Grammatical Rhymes in Polish Poetry: a Quantitative Analysis[1]

Karol Opara
Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warszawa, Poland
karol.opara@ibspan.waw.pl
tel. +48 223810393

Abstract

Analysis and interpretation of poetry is based on qualitative features of its text such as its semantics or means of expression as well as on general knowledge about the author and artistic period. Recent advances in automatic text processing allow for performing quantitative analysis of large sets of poetry. Their results may facilitate assessment of linguistic capabilities of its author or in other words his poetic mastery. This contribution presents a method of calculating the share of grammatical rhymes in Polish poetry. It is used to create a ranking of both historic and contemporary Polish poets. Comparative study and statistical analysis is developed using *Pan Tadeusz* by Adam Mickiewicz as a reference poem for Polish poetry. Assessment of technical mastery is one step towards the introduction of objective measures of poetic quality.

Keywords

rhyme detection, Computer-Aided Poetry, *Pan Tadeusz*

1 Introduction

Perception of poetry largely depends on prosodic features of its language such as intonation, meter or rhyme. Linguistic, qualitative analyses of poetry have been conducted for thousands of years now. However, quantitative evaluation of verse structure became much more effective after the introduction of automatic text processing tools.

Prosodic features of poetry are easier to analyse automatically than its semantics, whose subtlety and ambiguity sometimes make its interpretation challenging even for specialists. Their statistical study coupled with analysis of the verse syntax constitutes an important part of an interdisciplinary trend of applying modern Computer Science techniques in poetry generation, analysis, evaluation, translation and paraphrasing. This emerging field of study could be called Computer-Aided Poetry (CAP).

The interest of the scientific community for CAP is growing, which is best illustrated with anonymous reviewers giving their remarks in verse. For instance, comments on a paper by Genzel et

[1] *This is a pre-copyedited, author-produced version of an article accepted for publication in Digital Scholarship in the Humanities following peer review. The version of record "Opara, K. R. (2014). Grammatical rhymes in Polish poetry: A quantitative analysis. Digital Scholarship in the Humanities, 30(4), 589-598" is available online at:* https://dx.doi.org/10.1093/llc/fqu029*.*

al. (2010) on Statistical Machine Translation (SMT) begin in the following way (Anonymous Reviewer, 2010):

> *This paper has an admirable objective;*
> *one that would have most of us spewing invective.*
> *Can current SMT systems be hacked*
> *to translate verse with its form intact? (…)*

One of the main challenges of CAP is the introduction of objective quality measures for poetry. Although the first steps towards reaching this goal have been taken (Dalvean, 2013), attempts to evaluate poetry and its authors' mastery usually remain quite subjective.

Little research was done on the automated analysis of poetry in inflected languages, where rhymes depend on the morphology of words. In this contribution, I try to fill this gap for the case of Polish by investigating the frequency of grammatical rhymes in a verse.

The reception of grammatical rhymes in different poetic schools is discussed in section 2. It is followed by description of rhyme detection procedure. Next a notion of *Cz*-score is introduced to provide a quantitative measure of grammatical rhymes in (a set of) poems. In sections 5 and 6 a reference poem is introduced to facilitate comparative, statistical analysis of grammatical rhymes in both historical and contemporary Polish poetry. A discussion of promising areas of further study is concluded by a brief summary of the paper.

An earlier Polish version of this paper is currently in press in *Polonica* (Opara, 2013). The aim of this extended contribution, is to make the findings also accessible to international readers.

2 Grammatical rhymes

In inflected languages parts of speech in the same morphological form have common endings. This introduces grammatical rhymes, which do not necessary reflect the creativity of an author but are a consequence of the language structure. In English, this phenomenon occurs only to a very limited extend. For example, words 'greatest', 'slowest' and 'tallest', 'strongest', etc. rhyme only because all of them are superlative adjectives and therefore have the common ending 'est'. In highly inflected languages "simply matching morphological endings is a task which – truth to be told – does not require great ingenuity" (Wachtel, 2004). For instance, in second person singular most Russian verbs end with *ешь* (transliterated as 'yesh'') leading to multitude of rhyming words (Wachtel, 2004): *читаешь* 'chitayesh'' (read), *понимаешь* 'ponimayesh'' (rememeber), *знаешь* 'znayesh'' (know), *извиняешь* 'izvinyayesh'' (apologize), etc. The same applies to Polish (Pszczołowska, 1972). For instance, masculine adjectives typically end with 'ego', e.g. 'ładnego' (nice), 'dobrego' (good), 'zielonego' (green), 'obcego' (foreign).

What is the appropriate share of grammatical rhymes in a verse? The literary cannon varies between languages. In classical Latin all kinds of rhymes were avoided, with a strange exception in famous poem III, XXX (*Exegi monumentum*) by Horace, where a purely grammatical rhyme 'perennius' : 'altius' was used. To a larger extend, rhymes appeared in Latin only in the early middle ages (Podhorki-Okołów, 1925). On the other hand, in French medieval poetry grammatical rhymes were rather welcome (Shaw, 2003).

