and prose fragments.

3 Results

We analyzed reading times for the fragments. In all three experiments, reading times that exceeded the mean plus three standard deviations (per condition) were excluded as outliers. As a result, a very small percentage of the data was excluded (this never exceeded 6%). Table 1 shows the average reading times for texts with and without an error for both verse and prose fragments in Experiment I.

[1] Control condition: original texts with no errors; experimental condition: original texts with an error introduced.
[2] The standard deviation from the mean is given in parentheses.

Text type	Condition	Average reading time (sec)
prose	control	10.6 (1.5)
	experimental	12.9 (1.7)
	filler	11.6 (1.6)
poetry	control	11.2 (2.1)
	experimental	10.8 (1.8)
	filler	10.7 (1.9)

Table 2: Average reading times for prose and verse fragments with and without an error (Experiment II, 40 informants)

Text type	Condition	Average reading time (sec)
prose	control	9.3 (1.4)
	experimental	11.2 (1.5)
	filler	10.6 (1.6)
poetry	control	10.4 (1.5)
	experimental	10.2 (1.5)
	filler	9.8 (1.4)

Table 3: Average reading times for prose and verse fragments with and without an error (Experiment III, 30 informants)

We used RM ANOVAs to estimate whether the differences between the experimental and control conditions were statistically significant. Fragment reading time was treated as the dependent variable. According to this statistical analysis, the difference between the conditions was statistically significant for prose ($F(1, 192) = 3.69, p = 0.05$) but not for verse ($F(1, 182) = 0.02, p = 0.88$). This means that errors in prose were noticed and caused a significant slowdown, while in verse they remained undetected. The same pattern was found in Experiment II (Table 2).

The difference in reading times for the experimental and control conditions was statistically significant for prose ($F(1, 150) = 4.47, p = 0.04$) but not for verse ($F(1, 150) = 0.11, p = 0.75$). We again saw a significant slowing of the reading time for prose, which means that errors were noticed; in contrast, in verse, they remained undetected. The same picture was observed in Experiment III (Table 3).

For the prose texts, the difference between the experimental and control conditions was statistically significant ($F(1, 109) = 4.99, p = 0.03$), while for the verse ones, it was not ($F(1, 112) = 0.21, p = 0.88$). Again, errors in prose caused a considerable slowdown but this was not the case for verse. This shows that errors were discovered in the former but not in the latter. The observation of the same general trend in all three experiments is important proof of its validity.

We have recently completed another experiment that suggests that the same phenomenon applies when reading Turkish verse and prose with and without errors. We have selected Russian and Turkish verse as data points relatively dis-

tant from one another in order to prove that the influence comes from the verse structure itself irrespective of verse type, language type and cultural traditions. The Russian verse used in our experiments includes syllabic-accentual and non-classical verse derived from the syllabic-accentual system. Moreover, the Russian language is inflectional. The Turkish verse is syllabic, and the Turkish language is agglutinative. Russian and Turkish cultural traditions significantly differ. Still, our experiments suggest that in both cases, the effect of verse on cognitive processes differs from that of prose in the same way.

4 Conclusion

Our findings show that errors in verse remain undetected while the same errors are easily discovered in a prose counterpart text. We suggest that prose and verse differently activate two ways of processing information in the brain; the first way is logical and relies on critical thinking including error detection, while the second is associative and depends on mental imagery rather than sequential logic.

Acknowledgments

The study was enabled by RFBR grant 19-012-00534.

References

Clifton, Charles Jr., Lyn Frazier, and Patricia Deevy (1999). “Feature manipulation in sentence comprehension”. In: *Rivista di Linguistica* 11, pp. 11–39.

Dillon, Brian, Alan Mishler, Shayne Sloggett, and Colin Phillips (2013). “Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence”. In: *Journal of Memory and Language* 69.2, pp. 85–103. DOI: 10.1016/j.jml.2013.04.003.

Drummond, Alex, Titus von der Malsburg, Michael Yoshitaka Erlewine, and Mahsa Vafaie (2016). *Ibex farm*. URL: https://github.com/addrummond/ibex.

Just, Marcel A. and Patricia A. Carpenter (1980). “A Theory of Reading: From Eye Fixations to Comprehension”. In: *Psychological Review* 87, pp. 329–354. DOI: 10.1037/0033-295X.87.4.329.

Just, Marcel A., Patricia A. Carpenter, and Jacqueline D. Wooley (1982). “Paradigms and processes in reading comprehension”. In: *Journal of Experimental Psychology: General* 3, pp. 228–238. DOI: 10.1037//0096-3445.111.2.228.

Kostyuk, Alexander Edmundovich (2017). “Prosodiya stroki v russkom stixe”. In: *Trudy instituta Russkogo yazyka imeni V.V.Vinogradova* 14, pp. 27–41.

Krivnova, Olga Fedorovna and Alexander Edmundovich Kostyuk (2020). "Ramochnaya tonal'naya konstrukciya v foneticheskoj strukture stixa i prozy". In: *Vaprosy yazykoznaniya: Megasbornik nanostatej. Sb. st. k yubileyu V. A. Plungyana*. Ed. by Andrej Aleksandrovich Kibrik, Kseniya Pavlovna Semenova, Dmitrij Vladimirovich Sichinava, Sergej Georgievich Tatevosov, and Anna Yur' evna Urmanchieva. Moskva: Buki-Vedi, pp. 86–98.

Pearlmutter, Neal J., Susan M. Garnsey, and Kathryn Bock (1999). "Agreement Processes in Sentence Comprehension". In: *Journal of Memory and Language* 41.3, pp. 427–456. DOI: 10.1006/jmla.1999.2653.

Proxorov, Aleksandr Oktyabrinovich, ed. (2011). *Psixologiya sostoyanij*. Moskva: Cogito-center.

Skulacheva, Tatyana Vladimirovna (1996). "Lingvistika stixa: struktura stixotvornoj stroki". In: *Slavyanskij stix: stixovedenie, lingvistika i poe'tika*. Ed. by Mixail Leonovich Gasparov and Tatyana Vladimirovna Skulacheva. Moskva: Nauka, pp. 18–23.

Skulacheva, Tatyana Vladimirovna (2004). "Stix i proza: semanticheskie razlichiya". In: *Slavyanskij stix VII: lingvistika i struktura stixa*. Ed. by Mixail Leonovich Gasparov and Tatyana Vladimirovna Skulacheva. Moskva: Yazyki slavyanskoj kul'tury, pp. 167–178.

Skulacheva, Tatyana Vladimirovna (2014). "Verse and Prose: A Linguistic Approach". In: *Poetry and poetics: a centennial tribute to Kiril Taranovsky*. Ed. by Barry P. Sherr, James Bailey, and Vida T. Johnson. Bloomington (Indiana): Slavica, pp. 239–248.

Skulacheva, Tatyana Vladimirovna (2016). "Struktura stixa i ego vospriyatie". In: *Sed'maya mezhdunarodnaya konferenciya po kognitivnoj nauke*. Ed. by Aleksandrov Yurij Iosifovich and Anoxin Konstantin Vladimirovich. Moskva: Institut psixologii RAN, pp. 545–546.

Skulacheva, Tatyana Vladimirovna and Mariya Valeryevna Buyakova (2010). "Stix i proza: sochinenie i podchinenie". In: *Voprosy yazykoznaniya* 2, pp. 37–54.

Skulacheva, Tatyana Vladimirovna and Alexander Edmundovich Kostyuk (2018). "Stix: lingvisticheskie mexanizmy i ix funkcii". In: *Vos'maya mezhdunarodnaya konferenciya po kognitivnoj nauke. Tezisy dokladov*. Ed. by Andrej Konstantinovich Krylov and Valerij Dmitrievich Solov' ev. Moskva: Institut psixologii RAN, pp. 930–932.

Skulacheva, Tatyana Vladimirovna and Alexander Edmundovich Kostyuk (2019). "Verse and Prose: Linguistics and Statistics". In: *Quantitative Approaches to Versification*. Ed. by Petr Plecháč, Barry Paul Scherr, Tatyana Vladimirovna Skulacheva, Helena Bermúdez-Sabel, and Robert Kolár. Prague: ICL CAS, pp. 245–254.

Skulacheva, Tatyana Vladimirovna and Alexander Edmundovich Kostyuk (2020). "Lingvisticheskie osobennosti stixa i ix funkcii". In: *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2, Yazykoznanie* 19.3, pp. 155–168. DOI: 10.15688/jvolsu2.2020.3.14.

Wagers, Matthew W., Ellen F. Lau, and Colin Phillips (2009). "Agreement attraction in comprehension: Representations and processes". In: *Journal of Memory and Language* 61.2, pp. 206–237. DOI: 10.1016/j.jml.2009.04.002.

Yanko, Tatyana Evgenyevna (2010). “Prosodiya predlozhenij so «snyatoj» illokutivnoj siloj”. In: *Komp'yuternaya lingvistika i intellektual'nye texnologii: Tr. Mezhdunar. konf. po komp'yuternoj lingvistike i ee prilozheniyam “Dialog-2010”* 9.16, pp. 609–621.

Yanko, Tatyana Evgenyevna (2015). “Lingvisticheskie texnologii analiza zvuchashhego poe’ticheskogo teksta”. In: *Kognitivnye issledovaniya yazyka* 23, pp. 174–183.

Yarxo, Boris Isaakovich (2006). *Metodologiya tochnogo literaturovedeniya: Izbrannye trudy po teorii literaturovedeniya*. Moskva: Languages of the Slavic Cultures.

https://doi.org/10.51305/ICL.CZ.9788076580336.11
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

On Digital Comparative Editions and Textual Similarity Detection Tools: Towards a Hypertextual Cartography of a Rewritten Myth

Karolina Suchecka
ALITHILA, ULR 1061
Université de Lille, France
karolina.suchecka@univ-lille.fr
0000-0002-6485-7033

Nathalie Gasiglia
STL, UMR 8163
Université de Lille, France
nathalie.gasiglia@univ-lille.fr
0000-0001-7149-2274

Abstract

Our project aims to expose the intertextual relationships observable within a heterogeneous literary corpus. For this purpose, we examine the output of two text reuse detection tools, `Tracer` and `TextPAIR`. We suggest some solutions to overcome the specific limitations observed in those tools and to enhance data quality. We believe that automatic analysis of the rewriting process can make it more comprehensible if the analysis is combined with empirical research methods adapted to the corpus in question.

1 Introduction

Developments in computational technology and digital corpus accessibility are helping to improve the identification of similar passages, phrasing strategies, and references in different literary texts.[1]

In this article, we report on an investigation of rewritings and intertextual references within a corpus of French literary texts about the myth of Orpheus and Eurydice. Our aim is not only to identify text reuse but also to create a hypertextual and modular cartography to represent this phenomenon. At the same time, the heterogeneity of the corpus, which we discuss in Section 2, presents significant problems. Although the texts are closely related thematically, we observe numerous differences in their narrative structure and vocabulary. As works in the corpus were published at various points between the 15th and the 21st century, they include diverse forms of French, and thus, also very different constructions and lexical forms.

[1] Cf. for example Barzilay et al. (2001), Coffee et al. (2012), and Ganascia et al. (2014).

To detect similarities, we use two tools: `Tracer`[2] and `TextPAIR`.[3] Our study compares their results and capacities regarding literary texts. We first describe these tools briefly in Section 3. In Section 4, we then highlight some specific processing challenges related to the preparation of the corpus and the relevance and format of the results. Finally, in Section 5, we suggest a system that combines, reworks, and enriches the relationships detected. From morphosyntactic annotation to the word-level matching stage, we outline our method for adapting the general treatments offered by these tools to the specific requirements of our corpus. This method also focuses on several marked-out factors that can be interconnected.

2 Introducing the Corpus

Of the more than 30 representations of the myth of Orpheus and Eurydice in ancient Greek and Roman literature, *The Georgics* by Virgil (37–30 BC) and *The Metamorphosis* by Ovid (1st AD) are the most frequently reused in subsequent rewritings.[4]

Although these texts differ significantly, the main plot of the myth remains the same: inconsolable after the death of his wife, Eurydice, Orpheus descends into Hell and begs the deities to bring her back to life. Thanks to the beauty of his song, they grant him this favour, provided that he does not look back at Eurydice while she follows him out of the Underworld. Unfortunately, on reaching the surface, he turns around and loses her forever.

In addition to several French translation series,[5] our corpus contains more than 70 rewritings[6] that were created in different genres and during different periods.[7] Proximity to the canonical plot of the myth varies: while theatrical adaptations may be quite closely related, other works, especially poems and modern rewritings, only allude implicitly to the myth. The content and motifs used and their (re)interpretation also differ from one work to the next. Parodies, for example, often satirise Orpheus' music or his intense love for Eurydice. On

[2] Cf. https://www.etrap.eu/research/tracer/ (eTRAP Project, University of Göttingen) and Büchler (2013) and Büchler et al. (2012).

[3] Cf. https://artfl-project.uchicago.edu/text-pair (ARTFL Project, University of Chicago) and Allen et al. (2010) and Horton et al. (2010).

[4] This paper does not discuss the specific relationship between Ovid's and Virgil's versions of this myth, but it is important to note that these versions are not independent; Ovid deliberately alludes to and revises Virgil's text, cf. Kushner (1961) and Segal (1988).

[5] By "French translation series", we mean a subset of French translations of the same ancient source (either Ovid's *The Metamorphosis* or Virgil's *The Georgics*), which date, however, from different periods and reflect different translation practices (parallel, literary, adapted, etc.).

[6] Different theories provide different definitions of *intertextuality* and *rewriting*, especially in the mythological context (cf. Gignoux (2006) and Schnyder (2008)). This study does not discuss the complexity of these definitions and instead divides our corpus somewhat roughly into two subsets: (1) *translation series*, directly assigned to Ovid or Virgil and (2) *rewritings*, i.e. heterogeneous works that refer to the myth (adaptations, parodies, etc.).

[7] The texts in our corpus are available on https://github.com/karolinasuchecka/orphidys. We have also published our scripts there but please note that we are not yet at the final stage of our project and so the documentation is incomplete, particularly in English. We would be happy to receive any feedback and hope that developments in this study, which we will continue to update, will prove relevant and useful for other projects.

the other hand, modern rewritings introduce amalgams, new characters, and scenery that modify the meaning of the myth (Brunel 1997).

To counterbalance the diversity of these rewritings, it is helpful to begin by analysing the reuses detected within translations of Ovid and Virgil respectively and noting similarities and variations in these instances. Since they are very closely related, translation series of ancient sources have the potential to establish a common foundation and, at least indirectly, link different rewritings of the myth. These series, thus, seem to warrant special attention when developing a method to benchmark and improve detection results both for translations and rewritings. We hope to show that translation series play a crucial role in the adaptation of general treatments to the specific requirements of our corpus. For this purpose, we start by observing the initial detection results of `TextPAIR` and `Tracer`, which we introduce below.

3 Introducing the Tools

`TextPAIR` enables the detection of reuse based on a fairly adaptive parameterisation process. As such, the user can choose the minimum number of common words to be detected, submit a list of words to ignore, and determine whether the matching algorithm should take into account words, lemmas, or stemmas (word roots). The tool employs the sequence analysis techniques that are applied, for instance, to detect plagiarism. Initially, `TextPAIR` generates overlapping word sequences (*n-grams*)[8] for each text. It then compares these results with those from sequences in other texts.[9]

In contrast, `Tracer` requires the corpus to be submitted in text format and tokenised into syntactic units (sentences in the case of our corpus). Each unit obtains a unique ID that permits the regrouping of sentences belonging to the same text.[10] The parameters allowed include matching based on lemmas, synonyms, or word embedding. The tool also calculates a score that reflects the proximity between two segments.

Our treatment attempts to reconcile the parameterisation approaches of the two tools. This entails generating tri-grams and using lemmatisation, flattened accents, and minimum three-word matching as well as ignoring word order.

4 Processing the Corpus

As our corpus has been assembled not only for processing with reuse detection tools, but also for presentation within a digital scholarly edition, it is structured according to the XML-TEI P5 standard. The genre of each piece is taken into ac-

[8] Cf. Jurafsky et al. (2009) and, for applications to intertextuality detection, Forstall, Coffee, et al. (2015).

[9] Cf. https://github.com/ARTFL-Project/text-pair.

[10] For text-format corpus preparation and required segment ID formatting, cf. https://tracer.gitbook.io/-manual/manual/corpus-preparation.

count; a common mark-up vocabulary is used to digitise all subgenres, whether they are plays, opera librettos, poetry, or novels.[11]

4.1 Preparing the Corpus

While `TextPAIR` accepts XML format without exploiting mark-ups, `Tracer` requires plain text. The latter leads to the loss of marked-up enhancements which might have improved the quality of the results.

Furthermore, `Tracer` requires the corpus to be split into syntactic units which are then assigned identifiers. Some programming skills are needed for automatic sentence splitting, and the results should be reviewed manually at least for a French literary corpus like our own. Plays and poems seem to be particularly hard to process since they are full of interjections and follow specific capitalisation rules.[12] Processing is, thus, limited to matches between sentences only. In contrast, `TextPAIR` does not impose this initial preparation requirement and can therefore also detect similarities across sentence boundaries.

4.2 Relevance of the Detected Pairs

To assess the relevance of reuse detected by these tools, a sample of the results is evaluated by a human reader. The initial processing is performed without linguistic enhancements except for lemmas extracted from the corpus annotation by the `TreeTagger`[13] tool. Five hundred pairs are, thus, randomly retrieved from the results for each of `Tracer` and `TextPAIR`.[14] To evaluate the relevance of each match, we follow the 5-point grading schema proposed by Coffee et al. (2012, p. 392).[15] Below we describe all of the potential types of results:

Type 1 A false match caused by bad parameterisation of the tool or failure to adapt the corpus during preparation.

Type 2 A false or irrelevant match based on stop-words (articles, auxiliaries, etc.), very common constructions, or different contexts.

Type 3 A match based on lexical words. This is affected by different contexts, difficulties in evaluating the relationship effectively, or references to different episodes of the myth.

[11] Cf. TEI Guidelines, https://tei-c.org/guidelines/p5/.

[12] Since optimal splitting improves the quality of results, we ultimately decided to enrich the corpus with specific mark-up (`<milestone>`).

[13] Cf. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[14] We retain the same number of pairs for each tool rather than assigning it a percentage of the total number of detected relationships. This is because the fact that `Tracer` detects more reuse than `TextPAIR` (10 414 to 8851) does not imply that Tracer's cases are more relevant or should be more present within the evaluation sample.

[15] This schema is adapted to meet the specific requirements of our corpus. While Coffee et al. (2012) proposes general criteria related to formal similarity and context proximity, we also evaluate references to the same episodes in the myth, especially for matches of types 3–5.

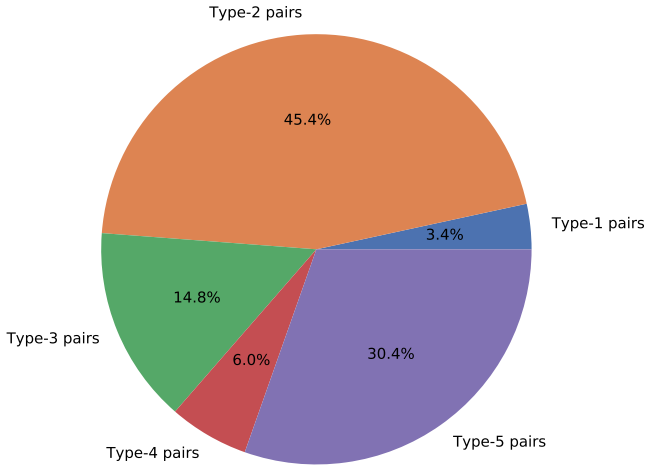


Figure 1: Human reader's evaluation of the results sample

Type 4 A relevant match where the two excerpts refer to the same episode. There are few identical words and expressions are different. The parallel is largely based on allusions.

Type 5 A relevant match that refers to the same episode. The majority of words and expressions are similar or identical.

As noted in Figure 1, irrelevant matches (types 1–2) represent 48.8 % of the evaluation sample, and type-2 pairs comprise the vast majority of these cases.[16] The analysis of these instances reveals several difficulties.

First, even though `TextPAIR` can detect long common segments, the results are most likely to be dominated by noisy data, especially for plays whose stage directions are filled with named entities (designations of speakers in the dialogue, etc.). The results are, thus, often false since the matching is only due to the presence of these named entities (Table 1).[17]

The majority of type-2 matches (259/454) arise within the `Tracer` results. Indeed, while the tool can filter certain words based primarily on their frequency[18] it does not allow us to ignore stop-words only. It is also not possible to exclude the most frequent words from our corpus (even if that would allow us to discard stop-words) since crucial elements for our research, particularly named entities, would then also be ignored.

[16] Type-1 pairs mainly reflect the issues described in Section 4.1. In the sections that follow, we therefore focus exclusively on pairs of types 2–5.

[17] All French excerpts are our own translations. We highlight common words in both the original quotes and their translations.

[18] The value of the minimum or maximum word frequency to be ignored is not customisable, however, cf. `Tracer` *Manual*, "Step 3: Selection. § Selection strategies" (https://tracer.gitbook.io/-manual/manual/configuration/step-3.-selection).