Jakobson (1960) notes that poets or poetic schools tend to be oriented towards or against grammatical rhyme – indifference to this aspect of verse is rather excluded, especially in Slavic

languages. Moreover, such preferences may change over time. For instance, in 18th century grammatical, perfect rhymes were demanded in Russian poetry, while in the 20th century they become much rarer (Wachtel, 2004). Properties of a language also play a role in constituting literary canons. Kiparsky (1973) argues that "a good number of what we think of as traditional and arbitrary conventions are anchored in grammatical form, and seem to be, at bottom, a consequence of how language itself is structured". Following this way of thinking, Wagner and McCurdy (2010) suggest that users of highly inflected languages in which word stress tend to fall on one of the last syllables are more tolerant towards grammatical rhymes.

In early Polish poetry grammatical rhymes were very popular (Głowiński, 1991) but their prevalence decreased significantly in the 19th century (Pszczołowska, 1970). Poets of the interbellum period inclined towards non-grammatical rhymes, which is clearly expressed in Podhorki-Okołów (1925) essay:

> *(…) im lepszy (rzadszy, trudniejszy) rym, tym lepsza, mocniejsza treść; im bliższe sobie a więc łatwiejsze do połączenia są słowa stanowiące rym, tym mniej skomplikowana i esencjonalna, a bardziej łatwa, płytka i monotonna jest treść.*
>
> *(…) the better (rarer, harder) a rhyme the better, stronger its content; the closer and hence easier to connect are the words in a rhyming pair the less elaborated and essential but easier, shallower and more monotonous the content is.*

This alleged banality of grammatical rhymes has led to their disregard in contemporary Polish poetry (Pszczołowska, 1970). It manifests in a pejoratively referring to trivial and grammatical rhymes as 'Częstochowa' rhymes. This notion origins from pious books with verses of questionable quality sold to pilgrims arriving in monastery in Częstochowa to pray to the Virgin Mary (Piersiak, 2008).

There is only a limited dictionary of non-grammatical rhymes. Most of them were exploited by modernist and interbellum poets making it difficult to find genuinely original rhyming pairs. This has lead to wider spread of masculine and inexact rhymes (Pszczołowska, 1970), while contemporary Polish poetry inclines towards blank verse.

Pszczołowska (1972) counted that in a set of 100 multi-syllable words from *Konrad Wallenrod* by Mickiewicz there were 12 exact rhymes, out of which only 2 were clearly non-grammatical. Complete abandonment of grammatical rhymes would therefore lead to strange, unnatural constructions. It would also mean the resignation of repetition as a means of expression. Hence, one need not completely resign from grammatical rhymes but rather avoid overusing them to the point of dominating a verse. Noteworthy, some contemporary authors take advantage of morphological similarities to express intended naivety, irony or enhance the sonic aspects of a verse. Consonance of form can be also contrasted with dissonance of meaning (Pszczołowska, 1972).

3 Rhyme Detection

Automatic investigation of grammatical rhymes requires detecting rhyming words. There is little literature related to this topic, probably due to the fact that Computer-Aided Poetry is only an emerging field of study. Genzel et al. (2010) worked on machine translation of poetry, which keeps meter and rhyme. However, technical details of their approach remain the intellectual property of a private company. Hirjee and Brown (2010) characterized rhyming style in rap music by automatic detection of rhymes with advanced tools from bioinformatics. Their analysis is helpful in authorship identification, style-based comparison and music recommendation. Greene et al. (2010) use word-stress patterns as well as rhyme and discourse models to generate English love poetry. However, all

these papers concentrate mainly on English or a pair of languages in case of translation tasks, which means that analysis of Polish verses would require adopting some of these approaches.

Poetry is usually stored as text, whereas rhymes depend on its pronunciation. Therefore, the first step in rhyme detection consists of converting text to its phonetic representation. In this paper I based it on phonemes from a database of Polish diphones CORPORA (Grocholewski, 1997). In Polish, text quite unambiguously defines its pronunciation, which enables conversion based on a simple rule-based system. In this study, only the line-final rhymes were detected, which are the most canonical and common rhymes in Polish (Głowiński et al., 1991).

The last word in each line was compared with the last words of the preceding four lines. Rhymes were detected based on the similarity in the last three phonemes of both words. This roughly corresponds to syllables, which, according to Śledziński (2008), are a good unit for speech analysis and synthesis. Similarities between phonemes from the two potentially rhyming words were calculated using the matrix obtained with UPGMA clustering method (UPGMA stands for Unweighted Pair Group Method with Arithmetic Mean). The matrix was developed by Gałka in his PhD thesis (2008) through spectral analysis using a six-level parameterization of discrete wavelet transform and 'dmey' wavelets for 5 different speakers. This similarity matrix allowed me to detect both perfect and imperfect rhymes. Further improvements could possibly be obtained by taking into account coarticulation of consecutive phonemes (Grocholewski ,1997; Śledziński, 2008).

4 *Cz*-scores

Automatically detected rhyming words underwent analysis to find their possible morphological tags. Table 1 presents an example of rhymes taken from the invocation to the Polish national poem *Pan Tadeusz* (Mickiewicz, 1834; translation by Marcel Weyland):

Panno Święta, co Jasnej bronisz Częstochowy
I w Ostrej świecisz Bramie! Ty, co gród zamkowy
Nowogródzki ochraniasz z jego wiernym ludem!
Jak mnie dziecko do zdrowia powróciłaś cudem
(Gdy od płaczącej matki pod Twoję opiekę
Ofiarowany, martwą podniosłem powiekę
I zaraz mogłem pieszo do Twych świątyń progu
Iść za wrócone życie podziękować Bogu),
Tak nas powrócisz cudem na Ojczyzny łono.