Offenbach et al. (1858)	Ovide (1702, trans. Duryer)
Eurydice : Une heure un quart ! Orphée : Au moins. Eurydice : Je n'écouterais pas !	Eurydice s'évanouit, et le malheureux Orphée n'embrassa que de l'air [...]. Cependant Eurydice qui mourut alors pour la seconde fois [...]
(Eurydice: An hour and a half! Orpheus: At least. Eurydice: I won't listen!)	(Eurydice vanishes and the unfortunate Orpheus embraces nothing but the air [...]. However, Eurydice dies a second time [...])

Table 1: False matching based on stage directions

Finally, while almost all type-5 matches are established within translation series (100 % for `Tracer` and 93 % for `TextPAIR`), rewritings mostly occur among the matches for types 3 and 4. For type-3 pairs, the contexts provided by both tools are not always sufficient to evaluate relevance. We therefore initially deal with relatively few pairs that contain rewritings (71) where we can be certain that the two parts are effectively related.

4.3 Output Formats

Another problem that we encounter is more formal in nature: working simultaneously with two differently designed tools creates the need for a treatment that can overcome these differences in order to arrive at comparable and compatible results. In addition, these tools generate an output consisting of either a list of properties associated with a value (Figure 2) or a selection of these values in a tabulated file[19]. These formats are machine-readable but difficult to interpret by a human.

Besides the data allowing the identification of each segment, no details are provided for the linguistic elements that enable matching. We would have liked to know, for example, which words are aligned and on what basis (common lemmas or stemmas or synonyms) as well as the number of common entities and the distance between them (i.e. the number of words that separate each common entity from the next one).

5 Post-Processing Results

To determine the optimal method for benchmarking results, we process the evaluation subset (Section 4.2) with a basic algorithm that detects unit-level

[19] Cf. *Tracer Manual*, "Results & computed files", https://tracer.gitbook.io/-manual/manual/results-and-computed-files.

```
{
    "source_passage": "Eurydice qui mourut alors pour la seconde fois par la
faute de son mari, ne s'en plaignit point en mourant ; et de quoi eût-elle pû
se  plaindre si ce n'étoit d'être trop aimée ? Elle lui dit seulement le
dernier adieu d'une voix foible, et qu'il ne pût presque entendre, et retomba
dans le gouffre d'où il venoit de la retirer. Orphée ne demeura pas moins
étonné de cette seconde mort de sa femme, que ce mal-heureux Berger qui vit
Cerbere",
    "source_author": "Ovide",
    "source_title": "Les Métamorphoses d'Ovide en latin et françois, divisées
en XV. Livres",

    "target_passage": "Mourant pour la seconde fois, elle ne se plaignit
point de son époux ; car de quoi eût-elle pû se plaindre, sinon qu'il l'avoit
trop aimée. Elle lui dit le dernier adieu d'une voix foible, et qu'il ne put
entendre qu'avec peine. Elle fut engloûtie pour la seconde fois dans le même
abîme dont il venoit de la retirer. Orphée demeura autant étonné de cette
seconde mort de sa femme, que le fut autrefois ce berger voyant Cerbere",
    "target_author": "Ovide",
    "target_title": "Les metamorphoses d'Ovide, avec des explications à la
fin de chaque fable"
}
```


Figure 2: Extract of `TextPair` results format

lexical relations (common forms or lemmas). We proceed to perform automatic synonym detection.[20]

We first observe that in order to increase the number and relevance of detected pairs, there is a need for either a more complex NLP pipeline that exploits the taggers trained on each French language variety, or manual correction of the POS-tagging results. Incorrect lemmas produce many false alignments and hinder the effective detection of common linguistic elements, especially among texts in old and modern French.

Synonym detection also produces a lot of irrelevant matches, especially for polysemous verbs such as *retourner* (to turn around, to return). Nevertheless, synonymy appears to have real potential for improving results. Translations in verse are, for example, less connected with their subset since they privilege synonyms, polylexical constructions, and circumlocutions and have to respect rhyme and rhythm structures. But these characteristics are precisely what enable more subtle matches with the subset of rewritings: Ovide (1687, trans. Corneille) is a unique translation of Ovid that relates to a comic opera by Offenbach et al. (1858). A comparison of these texts shows that among the 18 common entities detected within 4 couples, 10 are pairs of synonyms.

`Tracer` can take synonymy into account to detect reuse.[21] It seems, however, that to fully exploit this functionality, general language synonyms should first be adapted to the specific requirements of the corpus.

Finally, among type-5 pairs, the same words, especially modifiers of sentence constituents, are not always correctly aligned ("*il ne porte ni visage serein ni présage heureux*" [he does not have a serene face nor reveal any happy omen] / "*il n'apporte ni parole rituelle ni visage heureux*" [he brings neither a ritual word

[20] Synonym lists were first extracted from a cumulative synonym dictionary for general language (the *Dictionnaire Électronique des Synonymes*, Crisco, Université de Caen, cf. https://crisco2.unicaen.fr/des/).

[21] Cf. *Tracer Manual*, "Pos-tagging, lemmatisation and WordNets", https://tracer.gitbook.io/-manual/manual/pos-tagging-lemmatisation-and-wordnets

nor a happy face]). In order to discard irrelevant matches and improve the detection of common units, it therefore seems necessary to delimit lexical groups. The integration of sentence constituents at the post-processing stage may also enhance the alignment of descriptive paraphrases (*"les âmes nouvelles"* [the new souls]/ *"les ombres arrivés récemment"* [the shadows recently appeared]).

Based on these observations, we develop a three-step post-processing chain that aims to discard irrelevant matches and detail the common linguistic elements of each sufficiently-related pair. First, we convert the outputs obtained with `TextPAIR` and `Tracer` into an enriched format (Section 5.1). We then perform a detailed analysis of each pair to detect the common entities and determine their number and degree of proximity. We calculate a new similarity score and then exclude insufficiently related pairs (Section 5.2). Finally, we integrate the results and enhance the annotations into an XML file (Section 5.3).

5.1 Results Compilation

To compile the results of each treatment, we take advantage of the division of the corpus into the sentences imposed by `Tracer` at the pre-processing stage. Each sentence is then enriched with a multi-level annotation, both for word groups and for words alone. Consider, for example, the case of Figure 3, a fragment of sentence 28 from `duryer1702`.[22]

Contained within the `<s>` element, the identifier of this sentence is provided with the `@xml:id` attribute while the IDs of matching sentences are included in `@corresp`. The nominal, verbal, and adjectival groups are marked as `<phr>` with the `@select` that supplies the lemma of the headword. *"La faute de son mari"* (the fault of her husband) and *"son mari"* (her husband) are marked as nominal groups.[23] The second group nested in the first one is also marked as a named entity *Orphée* using the `<persName>` element, which provides the ID of that entity with `@corresp`. As for word annotation (`<w>`), after manually correcting the POS-tagging results, we propose grammatical (`@pos`) and inflexional (`@msd`) codes, lemmas (`@lemma`), and a selection of synonyms (`@sameAs`).[24] The inflexional codes and synonyms are not provided for stop-words.

5.2 Searching for Lexical Relations

The algorithm that searches for lexical relations compares each pair of reused elements that was detected by at least one of the tools. It takes into account the enrichment provided through the attribute values and performs multi-level matching.

[22] "[...]*Eurydice qui mourut alors pour la seconde fois par la faute de son mari* [...]" [Eurydice, who dies for the second time through the fault of her husband] (Ovide 1702, trans. Duryer).

[23] In fact, "*la faute de son mari*" and "*son mari*" are part of prepositional groups introduced by *par* and *de* respectively.

[24] DES (footnote 20) classifies synonyms in order of their score, which is thought to represent proximity to the headword (cf. "§ L'ordre des synonymes", *Présentation du DES*, http://crisco.unicaen.fr/dictionnaire-electronique-des-synonymes/presentation-du-des/). For our annotation, we initially choose the first 4 synonyms.

```
<s xml:id="2400028" corresp="3800039 7400048 1600020 6900027 2600008
  2000041 1700026">[...]
    <phr type="GN" select="faute">
        <w xml:id="2400028_11" n="11" lemma="le" pos="DET:ART">la</w>
        <w xml:id="2400028_12" n="12" lemma="faute" pos="NOM" msd="fs"
        sameAs="erreur bévue manquement pêché">faute</w>
        <w xml:id="2400028_13" n="13" lemma="de" pos="PRP">de</w>
        <persName corresp="orphee" type="périphrase">
            <phr type="GN" select="mari">
                <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
                <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM" msd="ms"
                sameAs="époux conjoint homme">mari</w>
            </phr>
        </persName>
    </phr>
[...]
</s>
```

Figure 3: Mark-up of an extract of sentence 28 from duryer1702

First, only the lemmas provided as head of sentence constituents are taken into account and we calculate the ratio of exact lexical identity or synonym equivalence.[25] The processing of the evaluation sample shows (Figure 4) that even though some type-2 pairs obtain relatively high ratios (greater than 40 %), in the majority of cases (282/454) less than 30 % of headwords can be matched. For the type-5 pairs, the ratio of only 33 of the 303 pairs is less than 30 %.

The roughly equal distribution of ratios for the type-4 pairs implies that the processing of headwords alone is not sufficient. A more specific adaptation of linguistic resources (lists of synonyms adapted to the corpus, keywords, named entities, and their circumlocutions, etc.) may improve the relevance of headword matching and scoring. For now, and given that the initial results are mainly explored to enhance future treatments, we decide to discard all pairs of ratios below 30 % since irrelevant synonym matches may inflate the score.

When applying our method to sentence 28 of duryer1702, we observe that TextPAIR and Tracer detect very close relationships with 9 translations of Ovid. One reuse is also detected with a novel by Ballanche (1809) ([C], Figure 5).

For sentence A from bellegarde1701 (Ovide 1701, trans. Bellegarde), the correspondence is almost total with 5 identical headwords (marked in red in Figure 5) (*mourir* [to die], *fois* [time], *plaindre* [to complain], repeated twice, and *aimer* [to love]) and 1 group (marked in green) linked by synonymic equivalence (*mari* [husband]∼*époux* [spouse]). Only 2 groups remain unaligned. As for sentence B from corneille1687 (Ovide 1687, trans. Corneille), we find 5 identical headwords (*mourir*, *fois*, *plaindre*, *mari*, and *aimer*) and 1 derivative relationship (*plaindre*∼*plainte* [complaint]; marked in blue). Although no matching is observed for 8 groups from corneille1687, the threshold of 30 % similarity is exceeded for both correspondences and they accede to further processing. However, this does not extend to sentence C from ballanche1809 (Ballanche 1809) since only 2 out of the 9 groups (22 %) can be matched (*plaindre* is repeated twice).

[25] The number of matching headwords is expressed as a percentage of the total number of headwords within a sentence.

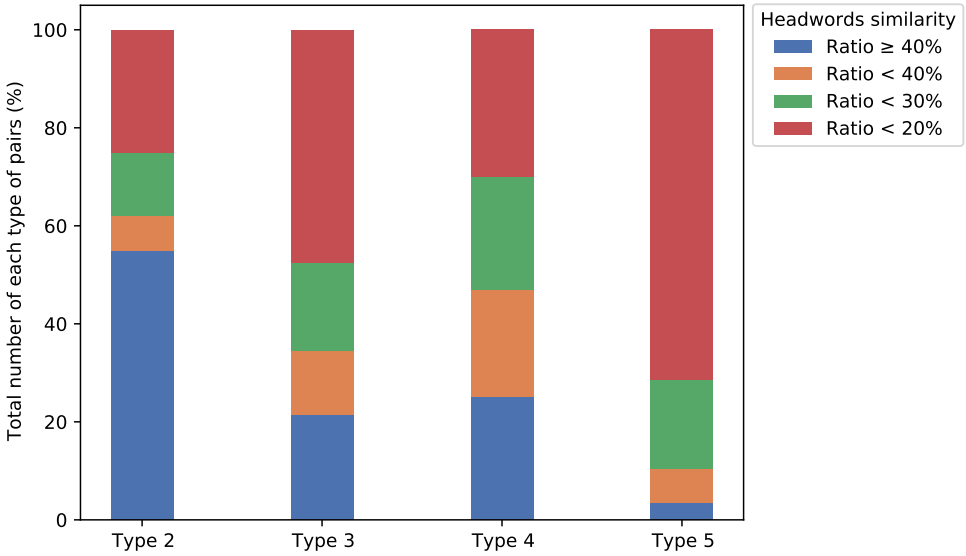


Figure 4: Headwords similarity ratios within evaluation sample

Excluding irrelevant pairs by benchmarking the results is indeed important. But our aim is also to focus on linguistic elements that enable matching and eventually to distinguish the degree of similarity between each pair (quotation, paraphrase, allusion, etc.). To do so, we proceed to word-level matching that excludes stop-words. Currently we consider 3 types of equivalence: (1) unit-level (form∼form, lemma∼lemma, stemma∼stemma), (2) synonym-level, and (3) mixed (form∼synonym, lemma∼synonym).

Several methods can evaluate the lexical proximity between text reuses.[26] Some of these approaches show the relevance of word frequency. However, as all the texts in our corpus share the same mythological theme, the relevant relationships can be based on words of highly variable frequencies. For example, *époux* (spouse) and *mari* (husband) are recurrent periphrases for Orpheus. Their respective frequencies in the corpus are relatively low (146 and 94 out of 432 078), but they appear regularly in the detection results. Indeed, within the evaluation sample, 55 pairs contain the word *époux*, among which 31 are of type 5. However, we also regularly find other keywords from the myth within the type 4–5 pairs derived from the first 100 most frequent words (*amour* [love], *dieu* [god], *mort* [death], *voix* [voice], and *femme* [women/wife]). Therefore, in addition to the headword proximity ratio, we propose a simple lexical proximity assessment method that focuses mainly on the type of equivalence.

First, if a unit-level match is found, 5 points are added to the sentence score and the word pair is no longer taken into account in further processing. For synonym-level and mixed relations, the processing continues in order to achieve optimal matching. Each synonym-level equivalence is worth 0.5 points and each

[26] Cf. for example Büchler (2013, pp. 116–119) and Forstall and Scheirer (2019).

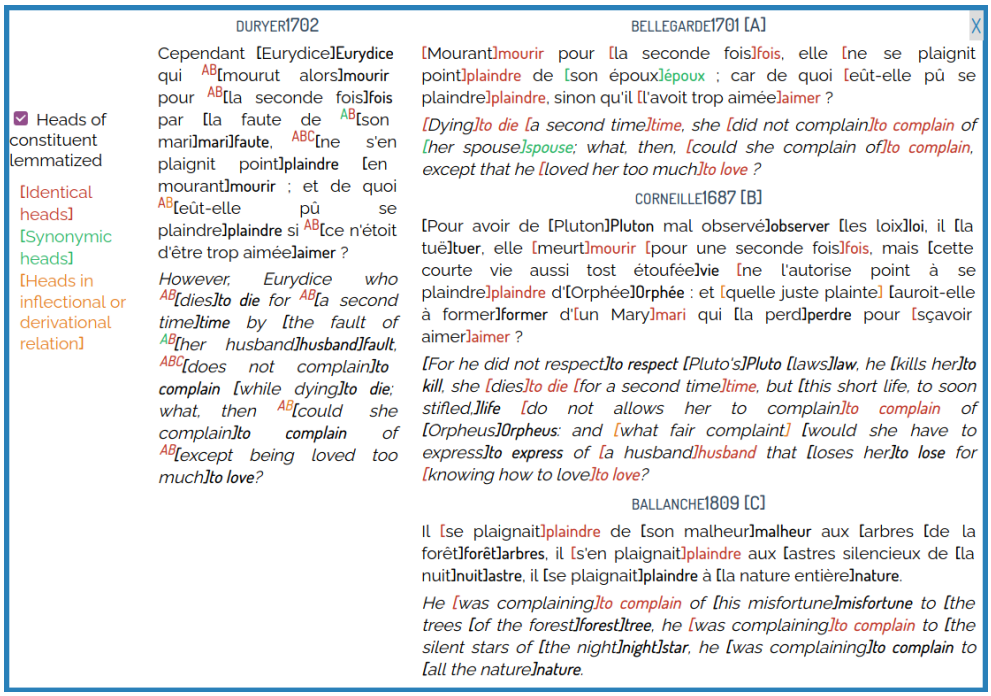


Figure 5: Example of matching headwords

mixed relationship counts for 1.5 points. Points may accumulate if multiple synonym or mixed matches are found. As 4 synonyms are provided for each word, a pair can obtain 3 points maximum (1 mixed and 3 synonym-level matches). The equivalence between *mari* and *époux*, for example, gets 2 points: this reflects 1 mixed match between the lemma *mari* found among the *époux* synonyms, and 1 synonym-level match for *conjoint* (≈marriage partner), which is common to both words. This scoring of synonym matches enables us to distinguish different degrees of proximity. The equivalence of *mari*∼*homme*, for example, accumulates 1.5 points (lemma∼synonym) and *époux*∼*ami* (friend) only 0.5 (synonym *compagnon* [companion]). Currently a synonym match counts for less than a unit-level one since the error rate for the former remains quite high as long as the lexical resource used is intended for general language.

As Figure 6 illustrates, the correspondence with `bellegarde1701`, thus, receives the score of 42 points (8 equivalences on a word-to-word basis and 1 between synonyms). In contrast, the match with `corneille1697` accumulates 40 points (4 same words, 3 same lemmas, and 1 equivalence between stemmas). A moderate difference between the scores for two texts may suggest a comparable degree of proximity. Nevertheless, while both `duryer1702` and `corneille1679` suggest Orpheus' responsibility for the second death of Eurydice ("*par sa faute*" [through his own fault]/ "*il la tuë*" [he kills her]), this relationship seems too complex to be detected automatically. As for "*Cette courte vie aussi tost étoufée*" (this short life, too soon stifled) and "*Pour avoir*

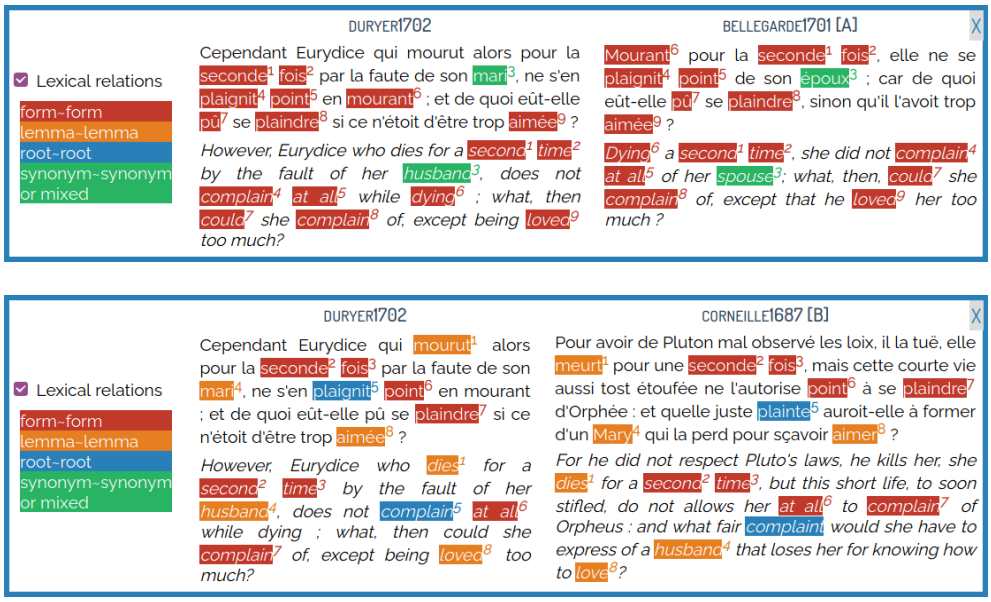


Figure 6: Example of word-level matching

de Pluton mal observé les loix" (for he did not respect Pluto's laws), they are stylistic additions to `corneille1687`, a verse translation.

To compare sentences of different lengths and complexities, we therefore calculate a lexical similarity ratio. This is expressed as a percentage of the maximum score a pair would accumulate if all the lexical words of the shortest sentence were matched at unit-level. The maximum score for `bellegarde1701` is 45 points, and so the relationship with `duryer1702` receives 93.3 % lexical ratio. As for `corneille1687`, the maximum score for the sentence is 110 points. It is, however, the shorter sentence in `duryer1702` that we take into account when calculating the lexical ratio. As 8 words out of 13 match, the lexical similarity of the pair is 65 % (in contrast with only 36 % when we calculate the ratio in `corneille1687`).

Applying this method to our evaluation sample suggests that the threshold for the lexical ratio is lower than the scoring based on sentence constituents.[27] If we assume that only pairs with a lexical similarity above 20 % (and a headwords similarity above 30 %) are relevant, then without losing type-4 and 5 pairs, we can discard 76 additional type-2 pairs. Meanwhile 96/454 remain.