Holy Virgin who shelters our bright Częstochowa
And shines in Ostra Brama! You, who yet watch over
The castled Nowogródek's folk faithful and mild;
As You once had returned me to health, a sick child,
(When by my weeping mother into Your care given,
I by miracle opened a dead eye to heaven,
And to Your temple's threshold could straightaway falter
For a life thus returned to thank God at the altar)
Thus to motherland's breast You will bring us again.

Possible morphological forms were obtained using the *Morfeusz* analyser (Woliński, 2006; Saloni et al., 2011). It may be described as "little more than an appropriately compressed morphological dictionary" (Przepiórkowski and Murzynowski, 2009). The dictionary contains a

significant share of the 19th century vocabulary, some of which is not used in current-day Polish anymore. Surprisingly, this is advantageous for the study, as it makes analysis of historical poetry more reliable.

Tags depicted in Table 1 describe morphological forms of a word allowed by Polish grammar. For instance the word 'Częstochowy' can be a noun (subst) in singular number (sg), in genitive (gen) and of feminine gender (f), which is written in short as subst:sg:gen:f. However it can be also a form of nominative, accusative or vocative (nom.acc.voc) in plural number, which is written as subst:pl:nom.acc.voc:f. Analogously, the word 'zamkowy' is tagged as adj:sg:nom.acc.voc:m1.m2.m3:pos, which means it is an adjective (adj) in singular number (sg), in nominative, accusative or vocative (nom.acc.voc) in one of the three masculine genders (m1.m2.m3) and in positive degree (pos). A detailed explanation of the morphological tags is given in (Woliński, 2003; Przepiórkowski, 2004).

To assess the share of grammatical rhymes in a verse I introduced a measure called *Cz*-score. In cases where both rhyming words have only one common tag, it must be a grammatical rhyme and it is assigned *Cz* = 100 percent. On the other hand, if the intersection of both tag lists is empty, the rhyme gets no scores *Cz* = 0. In the remaining cases it is unclear if the rhyme is grammatical. For instance, both words 'opiekę' and 'powiekę' may be feminine nouns in accusative. On the other hand, 'opiekę' can be a future form of a verb ('piec'). Hence, determining whether such a rhyme is grammatical requires performing tag disambiguation. This in turn requires syntactic analysis of the verse, which may be difficult in the case of poetry and is beyond the scope of this work. Consequently, in cases where the lists of possible tags for a rhyme contain both common and different elements, it is assigned a compromise value of *Cz* = 50 percent. Examples of such an approach are shown in Table 1, where manually disambiguated tags are typed in bold. I also report the respective number of points and indicate if the original rhymes are indeed grammatical.

Table 1. Morphological analysis of rhymes from invocation to *Pan Tadeusz* by Adam Mickiewicz witch their scores *Cz*; manually disambiguated tags are typed in bold

Word 1	Morphological tag 1	Word 2	Morphological tag 2	*Cz*-score	Gramm. rhyme
Częstochowy	**subst:sg:gen:f** subst:pl:nom.acc.voc:f	zamkowy	adj:sg:nom.voc:m1.m2.m3:pos **adj:sg:acc:m3:pos**	0%	No
ludem	**subst:sg:inst:m3**	cudem	**subst:sg:inst:m3** subst:sg:inst:n2	50%	Yes
opiekę	**subst:sg:acc:f** fin:sg:pri:perf	powiekę	**subst:sg:acc:f**	50%	Yes
progu	**subst:sg:gen:m3** subst:sg:loc:m3 ubst:sg:voc:m3	Bogu	**subst:sg:dat:m1** subst:sg:loc:m1	0%	No
biała	subst:sg:nom.voc:f **adj:sg:nom:f:pos** adj:sg:voc:f:pos	pała	subst:sg:gen:m3 subst:sg:nom:f **fin:sg:ter:imperf**	50%	No
ugoru	**subst:sg:gen:m3**	dworu	**subst:sg:gen:m3**	100%	Yes

5 *Reference poem and Cz-test*

Comparison of the rhyming style of Polish poets may be based on a reference poem, in which the share of grammatical rhymes is on a level characteristic of high-quality poetry. For this purpose I chose *Pan Tadeusz* by Adam Mickiewicz, which is regarded as a national epic. Moreover, it is one of the longest pieces of rhymed poetry in Polish literature, which allows for reliable statistical investigations.

The reference poem contains over 4800 rhymes (9850/2 lines is 4925 pairs, but the text contains also many triple rhymes). The power of statistical tests increases with sample size, which is explained in most statistics textbooks, e.g. (Koronacki and Mieliczuk, 2009). For this reason, I normalized *Pan Tadeusz* to use a standard number of $n = 100$ rhymes. Distribution of *Cz*-scores resulting from imposing such a constraint was investigated by drawing a million (1 000 000) random subsamples.