To enable this kind of global analysis, it is important to preserve the enriched results in the adapted format. Such an approach will allow the results to be exploited for information retrieval that can lead, often through traditional analysis, to a significant improvement in the initial results.[28]

[27] For 58 % of type 2 pairs, the ratio is less than 20 and for 15 %, it is between 20 and 30. Only 4 % of type 5 pairs do not exceed the threshold of 20, and 16 % are between 20 and 30.

[28] This also allows for the visualisation of the results, a process that we do not detail here. The examples provided in Figure 5 and Figure 6 are retrieved from an operating interface developed

```
<phr type="GN" select="mari">
    <w xml:id="2400028_14" n="14" lemma="son" pos="DET:POS">son</w>
    <w xml:id="2400028_15" n="15" lemma="mari" pos="NOM"
        sameAs="époux conjoint homme">mari
        <xr corresp="1300018_19" type="lem-syn" cert="2"/>
        <xr corresp="1600020_13" type="lem-syn" cert="2"/>
        <xr corresp="2000041_16" type="lem-syn" cert="2"/>
        <xr corresp="6700029_14" type="lem-syn" cert="2"/>
        <xr corresp="6900027_17" type="mot" cert="5"/>
        <xr corresp="7400048_48" type="lem" cert="5"/>
    </w>
</phr>
```

Figure 7: Enriched mark-up of an extract from sentence 28 in duryer1702

5.3 Enriched Results Exploitation

The results of the word matching are included in the initial annotation of the sentence using an <xr> element, child of <w>, which specifies the ID of the matching word (@corresp), the type of relationship detected (@type), and its score (@cert). In this way, the extract from sentence 28 in duryer1702 presented in Figure 3 is enriched by 6 elements <xr> (Figure 7). Indeed, 6 out of 10 correspondences contain an equivalence with the word *mari*, in most cases based on synonymy with *époux* (4 occurrences).

The enriched results are then included into an XML database that compiles all the matches detected beforehand by TextPAIR and Tracer. This compilation allows for a wide range of manipulation and analyses. Some of these are specific and focused on a chosen text, while others are general and include all the correspondences.

Sentence 28 of duryer1702, for example, can be included in a larger passage that consists of 6 sentences and relates the entire episode of Eurydice's second death, from Orpheus' fatal look to his dismay over his wife's vanishing. Based on the 156 lexical relations found in this excerpt, we can establish at least partial connections with 16 sentences from 11 different texts. Almost half of the matches (71) are found in sentence 28. The most frequently connected words are the proper names of the two lovers, their periphrases, and 3 verbs: *plaindre*, *mourir*, and *aimer*. These could be considered the most salient keywords for this episode if the same trend is confirmed through an analysis of the overall results obtained from other translations of Ovid. For each episode and each variant of the myth, we aim to determine the keywords and their synonyms based on the evidence within the translation series. Using these findings, we hope to improve the results for rewritings. To take one example, the keywords observed for the episode of Eurydice's second death could be exploited to establish more relevant matching between the translation by duryer1702 and the novel by ballanche1809 :

to facilitate exploring specific or complex reformulations. By the end of our project, this interface should be adapted and converted into an open access digital comparative edition.

> Mais, ô faiblesse d'un cœur qui aime ! [...] Vaincu par cette puissance contre laquelle l'homme lutte en vain, Orphée se retourne. [...] Eurydice s'évanouit [...] et sa parole plaintive, inarticulée, meurt dans le vague des airs [...].[29]

Working progressively and starting from the most explicit adaptations of the myth, we, thus, plan to extend our method to increasingly allusive and symbolic connections, parodies, reinterpretations, and modernisations. The objective is to improve the understanding of the role of the linguistic processes observed within these texts.

6 Conclusion

The automatic detection of intertextual relations is an exciting prospect for literary and linguistic researchers as well as for the creators of digital scholarly editions. False detection and recognition problems are integral to this process and confirm the importance of combining distant and close reading. An in-depth understanding of the corpus is essential not only for the analysis of the results, but also for their improvement and enrichment so that new treatments can be applied to the same texts. These new approaches may prove more fruitful and capable of revealing increasingly subtle and surprising relationships.

As part of our project, we are endeavouring to establish some shared practices for preparing digital editions in the humanities. To this end, we employ interoperability, open access data, and usage sharing. In addition, we propose an automated processing method that generates enriched files marked up to the XML-TEI standard. Nevertheless, we cannot claim that our treatment can be generalised easily. It may be adapted with varying success to different corpora with manual adjustments needed to improve the data quality. It will also take significant work and time to obtain the first meaningful results. For some texts, no relevant relationships will ever be found. Still all of this seems to us to be inherent to literary research, whether it is traditional and empirical or supported by computational techniques.

References

Allen, Timothy and Charles Cooney (2010). "Plundering Philosophers: Identifying Sources of the Encyclopédie". In: *Journal of the Association for History and Computing*.

Ballanche, Pierre-Simon (1809). "Sixième Fragment". In: *Œuvres de M. Ballanche*. Paris: J. Barbezat, pp. 488–492.

Barzilay, Regina and Kathleen R. McKeown (2001). "Extracting Paraphrases from a Parallel Corpus". In: *Proceedings of the 39th Annual Meeting of the ACL*. Toulouse: Association for Computational Linguistics, pp. 50–57. DOI: 10.3115/1073012.1073020.

[29] [But, O weakness of the loving heart ! [...] Defeated by this power against which man struggles in vain, Orpheus turns around. [...] Eurydice vanishes and her unarticulated complaints die in the air], Ballanche (1809, p. 491).

Brunel, Pierre (1997). “Orphée Moderne”. In: *Apollinaire Entre Deux Mondes. Le Contrepoint Mythique Dans Alcools. Mythocritique II*. Écriture. Paris: Presses Universitaires de France, pp. 63–82.

Büchler, Marco (2013). “Informationstechnische Aspekte Des Historical Text Re-Use”. PhD thesis. Leipzig: Université de Leipzig.

Büchler, Marco, Gregory Crane, Maria Moritz, and Alison Babeu (2012). “Increasing Recall for Text Re-Use in Historical Documents to Support Research in Humanities”. In: *Theory and Practice of Digital Libraries*. Ed. by P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides. Berlin: Springer Berlin Heidelberg, pp. 95–100.

Coffee, Neil, Jean-Pierre Koenig, Poornima Shakti, Roelant Ossewaarde, Christopher Forstall, and Sarah Jacobson (2012). “Intertextuality in the Digital Age”. In: *Transactions of the American Philological Association* 142.2, pp. 383–422. DOI: 10.1353/apa.2012.0010.

Forstall, Christopher, Neil Coffee, Thomas Buck, Katherine Roache, and Sarah Jacobson (2015). “Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level n-Gram Matching”. In: *Digital Scholarship in the Humanities* 30.4, pp. 503–515. DOI: 10.1093/llc/fqu014.

Forstall, Christopher and Walter Scheirer (2019). “Lexical Matching: Text Reuse as Intertextuality”. In: *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*. Cham: Springer International Publishing, pp. 55–78. DOI: 10.1007/978-3-030-23415-7_3.

Ganascia, Jean-Gabriel, Pierre Glaudes, and Andrea Del Lungo (2014). “Automatic Detection of Reuses and Citations in Literary Texts”. In: *Digital Scholarship in the Humanities* 29.3, pp. 412–421.

Gignoux, Anne-Claire (2006). “De l’intertextualité à La Récriture”. In: *Cahiers de Narratologie. Analyse et théorie narratives* 13. DOI: 10.4000/narratologie.329.

Horton, Russell, Mark Olsen, and Glenn Roe (2010). “Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections”. In: *Digital Studies / Le Champ numérique* 2.1. DOI: 10.16995/DSCN.258.

Jurafsky, Daniel and James H. Martin (2009). “N-Gram Language Models”. In: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New York: Prentice-Hall, Inc., pp. 189–232.

Kushner, Eva (1961). *Le Mythe d’Orphée Dans La Littérature Française Contemporaine*. Paris: A.G. Nizet.

Offenbach, Jacques and Hector Crémieux (1858). *Orphée Aux Enfers*. Paris: Calmann-Lévy.

Ovide (1687). *Les Metamorphoses, Mises En Vers François*. Trans. by Thomas Corneille. Vol. 2-3. Paris: Berthelemy Girin, Michel Brunet.

Ovide (1701). *Les Métamorphoses, Avec Des Explications à La Fin de Chaque Fable*. Trans. by Jean-Baptiste Morvan de Bellegarde. Paris: Michel David.

Ovide (1702). *Les Métamorphoses En Latin et François*. Trans. by Pierre du Ryer. Amsterdam: P. & J. Blaev, Janssons à Waesberge, Boome, & Goethals.

Schnyder, Peter, ed. (2008). *Métamorphoses Du Mythe: Réécritures Anciennes et Modernes Des Mythes Antiques*. Universités - Domaine Littéraire. Paris: Orizons.
Segal, Charles (1988). *Orpheus: The Myth of the Poet*. Baltimore: John Hopkins University Press.

https://doi.org/10.51305/ICL.CZ.9788076580336.12
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

On the Expected and Actual Rhythmical Grammar of Russian Iambic Tetrameter

Kseniya Tver'yanovich

Institute of Russian Literature (Pushkinskij Dom) /
Russian Academy of Sciences, Saint Petersburg, Russia
ksutver@gmail.com
0000-0001-9831-8129

Abstract

The combination of certain rhythmical forms and similar grammatical structures in Russian iambic tetrameter was first noted in the 1920s by scholars such as Sergey Bobrov and especially Osip Brik, who detected its presence across the oeuvres of different authors. Decades later, drawing on his own data about the parts of speech in Russian language and verse, Mikhail Gasparov claimed that rhythmical and grammatical stereotypes occurred in Russian classical meter because each ictus could naturally only accommodate certain words or grammatical forms based on their length or accentual structure. Together with his co-author, Tatiana Skulachyova, Gasparov discovered, for example, that Russian adjectives and verbs tended to be longer than an iambic foot and therefore to occupy ictuses that lacked any metrical stresses. By creating preferred verse locations for certain parts of speech, this also strongly influenced the syntax of the iambic line. This paper considers new data about parts of speech that behave atypically from the standpoint of Gasparov's schema. For some of the authors and periods concerned, longer forms tend to fit into shorter syllabic slots and vice versa. My overview and analysis are based on data regarding iambic tetrameter in the work of two Russian poets from the early 20th century. I conclude that verse is an intricate system in which the rhythmical vocabulary of parts of speech—that is, their typical syllabic length and stress locations—does not necessarily prevail over other important factors. Those factors may ultimately determine the content of rhythmical structures.