In this way I obtained a distribution of *Cz*-scores in *n*-rhyme subsamples from *Pan Tadeusz*, which is plotted in Fig. 1. To create a non-parametric statistical test at significance level α one must cut $\alpha/2$ percent of probability from each tail of the distribution, i.e. from each side of the bell curve. If the mean share of grammatical rhymes for a new observation, e.g. a new poem, falls into one of the cut tails, its share of grammatical rhymes significantly differs from the reference one. The choice of significance level is always somewhat arbitrary, as it should take into account the sample size and data characteristics. In this paper I chose significance $\alpha = 99\%$.

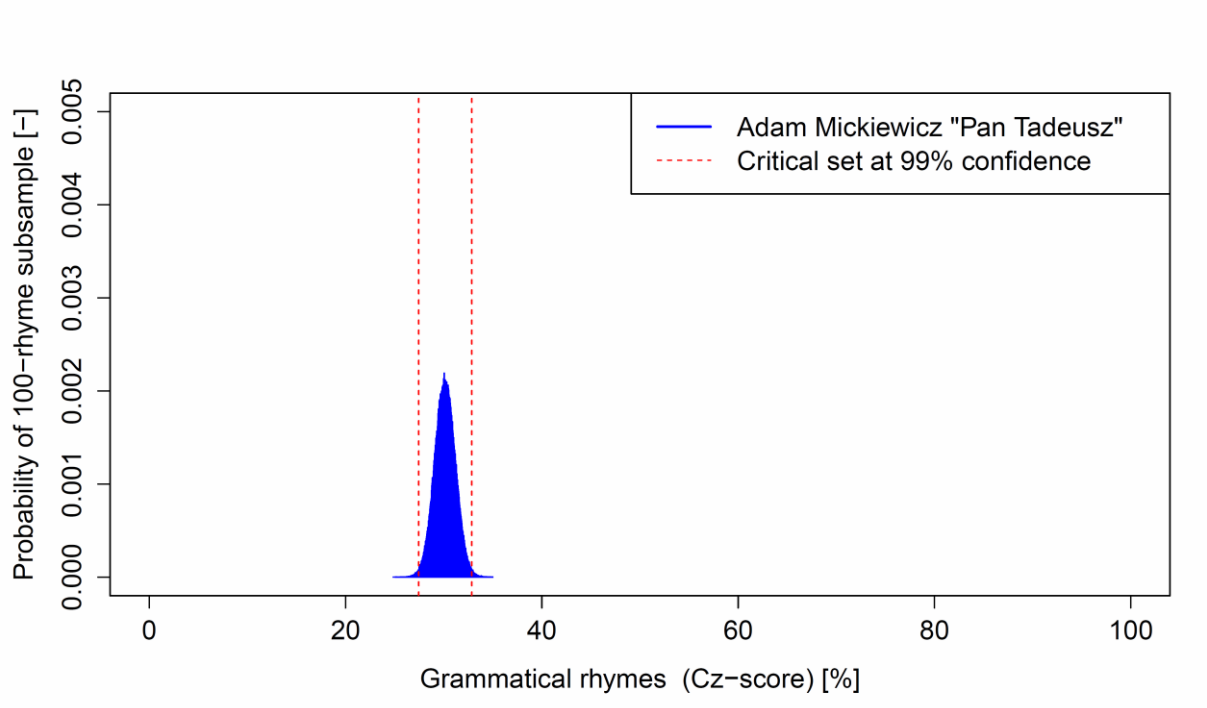

Fig. 1. Share of grammatical rhymes in 1 000 000 (a million) rhyme random samples of *Pan Tadeusz* by Adam Mickiewicz and the critical set of the *Cz*-test at significance level 99%

The *Cz*-test used in this paper is nearly equivalent to the common *t* test. According to the Central Limit Theorem, the distribution of the mean of *Cz*-scores in *n*-rhyme samples tends to normality. To see if *n* is a sufficiently large number, I checked million subsamples from *Pan Tadeusz* with seven normality tests. All of them failed to refuse the null hypothesis with confidence greater than 99.9%. Consequently, one can model the distribution of means of *n*-rhyme subsamples of a

poem with a normal distribution with expectation μ equal to the mean *Cz*-score for the whole poem and standard deviation σ given by the following formula

$$\sigma = \sqrt{\frac{p_0 \cdot 0^2 + p_{50} \cdot 50^2 + p_{100} \cdot 100^2 - \mu^2}{n}} = \sqrt{\frac{2500 p_{50} + 10000 p_{100} - \mu^2}{n}} . \quad (1)$$

Numerator in equation (1) describes standard deviation of *Cz*-scores in the whole poem. This is a discrete distribution, so its standard deviation can be computed from the definition by counting ratios p_0, p_{50} and p_{100} of rhymes for each *Cz*-score

$$p_i = \frac{\text{number of rhymes, for which } Cz = i}{\text{total number of rhymes}}, \quad \text{for } i = 0, 50, 100. \quad (2)$$

The square root of *n* in the denominator of equation (1) reflects the calculation of a mean from a sample of *n* independently chosen rhymes.

Analysis of the reference poem shows that the share of grammatical rhymes is approximately 30%. Therefore, I adopt value Cz_{ref} = 30% as the reference amount of grammatical rhymes in Polish poetry. Comparing the amount of grammatical rhymes in a verse with the reference value Cz_{ref} is facilitated by the statistical test, which shows if observed differences are significant. To balance for different lengths of works the analysis is based on subsamples from *Pan Tadeusz* of a fixed length of *n* = 100 rhymes.