1 Introduction

In the late 1910s, Russian scholars began to show a growing awareness that the vocabulary of each iambic foot naturally favored certain accentual and syllabic word types. In his articles on Pushkin's iambic tetrameter (1918) and pentameter (1919–1923), Boris Tomashevskij (2007a,b) presented data about the various rhythmical and syllabic word variations encountered in each foot of the two iambic meters. A few years later, Sergej Bobrov (1922, 1925) argued that the prevalence of certain words in certain parts of iambic verse reflected syntactic and rhythmical stereotypes. Osip Brik advanced similar ideas. In a

famous article on rhythm and syntax, Brik (1927) claimed that the repetition of syntactic structures in Russian iambic tetrameter was not the result of literary borrowings or influences. Rather these patterns were imminent to the structure of the verse itself since the rhythm provided for only limited possibilities or lexical and syntactic structures.

Although initiated around a century ago, this discussion did not produce any significant developments until the publication of the works of Mixail Gasparov in the 1980s, which led to further studies with his co-author, Tatyana Skulachyova. In an influential paper on verse linguistics, Gasparov (1996) applied Boris Yarxo's (1927) thesis that verse consists of a complex of various rhythmical levels. On this basis, he argued that verse theory should focus on the interrelationship of different levels rather than on any isolated trends at specific levels of rhythm. Of all the interconnections, those between accentual and grammatical rhythms attracted most attention since they also affected the vocabulary of verse and especially that of rhyme. Among other fruitful ideas, Gasparov (1984, 2000, 2004) offered a typology of rhythmical and grammatical verse stereotypes and a general explanation of their origins. Gasparov and Skulachyova (2004, pp. 51–61) also presented a "rhythmical dictionary of the parts of speech". This used prose works by Pushkin, Gogol', Turgenev, Tolstoj, and Chexov to estimate the average syllable lengths of different parts of speech and the average locations of stressed syllables in those various parts.

In the Russian language, the average word contains close to three syllables although this figure can range from one to around a dozen. Because the full syllabic count of a word includes not only its root but also any affixes and inflexions, this count depends on the word or word form's grammatical features. The same is true of the location of the stress, which is mobile in Russian. As such, the accentual and syllabic structures of Russian words can often be predicted based on their grammatical features.

According to Gasparov and Skulachyova (2004, p. 63), verbs and adjectives have especially distinct profiles in this respect. Not only are they the longest parts of speech in Russian but they also inversely reflect one another: adjectives tend to bear the stress on one of their initial syllables while verbs, in contrast, tend to carry the stress closer to their ending.

In Russian iambic verse, adjectives and verbs are therefore often expected to occupy pyrrhic feet so as to reserve that space for their long unstressed beginnings or endings. One important difference between these two parts of speech is that while in adjectives, the stressed syllable tends to precede the pyrrhic foot, in verbs, the stress tends to follow the pyrrhic foot. In their work on the grammar of Russian verse, Gasparov and Skulachyova (2004, pp. 62–90) highlight this "inverse relationship" between the accentual structures of adjectives and verbs as a generally expected feature of the iambus.

This paper, however, focuses on examples where this expectation is not met and then looks into possible reasons for that failure. To that end, I focus on the poetry of Anna Axmatova and Osip Mandel'shtam, two prominent authors of the Silver Age, a period of extensive formal experimentation and great revision of 19th-century traditions. The discussion that follows is based on data for the parts of speech found in these poets' use of iambic tetrameter, generally the most

popular and best-studied Russian classical meter. Although the poetry of the Silver Age tended to break with the classical iambic tradition, these two authors often applied it since it was, among other things, a way of contextualizing their own poems and themes.

2 Materials and Methods

Because grammatical structures and relationships may depend on factors such as strophic composition, my research for this paper was limited to quatrains of iambic tetrameter with alternate rhymes. In particular, I considered cases where the feminine clausulae in odd lines alternated with the masculine clausulae in even lines (I4 AbAb).

My selection of appropriately structured poems was based on the Russian National Corpus (RNC). For Axmatova, this produced 26 poems or 360 lines with composition dates ranging from 1910 to 1964. For Mandel'shtam, it generated 33 poems or 564 lines whose composition dated from 1909 to 1937.

For all of these poems, each line was manually annotated based on the following parameters:

- rhythmic form of iambic tetrameter (location of the pyrrhic foot),
- types of word boundaries (number of unstressed syllables following a stress), and
- parts of speech in each stressed foot.

The annotation of parts of speech was based mainly on the universal typology developed by the authors of the Universal Dependencies (UD) project, which relies on Universal Stanford Dependencies (Marneffe et al. 2014). In addition, I used Google's universal tags system for parts of speech (Petrov et al. 2012) and Daniel Zeman's (2008) universal conversion method for tag systems.[1] Nevertheless, a number of modifications were required in order to adapt the system to the task of tracking phonetic words (i.e. content words together with their clitics) with a focus on only those cases where the main stress falls on an ictus. These modifications also needed to ensure compatibility with the results previously obtained by Gasparov and Skulachyova. In particular, the grammatical annotation excluded 1) all function words whatsoever and 2) any content words located between ictuses. Furthermore, in line with Gasparov's approach, Russian participles were tagged as adjectives rather than verbs since the syllabic and accentual structure of participles is generally the same as that of adjectives. This also makes their behavior adjective-like in the iambic environment.

[1] The following abbreviations are used in tables in this paper: PoS = part of speech; ADJ = adjectives, participles, and ordinal numerals; ADP = adpositions; ADV = adverbs; AUX = auxiliary verbs; CCONJ = coordinate conjunctions; DET = pronominal adjectives and numerals; INTJ = interjections; NOUN = nouns; NUM = cardinal numerals; PART = particles; PRON = pronouns; PROPN = proper nouns; SCONJ = subordinate conjunctions; VERB = verbs; X = other.

Author		Rhythmic form 1	2	3	4	5	6	Total
Axmatova	no.	97	26	44	148	6	39	360
	%	26.9	7.2	12.2	41.1	1.7	10.8	100
Mandel'shtam	no.	89	60	81	241	20	73	564
	%	15.8	10.6	14.4	42.7	3.5	12.9	100

Table 1: Rhythmic profile of iambic tetrameter in AbAb stanzas by Axmatova and Mandel'shtam

3 Results and Discussion

In Russian iambic tetrameter, rhythm depends primarily on the distribution of actual word stresses along ictuses in an iambic line. Based on the location of ictuses without any stress (i.e. pyrrhic feet), Kirill Taranovskij (1953) offered the following classical typology of rhythmical forms in iambic tetrameter:

- form 1: ∪– ∪– ∪– ∪– (∪);
- form 2: ∪∪ ∪– ∪– ∪– (∪);
- form 3: ∪– ∪∪ ∪– ∪– (∪);
- form 4: ∪– ∪– ∪∪ ∪– (∪);
- form 5: ∪– ∪∪ ∪∪ ∪– (∪);
- form 6: ∪∪ ∪– ∪∪ ∪– (∪).[2]

Table 1 highlights the frequency of these forms in the material considered in this paper.

The current study explores the distribution of verbs and adjectives in each of these forms with one pyrrhic foot, i.e. forms 2, 3, and 4. In the following sections, I therefore compare my new data about Axmatova and Mandel'shtam with Gasparov and Skulachyova's (2004, pp. 62–90) data about "Evgenij Onegin" by Aleksandr Pushkin. I, thus, highlight where these scholars' conclusions based on Pushkin's material apply to the 20th-century poets and where differences arise.[3]

3.1 Form 2: ∪∪ ∪– ∪– ∪– (∪)

Gasparov and Skulachyova argue that in form 2, it is very likely that adjectives will avoid foot 2 because there is no space for their stressed "head" before the pyrrhic foot. This is also due to the very strong syntactic link that adjectives

[2] ∪ represents an unstressed syllable and – represents a stressed ictus.

[3] It must be noted that the strophic structure of Pushkin's poem is very peculiar (14 lines of iambic tetrameter in a scheme AbAbCCddEffEgg). As such, it clearly differs from the structure of the poems by Axmatova and Mandel'shtam considered in this paper.

		Foot			
Author	PoS	2	3	4	Total
Axmatova	ADJ	2	3	3	8
	VERB	13	4	7	24
Mandel'shtam	ADJ	12	8	6	26
	VERB	24	7	11	42
Pushkin	ADJ	32	41	39	112
	VERB	162	41	52	255

Table 2: Form 2—No metrical stress in foot 1

create with the nouns to which they are attributed. Since such strong links tend to close the line, foot 4 is often occupied by nouns while foot 3 provides a natural home for adjectives. On the other hand, the rhythmical environment of foot 2 is perfectly suited for verbs with their long unstressed head and stressed ending. In form 2, foot 2 is therefore occupied by verbs about twice as often as any other part of speech. Moreover, verbs appear in this position three to four times as often as they do in any other foot (Gasparov and Skulachyova 2004, p. 76). The pattern is well illustrated by the data for Pushkin's "Yevgenij Onegin" (Gasparov and Skulachyova 2004, p. 85), as cited in Table 2.

Some of the differences among the three poets noted here and further on in this paper may be due to random distributions. Furthermore, since the selections from Axmatova and Mandel'shtam are relatively small, the conclusions reached about them may not be as indisputable as those about Pushkin. Axmatova's verse in form 2 contains very few adjectives and the data are therefore insufficient to trace any trends. In Mandel'shtam'scase, however, not only do adjectives prefer foot 2 over any other location in form 2 but they are only outnumbered twofold in that slot by verbs. Moreover, around half of Mandel'shtam's adjectives prefer foot 2, roughly the same frequency we see in his verbs. Such a result, however, contradicts the rhythmical and grammatical nature of adjectives proposed by Gasparov and Skulachyova.

To occupy foot 2 in this form, an adjective must be four or five syllables long with the stress on the fourth syllable (∪∪∪– or ∪∪∪–∪). This is a structure more typical of verbs, the so-called rhythmical opposites of adjectives. The question, thus, arises how this situation is possible when the average Russian adjective has a short stressed head and a long unstressed ending. To find an answer, we need to consider Mandel'shtam 's samples more closely.

To begin with, we find that 10 out of the 12 adjectives located at the start of form 2 extend their rhythmical heads by one additional syllable. This occurs through the incorporation of enclitics (Table 3).