6 Grammatical Rhymes in Polish Poetry

The introduction of *Cz*-scores and the discussion of their statistical properties allows one to compare the amount of grammatical rhymes that is typical of the styles of various poets. For this purpose, I selected a set of Polish poetry from a few artistic periods. Apart from verses of national poets (referred to in Polish as *wieszcz*) Mickiewicz and Słowacki, I included Nobel Prize winners Miłosz and Szymborska, poets Krasicki, Staff and Tuwim, songwriters Osiecka and Kaczmarski, authors of poetry for children Konopnicka and Makuszyński as well as the popular disco band 'Weekend'. I also added a collection of folk verses gathered by ethnographer Kolberg and translation of Shakespeare's *Midsummer Night's Dream* by Barańczak. A more detailed description of the corpus is given in Table 2.

Table 2. Description of the corpus analysed in this contribution

Author	Pieces	Artistic period
Józef Baka	*Uwagi śmierci niechybnej*	Baroque
Ignacy Krasicki	*Monachomachia* and *Antymonachomachia*	Enlightenment
Adam Mickiewicz	*Pan Tadeusz*	Romanticism
Juliusz Słowacki	*Beniowski*	Romanticism
Oskar Kolberg	Collection of folk poetry	18th – 19th century
Maria Konopnicka	Poetry for children	Modernism
Leopold Staff	Verses	Interbellum
Julian Tuwim	*Bal w operze*	Interbellum
Kornel Makuszyński	*Koziołek Matołek*	Interbellum

Czesław Miłosz	Verses	Contemporary
Wisława Szymborska	Verses	Contemporary
Agnieszka Osiecka	Songs	Contemporary
Jacek Kaczmarski	Sung poetry	Contemporary
Stanisław Barańczak	Translation of *A Midsummer Night's Dream*	Modern day
'Weekend'	Discography	Modern day

Figure 2 shows the mean share of grammatical rhymes, which is characteristic of the style of each poet. Accuracy of estimation of that mean grows with the number of analysed rhymes. It is represented by a horizontal line, which covers this mean with 99% probability. Confidence intervals plotted in Fig. 2 were obtained with a formula differing from relation (1) by substituting the constant n in the denominator with the number of rhymes r detected in the investigated verses. For normal distribution, the range within two standard deviations from the mean (i.e. from μ-3σ to μ+3σ) covers approximately 99.7% of observations and hence approximately correspond to the confidence intervals plotted in Fig. 2. The length of vertical lines showing accuracy of the *Cz*-score estimation is therefore inverse proportional to the root of the number of rhymes r.

For each author, I also marked the lower and upper bound corresponding to the most pro-grammatical and anti-grammatical way of disambiguating morphological tags. In other words, all rhymes that obtained 50 *Cz*-scores are counted in the former case as if they were grammatical rhymes (*Cz* = 100) while in the latter as if they were non-grammatical (*Cz* = 0).

To enable a visual representation of the *Cz*-test, its critical set at a confidence level of 99% is plotted in Fig. 2 with vertical, dashed lines. Their location is the same place as in Fig. 1. A verse is claimed to be statistically significantly different from the reference poem, if its mean share of grammatical rhymes lay outside of the critical set, i.e. in the tails of the bell-shaped distribution depicted in Fig. 1.

The highest share of grammatical rhymes in the analysed verses goes to the disco band 'Weekend', yet even in this case at least 30% of rhymes are not grammatical (Liszewski, 2002):

Bum, bum, bum,
na faceta tak jak rum,
tak bardzo działa ruch damskiego ciała.
Bum, bum, bum,
facetowi tak jak rum
chodź ci pokażę swoje tatuaże.

Bum, bum, bum,
to a guy like strong rum,
for dancer's body there falls everybody.
Bum, bum, bum,
for a guy like strong rum,
come babe woo, I'll show you my tattoo.

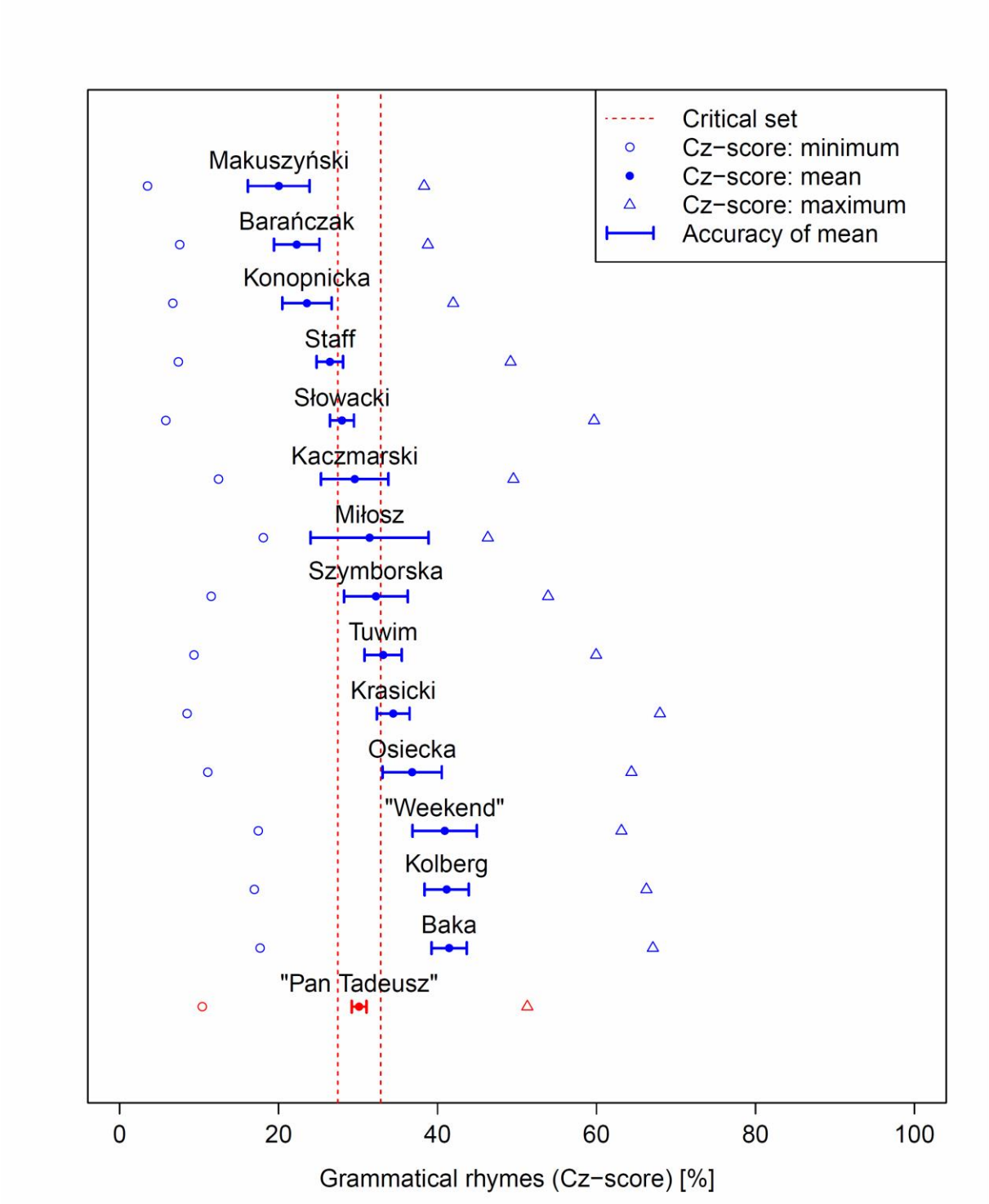

Fig. 2. Ranking of selected authors based on the amount of grammatical rhymes in their poetic works

Results of the *Cz*-test suggest that this kind of poetry may not be very refined, despite its popularity. However, the artistic capabilities of Radosław Liszewski, who is the leader and lyricist for this disco band, seem to develop over time. If one constrains analysis only to the newest album *Ona tańczy dla mnie* (*She's dancing for me*), rhymes are less grammatical and very close to the reference value Cz_{ref}.

Among the poets who are already in the Polish literary canon, priest Józef Baka in his *Comments on certain death* (Baka, 1776) leads in the use of grammatical rhymes:

Cny młodziku, migdaliku,
Czerstwy rydzu, ślepowidzu (...)
Śmierć jak kot wpadnie w lot!

Chaste youngster, dandy youngster,
rooty fungus, you blind-seer (...)
Death like cat grabs your hat!

His baroque verses were for a long time synonymous with graphomania and poor taste. Only in the 20th century was he recognized as a precursor of surrealism and linguistic poetry. This seems to explain the statistically significant excess of the reference share of *Cz*-scores.

Folk poetry is known for common use of grammatical rhymes. High position in the ranking of verses gathered by Kolberg confirms this statement.

The next place in the ranking goes to Agnieszka Osiecka, an author of popular songs and essays, which seems quite surprising taking into account generally positive reception of her poetry. This shows that the amount of grammatical rhymes is only one of many factors influencing the quality of verses.

Cz-scores for most poets are not significantly different from those of a *n*-rhyme subsamples of *Pan Tadeusz*. Is there, perhaps, a vital share of grammatical rhymes in Polish poetry, which follows from the structure of the language?

Juliusz Słowacki tends to develop his rhymes very carefully and obtains quite low *Cz*-scores. Adam Mickiewicz informally reflected upon this fact when characterizing his rival's poetry as:

Gmach piękną architekturą stawiony, jak wzniosły kościół – ale w kościele Boga nie ma.

A great piece of architecture, as a sublime church – but there is no God in the church.

Automatic analysis of the verse structure shows that architecture of Słowacki's rhymes is indeed impressive. Nevertheless, evaluating the spiritual value of his verses requires in-depth semantic analysis of the poems, which will for a long time remain beyond the reach of Computer-Aided Poetry.

In her seminal monograph about rhymes, Pszczołowska (1972) often recalls Leopold Staff as a non-grammatically rhyming poet, which fully agrees with his low position in the ranking.

Barańczak's translation of Shakespeare is characterized by significantly lower *Cz*-score than its value in the reference poem. There might be various underlying reasons, such as the influence of the English original, current trends in Polish poetry or Barańczak's individual style. Further study is required to decide which of those factors are most important.

The lowest share of grammatical rhymes is due to Maria Konopnicka and Kornel Makuszyński. Their verses for children make an example of poet's technical mastery which is of a great didactic value for the youngest.

Assessment of the Nobel Prize winners Czesław Miłosz and Wisława Szymborska is difficult because of the scarcity of rhymes in their verses, nevertheless they seem to fall into the range typical for most poets.

7 Discussion and further study

Statistical analyses of the rhyming style of different poets were already performed by Podhorski-Okołów (1925) or Pszczołowska (1972, 1973). Their credibility is based on semantic analysis of respective words and phrases as well as expert skills in determining morphological form. However, manual investigations require extensive amount of work and analysis of longer pieces of poetry becomes tedious and tiring. This leads to limiting the scope of research to a few verses as in

(Podhorski-Okołów, 1925) or a few hundred rhyming pairs as in (Pszczołowska, 1972, 1973). Quantitative information obtained in this way does not necessarily represent the patterns that are found across the whole text (Mahlberg and Smith, 2012).

Novelty of the approach proposed in this paper consists in automating the procedure of extracting statistical information from rhymed texts. This widens the scope of traditional analyses from selected pieces to whole corpora of verses. An overview of the linguistic phenomena in a large (or even complete) group of authors provides basis for stating much stronger conclusions and more accurate generalizations.

Variation of the rhyming style among Polish poets is evident. However, there is still space for improvement of the method proposed in this paper. The analysis is affected by choosing the compromise value of the *Cz*-score = 50 in case when it is not clear if the rhyming pair is grammatical. This biases the resulting share of grammatical rhymes towards 50%. Hence it is more meaningful to draw conclusions from relative position of a poet among different authors rather than the absolute amount of grammatical rhymes. This was one of the reasons to introduce the reference poem.

To reduce the bias one should perform disambiguation of morphological tags, which is a topic of further studies. Choosing appropriate tag is difficult due to complicated form of poetic texts characterized by numerous inversions and unusual grammatical constructions. Another problem is a large number of possible tags, reaching in theory 4 000. Morphological analyser *Morfeusz* used in this study was also a basis for development of the National Corpus of Polish (Przepiórkowski et al., 2012). In the manually disambiguated 1-million word subcorpus (Przepiórkowski, 2009) there are 935 different tags, out of which 184 most frequent cover 95% of words. This poses a challenge to taggers of Polish, the best of which currently have accuracy of 91% to 92%. Importantly, distribution of morphological forms in a rhymed poetry differs significantly from blank verse (Pszczołowska, 1972). Furthermore, the manually annotated 1-million word subcorpus, which is the main linguistic resource for developing and evaluating taggers of Polish, contains no poetry, therefore one should expect higher error rate.

Fine granularity of the tagset narrows the notion of a grammatical rhyme. Rhymes which consist of different parts of speech are clearly non-grammatical, but for some distinct morphological tags may be quite similar. For instance, in current Polish in most contexts vocative can be substituted with nominative (which overall occurs 65 times more often in the 1-million corpus). Even more subtle difference occurs between three variants of masculine gender, denoted m1, m2 and m3. Moreover, Polish nouns do not inflect for gender. Woliński (2006) included it in the tags because “it is an important attribute of nominal lexemes describing their syntactic features”, such as agreement in gender between adjectives and nouns in a phrase. It seems therefore, that a rhyme may be grammatical to some extent, depending on the morphological similarity within a pair of words. Quantifying this value remains an open problem.

8 Summary

In this contribution I present a method and results of an automated evaluation of the technical quality and subtlety of the rhymes of select Polish poets. I use morphological tagging to detect and extract grammatical rhymes and calculate their share in the investigated works. The Polish national poem, *Pan Tadeusz* by Adam Mickiewicz, is used as a reference text to develop a statistical test and rank poets according to the frequency of grammatical rhymes characteristic of their style. The results generally agree with the existing knowledge in this field. Nevertheless, an automated evaluation of

rhyme based on transparent criteria enables a broad-based evaluation of writing styles of Polish poets.

Computer Science enters into Linguistics in many ways, such as in the correction of spelling and grammatical errors, machine translation tools or methods for synthesis and recognition of speech. This process will continue and further develop. More and more literary works are available online, which allows researchers to accelerate various kinds of linguistic analyses, especially these which concern formal matters. This paper falls into that interdisciplinary trend, as it uses natural language processing tools to investigate selected elements of verse structure.

Funding

This work was supported by research fellowship funded by European Social Fund within the Human Capital programme project "Information technologies: research and their interdisciplinary applications" agreement number POKL.04.01.01-00-051/10-00.

References

Anonymous Reviewer (2010). Paper review in verse. http://research.google.com/archive/papers/review_in_verse.html. Accessed 17 July 2013.

Baka, J. (1766). *Uwagi o śmierci niechybnej, wszystkim pospolitej, wierszem wyrażone, a sumptem Jmci pana Ksawerego Stephaniego, obywatela miasta J.K.M. Wilna, do druku na pożytek duchowny podane w roku 1766* (*Comments on inevitable Death…*, a poem).

Dalvean, M. (2013). Ranking contemporary American poems *Literary and Linguistic Computing*, DOI:10.1093/llc/fqt03 (online advance access).

Gałka, J. (2008). *Optymalizacja parametryzacji sygnału w aspekcie rozpoznawania mowy polskiej* (*Optimization of signal parameterization for Polish speech recognition*). PhD thesis, AGH University of Science and Technology, Cracow.

Genzel, D., Uszkoreit, J. and Och, F. (2010). "Poetic" statistical machine translation: Rhyme and meter. In *proceedings of 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP'10, 158–166, Stroudsburg, PA, USA. Association for Computational Linguistics.

Głowiński, M., Okopień-Sławińska, A. and Sławiński J. (1991). *Zarys teorii literatury* (*Outline of the theory of literature*), Wydawnictwa Szkolne i Pedagogiczne.

Greene, E., Bodrumlu, T. and Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP'10*, 524–533, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Grocholewski, S. (1997). Corpora - speech database for Polish diphones. In *proceedings of EUROSPEECH*.

Hirjee, H. and Brown, D. G. (2010). Using automated rhyme detection to characterize rhyming style in rap music. *Empirical Musicology Review*, 5(4): 121–145.

Jakobson, R. (1960). Closing statement: Linguistics and poetics. In T. Sebeok (Ed.), *Style in language*, 350–377.

Kiparsky, P. (1973). The role of linguistics in a theory of poetry. *Daedalus* 102(3), 231–244.

Koronacki, J. and Mielniczuk, J. (2009). *Statystyka dla studentów kierunków technicznych i przyrodniczych* (*Statistics for students of technical and natural disciplines*). Wydawnictwo Naukowo-Techniczne.

Liszewski, R. (2002). A Ja To Lubię (I like it), song from *album "Playboy"* by group Weekend issued by Green Star.
Mahlberg, M. and Smith, C. (2012). Dickens, the suspended quotation and the corpus, *Language and Literature* 21(1): 51–65.
Mickiewicz, A. (1834). *Pan Tadeusz czyli Ostatni zajazd na Litwie: historia szlachecka z roku 1811 i 1812 we dwunastu księgach wierszem* (*Sir Thaddeus, or the Last Lithuanian Foray: A Nobleman's Tale from the Years of 1811 and 1812 in Twelve Books of Verse*, a poem).
Opara, K. (2013). Rymy częstochowskie w poezji polskiej – ujęcie ilościowe (*Częstochowa rhymes in Polish poetry – a quantitative approach*)**,** *Polonica* XXXIII, in print.
Piersiak, T. (2008). Z „pana Częstochowy" nie możemy być dumni ("Mr Częstochowa" brings us no pride, newspaper article). *Gazeta Wyborcza*, 23 October 2008.
Podhorki-Okołów, L. (1925). O rymowaniu (About rhyming), *Skamander* 37(5), 26–39.
Przepiórkowski, A. (2004). *The IPI PAN Corpus. Preliminary version.* Institute of Computer Science, Polish Academy of Sciences, Warsaw.
Przepiórkowski, A. and Murzynowski, G. (2009). Manual annotation of the National Corpus of Polish with Anotatornia. *The proceedings of Practical Applications in Language and Computers (PALC-2009). Frankfurt: Peter Lang*, 1–9.
Przepiórkowski, A., Bańko, M., Górski, R. and Lewandowska-Tomaszczyk B. (2012). *Narodowy Korpus Języka Polskiego (National Corpus of Polish).* Wydawnictwo Naukowe PWN, Warszawa. Available online at http://www.nkjp.pl.
Pszczołowska, L. (1970). Boje o rym (*Battles for rhyme*). *Pamiętnik Literacki: czasopismo kwartalne poświęcone historii i krytyce literatury polskiej*, 61(4), 161–177.
Pszczołowska, L. (1972). Rym (*Rhyme*). Zakład narodowy im. Ossolińskich.
Pszczołowska, L. (1973). Sound distribution in rhyme. *Slavic Poetics: Essays in honor of Kiril Taranovsky*, Jakobson et al. (Ed.), Mouton.
Saloni, Z., Gruszczynski, W., Wolinski, M., Wołosz, R. and Skowrońska, D. (2011). *Analizator morfologiczny Morfeusz* (*Morphological analyser 'Morfeusz'*, software). http://sgjp.pl/morfeusz/. Accessed 1 May 2013 .
Shaw, M. L. (2003). *The Cambridge introduction to French poetry*. Cambridge University Press.
Śledziński, D. (2008). Fonetyczno-akustyczna analiza struktury sylaby w języku polskim na potrzeby technologii mowy (Acoustic-phonetic analysis of syllable in Polish for use in speech technology). *Investigationes Linguisticae*, 16: 219–240.
Wachtel, M. (2004). *The Cambridge introduction to Russian poetry*. Cambridge University Press.
Wagner, M., and McCurdy, K. (2010). Poetic rhyme reflects cross-linguistic differences in information structure. *Cognition*, 117(2), 166–175.
Woliński, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN (System of morphological taggers in IPI PAN Corpus). *Polonica*, XXII–XXIII: 39–55.
Woliński, M. (2006). Morfeusz—a practical tool for the morphological analysis of Polish. In *Intelligent information processing and web mining*. Springer Berlin Heidelberg, 511–520.