In contrast, the ending is shortened because of the use of the short forms of the adjectives, as can be seen in five out of the 12 lines. This also explains why the above samples are hardly affected by the syntactic attraction of adjectives to the end of the line, i.e. the second major explanation for Russian adjectives' tendency to avoid feet 1 and 2. Specifically, short adjectives play a different syntactic role

	Ictus			
Line	1	2	3	4
И голубая нитка славы	0	ADJ	NOUN	NOUN
И широка моя стезя –	0	ADJ	DET	NOUN
Средь голубых шумят стихий.	0	ADJ	VERB	NOUN
На громовой призыв скрепясь:	0	ADJ	NOUN	VERB
Изображен грядущий день.	0	ADJ	ADJ	NOUN
Свой изначальный сон дробя.	0	ADJ	NOUN	VERB
Хоть говоривший мне о Риме	0	ADJ	PRON	NOUN
Кто незнаком с буфетным знаком	0	ADJ	ADJ	NOUN
Мы недовольны светом солнца,	0	ADJ	NOUN	NOUN
На деревянных лавках спят.	0	ADJ	NOUN	VERB
Как нежилого сердца дом, —	0	ADJ	NOUN	NOUN
Озарены луной ночевья	0	ADJ	NOUN	NOUN

Table 3: Form 2—Adjectives in foot 2 in Mandel'shtam's work

		Foot			
Author	PoS	1	3	4	Total
Axmatova	ADJ	9	12	5	26
	VERB	10	11	6	27
Mandel'shtam	ADJ	17	20	6	43
	VERB	19	17	5	41
Pushkin	ADJ	68	82	42	192
	VERB	118	154	56	328

Table 4: Form 3—No metrical stress in foot 2

in Russian: unlike longer adjectives, which are generally attributes, the shorter forms serve as predicates. They therefore tend to behave more like verbs and skew closer to the beginning of the line.

3.2 Form 3: ∪– ∪∪ ∪– ∪– (∪)

According to Gasparov and Skulachyova, in form 3, verbs are very common in the pyrrhic foot and their stressed syllable tends to fall in foot 3. Adjectives occupy the pyrrhic foot less often with their unstressed ending—they occur in foot 1 about half as often as verbs occur in foot 3—and are more attracted to foot 3 for syntactic reasons (Gasparov and Skulachyova 2004, p. 77); cf. Table 4.

In our samples from the two 20th-century poets, there is also clear evidence of the adjectival attraction to foot 3; this is not the case for verbs. In fact, in Mandel'shtam's case, there are fewer verbs in foot 3 than in foot 1, a pattern contrary to the average rhythmical structure. In Axmatova's case, verbs occur

in feet 1 and 3 at around the same frequency. In Table 5 we find examples from both poets.

Interestingly, for Mandel'shtam, about every third line in the list above (seven lines out of 19) starts with a verb form that has an atypical rhythmic structure where the stress occurs closer to the beginning (∪–∪∪ or ∪–∪∪∪). Of the remaining 12 verb forms that start lines, 11 have a symmetrical structure (∪–∪) and one includes only two syllables with the stress on the second one (∪–). For Axmatova, on the other hand, the picture is quite different: of the lines listed in Table 5, only one starts with a verb form that has an extended unstressed ending (I slushala yazy'k rodnoj ∪–∪∪) while eight initial verbs are symmetrical (∪–∪). In other words, for Axmatova, unlike for Mandel'shtam, it is not the ending of the verb form in foot 1 that occupies the pyrrhic foot 2, but rather the beginning of whichever part of speech has its stressed syllables in foot 3 (in most cases, a noun).

Another difference between Mandel'shtam's verbs and Axmatova's respective verb choices in foot 1 of form 3 is that whereas Mandel'shtam clearly prefers the present tense (seven lines out of 19), Axmatova is more inclined to use the past tense (four lines out of 10). Not surprisingly, the same is true of the verbs in foot 3, as can be seen from Table 6 where Axmatova's verbs are distributed evenly across the past, present, and all other verb forms. Meanwhile for Mandel'shtam, eight verbs are in the present tense and the remaining nine take either past, future, infinitive, imperative, or adverbial forms.

In another intriguing difference, we find that for Axmatova, 10 of the 11 verb forms that occupy foot 3 have a typical rhythmic structure, i.e. a longer unstressed head and shorter stressed ending; the most frequently encountered version of this is ∪∪–(∪) (9 verb forms of 11). For Mandel'shtam, however, only seven of the 17 verb forms have a typical rhythmic structure. In contrast, 10 of the verbs are either symmetrical ∪–∪ (four instances) or shorter forms with only two syllables (∪– or –∪; six instances in total). The seven typical ones have the structure ∪∪–(∪) and ∪∪∪–∪ of which there are five and two cases, respectively.

3.3 Form 4: ∪– ∪– ∪∪ ∪– (∪)

In form 4, adjectives generally predominate in the pyrrhic foot 3 and locate their stressed syllable in foot 2 for both rhythmical and syntactical reasons (Gasparov and Skulachyova 2004, pp. 67, 78, 79). This is the case for all three of the poets considered (see Table 7).

As for verbs, for rhythmical reasons, they are less easily accommodated in foot 2 than foot 4. The latter is, in contrast, well suited to their stressed ending while the preceding pyrrhic easily contains their long unstressed head. Foot 1 is also quite attractive for verbs although this is largely for syntactic reasons (Gasparov and Skulachyova 2004, p. 66). These trends are well illustrated by the data for Pushkin and Axmatova in Table 7 and. In Mandel'shtam's case, however, syntactic considerations seem somehow to override rhythmical ones. As such, although foot 4 is especially suited to the typical verb structure, i.e. a

(a) Mandel'shtam

	Ictus			
Line	1	2	3	4
Мы ринулись в зеленый омут.	VERB	0	ADJ	NOUN
И вскакивать на жесткой койке,	VERB	0	ADJ	NOUN
Вернуться на родной фрегат!	VERB	0	ADJ	NOUN
Работает в табачной мгле –	VERB	0	ADJ	NOUN
И дышит в роковых страстях.	VERB	0	ADJ	NOUN
Пустеет понемногу сад.	VERB	0	ADV	NOUN
Чтоб ладилась моя работа	VERB	0	DET	NOUN
И крепла – на борьбу с врагом.	VERB	0	NOUN	NOUN
Я слышу отреченья скрежет:	VERB	0	NOUN	NOUN
Кто выменял коня – событий	VERB	0	NOUN	NOUN
Косится на бочонок вождь.	VERB	0	NOUN	NOUN
Не хватит на мешки рогож, –	VERB	0	NOUN	NOUN
Плетется на асфальте воз.	VERB	0	NOUN	NOUN
Чернеет на скале гранитной	VERB	0	NOUN	ADJ
Напомнила твой образ, скиф!	VERB	0	NOUN	NOUN
Обманывает нас в мечтах,	VERB	0	PRON	NOUN
Я шел, не опуская глаз.	VERB	0	VERB	NOUN
И, мнится, заворкует вдруг.	VERB	0	VERB	ADV
Рыдая, обнимает дочь	VERB	0	VERB	NOUN

(b) Axmatova

	Ictus			
Line	1	2	3	4
Но знайте: не пройдет вам даром	VERB	0	VERB	ADV
Приходит долгожданный час.	VERB	0	ADJ	NOUN
Прощаясь, помахал рукой	VERB	0	VERB	NOUN
Звенели голоса детей,	VERB	0	NOUN	NOUN
Качаясь на волнах эфира,	VERB	0	NOUN	NOUN
И знаем, что в оценке поздней	VERB	0	NOUN	ADJ
Смотри, как глубоко ныряю,	VERB	0	ADV	VERB
Истлело в глубине зеркал…	VERB	0	NOUN	NOUN
И слушала язык родной.	VERB	0	NOUN	ADJ
Увидел с высоты Кремля,	VERB	0	NOUN	NOUN

Table 5: Form 3—Verbs in foot 1

(a) Mandel'shtam

	Ictus			
Line	1	2	3	4
Цветочную ли холить грядку,	ADJ	0	VERB	NOUN
Кандальную дробите цепь!	ADJ	0	VERB	NOUN
Расширенный пустеет взор,	ADJ	0	VERB	NOUN
Клубящаяся стынет пена,	ADJ	0	VERB	NOUN
Прозрачными стоят деревья,	ADJ	0	VERB	NOUN
Такую причинить обиду	DET	0	VERB	NOUN
Товарищи лежат в бреду.	NOUN	0	VERB	NOUN
Был деятель. Глядясь в себя,	NOUN	0	VERB	PRON
И голубь не боится грома,	NOUN	0	VERB	NOUN
В палатки призывал народ.	NOUN	0	VERB	NOUN
У вечности ворует всякий,	NOUN	0	VERB	PRON
Нас пеною воздвигнул случай	NOUN	0	VERB	NOUN
И мне повиновалось пламя –	PRON	0	VERB	NOUN
Я шел, не опуская глаз.	VERB	0	VERB	NOUN
И, мнится, заворкует вдруг.	VERB	0	VERB	ADV
Рыдая, обнимает дочь…	VERB	0	VERB	NOUN
Мне холодно, я спать хочу;	X	0	VERB	VERB

(b) Axmatova

	Ictus			
Line	1	2	3	4
Родимый охраняет край,	ADJ	0	VERB	NOUN
Ты призрачным сияла светом,	ADJ	0	VERB	NOUN
Пусть горько улыбнутся губы,	ADV	0	VERB	NOUN
Зачем вы отравили воду	ADV	0	VERB	NOUN
Я горько вспоминаю вас.	ADV	0	VERB	PRON
Чем нынче и живет и дышит	ADV	0	VERB	VERB
Сегодня показался мне.	ADV	0	VERB	PRON
И дом припоминая темный	NOUN	0	VERB	ADJ
Как все здесь говорит о мире,	PRON	0	VERB	NOUN
Но знайте: не пройдет вам даром	VERB	0	VERB	ADV
Прощаясь, помахал рукой	VERB	0	VERB	NOUN

Table 6: Form 3—Verbs in foot 3

		Foot			
Author	PoS	1	2	4	Total
Axmatova	ADJ	18	68	18	104
	VERB	25	13	23	61
Mandel‘shtam	ADJ	39	83	24	146
	VERB	50	38	35	123
Pushkin	ADJ	254	719	330	1303
	VERB	464	344	547	1355

Table 7: Form 4—No metrical stress in foot 3

long unstressed head and stressed ending, it appears to accommodate verbs even less often than foot 2 (Table 8).

Foot 4 is the only location considered in form 4 where Mandel'shtam's verbs occur more often in the past tense than the present (11 vs. nine instances). This is also very close to the picture for Axmatova (Table 9, seven vs. five instances). For both poets, the verb forms in foot 4 have a typical rhythmical structure in about one half of all of their lines in form 4. In Mandel'shtam's case, this amounts to slightly less than one half of the lines (17 cases out of 35) while for Axmatova, it is slightly more than one half (13 cases out of 23). Among other factors, this is because the rhythmic head of Axmatova's verbs is more often extended owing to the use of clitics (this occurs in seven out of 23 lines for Axmatova compared to four out of 35 lines for Mandel‘shtam).

3.4 Conclusion

Although Gasparov and Skulachyova's conclusions were based on just one author's use of iambic tetrameter, their explanations of the trends they observed referred to the general properties of Russian parts of speech, primarily verbs and adjectives. Moreover, they explicitly argued that the effects of grammar on rhythm and vice versa were determined by language. Their arguments, thus, implied that the trends seen in Pushkin's work could be expected in any other Russian poet's use of iambic tetrameter. This assumption persisted despite their acknowledgment that a complete understanding of interactions between rhythm and grammar in verse would require more extensive samples of poetry in Russian and other languages (Gasparov and Skulachyova 2004, p. 80).

Against this, my comparison of the grammar and rhythm of the iambic tetrameter samples in forms 2, 3, and 4 for Axmatova and Mandel'shtam shows that although the patterns discovered in Pushkin's work are well supported by Gasparov and Skulachyova's "grammatical-rhythmical dictionary" (2004, pp. 51–61), they only apply with significant limitations. In particular, they hold when dominant verbs and adjectives have a standard rhythmical structure, direct word order is followed, and syntactic inversions are avoided. These conditions are not always met, however, and they tend to be missing from the

	Ictus			
Line	1	2	3	4
И светом ласковым сиял.	NOUN	ADJ	0	VERB
Священник римский уцелел.	NOUN	ADJ	0	VERB
В стропилах каменных исчез.	NOUN	ADJ	0	VERB
Копыта конские твердят.	NOUN	ADJ	0	VERB
Оркестр торжественный настройте,	NOUN	ADJ	0	VERB
Того, кто вовремя застыл.	DET	ADV	0	VERB
Как кони медленно ступают,	NOUN	ADV	0	VERB
Душа томительно живет.	NOUN	ADV	0	VERB
И, если подлинно поется	SCONJ	ADV	0	VERB
Не мог сильнее тосковать!	VERB	ADV	0	VERB
Листы, которые умрут,	NOUN	DET	0	VERB
Лесной вершине передать.	ADJ	NOUN	0	VERB
Ты желтый сумрак предпочла.	ADJ	NOUN	0	VERB
Слух чуткий парус напрягает,	ADJ	NOUN	0	VERB
В священном сумраке исчез!	ADJ	NOUN	0	VERB
Одна пустыня пролегла.	ADJ	NOUN	0	VERB
Впервые силой изошла.	ADV	NOUN	0	VERB
Обратно в степи привела…	ADV	NOUN	0	VERB
Когда рябина, развивая	ADV	NOUN	0	VERB
И будешь сталинкою зваться	AUX	NOUN	0	VERB
Которым церковь говорит;	DET	NOUN	0	VERB
Ее лица ни покрывайте —	DET	NOUN	0	VERB
В театре публики лежало	NOUN	NOUN	0	VERB
Сквозь рощу портиков идешь.	NOUN	NOUN	0	VERB
У Чарльза Диккенса спросите,	NOUN	NOUN	0	VERB
О, время, завистью не мучай	NOUN	NOUN	0	VERB
Он только сердце веселит.	PART	NOUN	0	VERB
Как трудно раны врачевать!	X	NOUN	0	VERB
И, как ее ни называйте	ADV	PRON	0	VERB
И мы его обороним:	PRON	PRON	0	VERB
А мне уж не на кого дуться	PRON	PRON	0	VERB
О, как мы любим лицемерить	ADV	VERB	0	VERB
И небо падает, не рушась,	NOUN	VERB	0	VERB
И море плещет, не пенясь.	NOUN	VERB	0	VERB
Зарделся, вспыхнул и погас.	VERB	VERB	0	VERB

Table 8: Form 4—Verbs in foot 4 (Mandel'shtam)

poetry of historical periods or literary movements preoccupied with formal experimentation.

Although the average Russian adjective, for example, is more than three syllables long and its main stress tends to fall close to the beginning of the word, this does not necessarily mean that these adjectives are always the most

	Ictus			
Line	1	2	3	4
За то, что я не говорила	DET	PRON	0	VERB
И дом, в котором не живем,	NOUN	DET	0	VERB
В ворота черные стучит.	NOUN	ADJ	0	VERB
И мнится мне, что уцелела	VERB	PRON	0	VERB
О смерти господа моля.	NOUN	NOUN	0	VERB
Страна великая живет,	NOUN	ADJ	0	VERB
Такой, что мне не разобрать,	DET	PRON	0	VERB
За то, что я не издевалась	DET	PRON	0	VERB
И в косах спутанных таится	NOUN	ADJ	0	VERB
Кто стать звенящими поможет	AUX	ADJ	0	VERB
Остаток юности губя,	NOUN	NOUN	0	VERB
Тростник оживший зазвучал.	NOUN	ADJ	0	VERB
Ничьих я слов не повторяю	DET	NOUN	0	VERB
Как ты до мая доживешь?”	PRON	NOUN	0	VERB
Оркестр веселое играет,	NOUN	ADJ	0	VERB
Где скромно ночи провожу,	ADV	NOUN	0	VERB
Мы что-то мудрое решали,	PRON	ADJ	0	VERB
Скрипач безносый заиграл.	NOUN	ADJ	0	VERB
Чтоб мне таинственно помочь.	PRON	ADV	0	VERB
Но Лишней я не назову.	ADJ	PRON	0	VERB
А ты мой дом благослови,	PRON	NOUN	0	VERB
От русской Церкви отлетал,	ADJ	NOUN	0	VERB
Я новым именем покрою	ADJ	NOUN	0	VERB

Table 9: Form 4—Verbs in foot 4 (Axmatova)

prevalent. The actual length of a Russian adjective can vary from one syllable to a dozen, and, although such extremes are rare, there is plenty of space for variation in between. The same holds true for other part of speech as well—the Russian vocabulary includes plenty of words that do not conform with the standard rhythmical scheme for their respective part of speech. Moreover, as can be seen from many of the examples in this paper, there are easy and logical ways to extend the rhythm structure of many parts of speech with clitics; these make the actual rhythmical structure of a word even less predictable since they add more variables. Due to its generally free word order, particularly in the case of poetry, Russian syntax also imposes no severe restrictions on rhythmical grammar.

Clearly the rhythmical-grammatical trends that Gasparov and Skulachyova discovered were apparent in their data just as their explanations were based in their rhythmical dictionary of parts of speech. Nevertheless my comparison reveals a more complex picture in which rhythmical grammar appears to belong to a more intricate system. The typical structures and syntactic roles of parts of speech are undoubtedly important but they are only some of a number

of factors whose relative weights still need to be determined from more extensive and diverse material. These other significant factors may, for example, include general historical changes in language or style, an author's individual preferences, and strophic structure.

Finally, it must be noted that this paper's observations and conclusions are based on relatively limited selections from just two 20th-century poets. Further extension of the data and inclusion of other authors may eliminate the risks related to random distributions. This may also help us to better understand the nature and causes of the reported differences between the expected and real distributions of parts of speech in Russian iambic tetrameter.

Acknowledgments

The author is deeply indebted to Kirill Maslinskij and Kirill Muxin for their assistance with the automatic calculation of data for this paper and to the anonymous reviewers for their helpful comments.

References

Bobrov, Sergej (1922). "Zaimstvovaniya i vliyaniya: popy'tka metodologizacii voprosa". In: *Pechat' i revolyuciya* 8, pp. 72–92.

Bobrov, Sergej (1925). "Zaimstvovaniya stixotvorny'ye". In: *Literatutnaya e'nciklopediya: Slovar' literaturny'x terminov 1*. Moskva, Leningrad: Izdatel'stvo L. D. Frenkel', pp. 255–258.

Brik, Osip (1927). "Ritm i sintaksis: (Materialy' k izucheniyu stixotvornoj rechi)". In: *Novy'j LEF*, pp. 3–6.

Gasparov, Mixail Leonovich (1984). "Ritmicheskij slovar' i ritmiko-sintaksicheskiye klishe". In: *Problemy' strukturnoj lingvistiki*. Moskva, pp. 169–185.

Gasparov, Mixail Leonovich (1996). "Lingvistika stixa". In: *Slavyanskij stix: stixovedeniye, lingvistika I poe'tika*. Moskva, pp. 5–17.

Gasparov, Mixail Leonovich (2000). ""Ty'-ty' rifmy'": ritmiko-sintaksicheskiye klishe u Pushkina". In: *Posle yubileya*. Ed. by Samuel Schwarzband and Roman Timenchik. Jerusalem: The Center for the Study of Slavic Languages and Literatures at the Hebrew University of Jerusalem, pp. 71–84.

Gasparov, Mixail Leonovich (2004). "Ritmiko-sintaksicheskiye klishe i formuly' v e'piloge "Ruslana i Lyudmily"". In: *Slavyanskij stix VII: Lingvistika i struktura stixa*. Ed. by Mixail Leonovich Gasparov and Tatyana Vladimirovna Skulachyova. Moskva: Yazy'ki slavyanskoj kul'tury', pp. 149–166.

Gasparov, Mixail Leonovich and Tatjana Vladimirovna Skulachyova (2004). *Stat'i o lingvistike stixa*. Moskva: Yazy'ki slavyanskoj kul'tury.

Marneffe, Marie-Catherine de, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford dependencies: A cross-linguistic typology". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA),

pp. 4585–4592. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). “A Universal Part-of-Speech Tagset”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul: European Language Resources Association (ELRA), pp. 2089–2096. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

RNC (2020). *Russian National Corpus*. URL: https://ruscorpora.ru/new/search-poetic.html (visited on 11/15/2020).

Taranovskij, Kiril (1953). *Ruski dvodelni ritmovi*. Beograd: Srpska akademija nauka.

Tomashevskij, Boris Viktorovich (2007a). “Pyatistopny'j yamb Pushkina”. In: *Izbranny'ye raboty' o stixe*. Sankt-Peterburg, Moskva: Filologicheskij fakul'tet SPbGU, Izdatel'skij centr “Akademiya”, pp. 140–242.

Tomashevskij, Boris Viktorovich (2007b). “Ritmika chety'ryoxstopnogo yamba po nablyudeniyam nad stixom “Yevgeniya Onegina””. In: *Izbranny'ye raboty' o stixe*. Sankt-Peterburg, Moskva: Filologicheskij fakul'tet SPbGU, Izdatel'skij centr “Akademiya”, pp. 101–139.

UD (2020). *Universal Dependencies*. URL: https://universaldependencies.org/ (visited on 11/15/2020).

Yarxo, Boris (1927). “Prostejshiye osnovaniya formal'nogo analiza”. In: *Ars Poetica I*. Moskva: Gosudarstvennaya akademiya xudozhestvenny'x nauk, pp. 7–29.

Zeman, Daniel (2008). “Reusable Tagset Conversion Using Tagset Drivers”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf.