Raport Badawczy

Research Report

Grammatical rhymes in Polish poetry: a quantitative analysis

K. Opara

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.:   (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2013

Grammatical Rhymes in Polish Poetry: a Quantitative Analysis

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, known as 'Częstochowa' rhymes. It is used to create a ranking of both historic and contemporary Polish poets based on technical quality of their rhymes. 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*, Częstochowa

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).

One of the main challenges of CAP is the introduction of objective quality measures for poetry. Although recently the first steps have been taken towards reaching this goal (Dalvean, 2013), attempts to evaluate poetry and its authors are usually quite subjective. This may lead even to personal conflicts, which is well illustrated in a dispute between teacher and pupil presented in a novel *Ferdydurke* (Gombrowicz 1937):

> — *A zatem dlaczego Słowacki wzbudza w nas zachwyt i miłość? (...) Dlotego, panowie, że Słowacki wielkim poetą był! (...)*
> — *Boże, ratuj, jak ta mnie zachwyca, kiedy mnie nie zachwyca?*
> — *Jak to nie zachwyca Gałkiewicza, jeśli tysiąc razy tłumaczyłem Gałkiewiczowi, że go zachwyca.*

— So why do we love and admire Słowacki? (...) Because a great poet Słowacki was!
— For God's sake, how can I admire his works when I don't admire?
— How can't you admire them, Gałkiewicz, after I told you a thousand times you do admire.

Was Słowacki indeed a great poet? To address this question one can investigate the quality and subtlety of his rhymes, which may indicate the level of his technical mastery.

Polish is a highly inflected language, so parts of speech in the same morphological form have common endings. This introduces trivial, unsophisticated rhymes, which do not reflect the creativity of an author but are a consequence of the inflected structure of Polish. To a very limited extend, this phenomenon occurs also in English. 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'.

Grammatical rhymes in Polish are referred to as 'Częstochowa' rhymes. Piersiak (2008) suggests that this name comes from pious books with verses of questionable quality sold to pilgrims arriving in Częstochowa. This town is famous for the icon of Black Madonna, whose numerous miracles were recognized already in 1716 by pope Clement XI. Does it mean that the poetic vein may depend on the distance from the holy icon? Addressing this question is difficult as current-day Częstochowa poets hardly use rhymes in their verses at all. This is easily noticeable for instance in the magazine Galeria (2012) issued by the Częstochowa Literary Society of Mutual Adoration 'Li-TWA'.

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 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? (...)

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. In section 2, I introduce a quantitative measure of grammatical rhymes in (a set of) poems, which I call 'Częstochowa' score or *Cz*-score. Section 3 provides a comparative, statistical analysis of Częstochowa rhymes in both historical and contemporary Polish poetry. A brief summary of the paper is followed by two appendices covering computational the details of rhyme detection and details of the statistical testing procedure.

2 *Cz*-scores and the Reference Poem

The first step of the investigation consists of detecting rhymes and performing a morphological analysis, which is described in appendix A. In this way I obtained pairs of rhyming words and lists of possible morphological tags, which I present in Table 1. They were taken from the invocation to the Polish national poem *Pan Tadeusz* (Mickiewicz, 1834; translation by George Rapall Noyes):

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 (...)

Holy Virgin, who protectest bright Czenstochowa
and shinest above the Ostra Gate in Wilno!
Thou who dost shelter the castle of Nowogrodek with its faithful folk!
As by miracle thou didst restore me to health in my childhood (...)

Tags depicted in Table 1 describe morphological forms of a word allowed by Polish grammar obtained using the *Morfeusz* tagger (Saloni et al. 2011). For instance '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).

To assess the share of grammatical rhymes in a verse I introduced a measure called 'Częstochowa' score or *Cz*-score. In cases where both rhyming words have only one common tag, it must be a grammatical rhyme and it is assigned $Cz = 100$ points. On the other hand, if the intersection of both tag lists is empty, the rhyme gets no 'Częstochowa' 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 the 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 $Cz = 50$ points. 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 'Częstochowa' 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 'Częstochowa' 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

What is the appropriate share of 'Częstochowa' rhymes in a verse? The literary canon clearly discourages their overuse. On the other hand, elaborate stanza free of grammatical rhymes is found in short verses only. In longer poems, some 'Częstochowa' rhymes are practically unavoidable as their complete abandonment would lead to strange, unnatural constructions. It would also mean the resignation of repetition as a means of expression. Hence, the point is not to avoid grammatical rhymes completely but to avoid overusing them to the point of dominating a verse. Consequently, a question that arises is, what share of 'Częstochowa' rhymes should be acceptable? Choice of this amount is somewhat arbitrary. I decided to base it 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, having 9850 lines, which allows for reliable statistical analyses.

Analysis of the reference poem shows that the share of grammatical rhymes is 29%. Therefore, I adopt value $Cz_{ref} = 29$ as the reference amount of 'Częstochowa' rhymes in Polish poetry. Comparing the amount of grammatical rhymes in a verse with reference value Cz_{ref} requires introduction of a statistical test, which shows if observed differences are significant. For computational reasons, fair tests should balance for different lengths of works. Therefore, statistical testing is based on subsamples from *Pan Tadeusz* of a fixed length of forty-four rhymes. This somewhat arbitrary constant reflects the choice the national poem's author made in *The Great Improvisation* (Mickiewicz 1832):

> *Wybawca narodu,*
> *Z matki obcej, krew jego - dawne bohatery*
> *A imię jego bedzie czterdzieści i cztery.*

> *Reviver of the nation,*
> *Of foreign mother, son of ancient heroes*
> *His name will be forty and four.*

Details of the statistical test are described in Appendix B. Figure 1 shows the distribution of *Cz*-scores in 44-rhyme subsamples of *Pan Tadeusz* as well as the boundaries of the critical set at a confidence level of 44%. Verse is claimed to be statistically significantly different from the reference poem, if its mean share of 'Częstochowa' rhymes lay outside of the critical set, i.e. in the tails of the bell-shaped distribution.
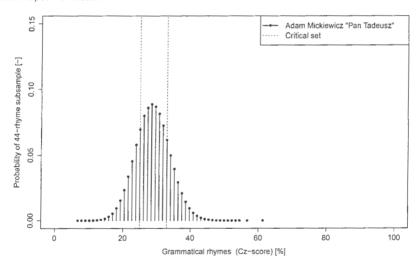

Fig. 1. Share of grammatical ('Częstochowa') rhymes in 10^6 (a million) 44-rhyme random samples of *Pan Tadeusz* by Adam Mickiewicz and the critical set of the 'Częstochowa' test at significance level 44%

3 'Częstochowa' Rhymes in Polish Poetry

The introduction of *Cz*-scores and the discussion of their statistical properties allows one to compare the amount of 'Częstochowa' 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 (in Polish *wieszcz*) Mickiewicz and Słowacki, I included Nobel Prize winners Miłosz and Szymborska, poets Kochanowski and Tuwim, songwriters Osiecka and Kaczmarski as well as the popular disco band 'Weekend'. A more detailed description of the 'Częstochowa' corpus is given in Table 2.

Figure 2 shows the mean share of 'Częstochowa' rhymes 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 indicating an interval, which covers this mean with 99% probability. 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 statistical 'Częstochowa' test, its critical set is denoted with vertical, dashed lines.

The highest share of 'Częstochowa' 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 domskiego ciała.
Bum, bum, bum,
facetowi tak jak rum
chodź ci pokażę swoje totuaże.

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.

Results of the 'Częstochowa' 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 become comparable to the traditional song *Morskie opowieści* (a shanty to the tune of *What shall we do with a drunken sailor?*).
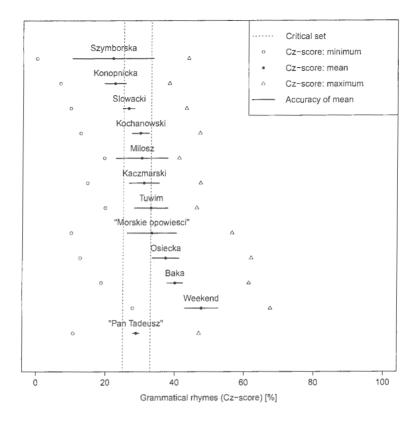
Fig. 2. Ranking of selected authors according the amount of grammatical ('Częstochowa') rhymes in their poetic works

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

Cny młodziku, migdaliku,
Czerstwy rydzu, ślepawidzu (...)
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 'Częstochowa' rhymes.

Table 2. Description of 'Częstochowa' corpus analysed in this contribution

Author	Pieces	Artistic period	Remarks
Jan Kochanowski	*Fraszki*	Renaissance	'Master of Czarnolas'
Józef Baka	*Uwagi śmierci niechybnej*	Baroque	Catholic priest
Adam Mickiewicz	*Pan Tadeusz*	Romanticism	National poet
Juliusz Słowacki	*Beniowski*	Romanticism	„a great poet was"
Unknown	*Morskie opowieści*		Popular shanty
Maria Konopnicka	Poetry for children	Modernism	Poet
Julian Tuwim	*Bal w operze*	Interbellum	Parliament proclaimed 2013 Tuwim's year
Czesław Miłosz	Verses	Contemporary	Nobel Prize 1980
Wisława Szymborska	Verses	Contemporary	Nobel Prize 1996
Agnieszka Osiecka	Songs	Contemporary	Poet, writer
Jacek Kaczmarski	Sung poetry	Contemporary	Poet, composer, bard of 'Solidarity' movement
'Weekend'	Discography	Modern day	Disco stars

The next place in the ranking goes to Agnieszka Osiecka, an author of popular songs and essays. This result seems quite surprising, as her poetry is highly regarded. For instance, in year 2002 foundation 'Okularnicy' was established to protect and promote her heritage. This shows that the amount of grammatical rhymes is only one of many factors influencing the quality of verses.

Czestochowa scores for most poets are not significantly different from those of a 44-rhyme subsamples of *Pan Tadeusz*. Interestingly, works of Adam Mickiewicz and Jan Kochanowski have very close *Cz*-scores, despite two and a half centuries of language development. 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 orchitekturą 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.

The question from *Ferdydurke* by Gombrowicz here resurfaces: was Słowacki a great poet? The architecture of his rhymes is indeed impressive, but 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.

The lowest share of Czestochowa rhymes is due to Maria Konopnicka. Her 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 poetry and hence the uncertainty in determining their mean share.

4 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, called 'Częstochowa' rhymes, and calculate their share in the investigated works. The Polish national poem, *Pan Tadeusz* by Adam Mickiewicz, was used as a reference text to develop a statistical 'Częstochowa' 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.

Analyses of this kind can be done manually. Their credibility is then supported by semantic analysis of respective words and phrases. However, this requires extensive amount of work and analysis of longer pieces of poetry would become tedious and tiring. Limiting the analysis to some selected sections reduces the time burden, but quantitative information obtained in this way does not necessarily represent the patterns that are found across the whole text (Mahlberg and Smith, 2012).

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

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).
Galeria (2012). *Częstochowski magazyn literacki Literackiego Towarzystwa Wzajemnej Adoracji „Li-TWA" (Częstochowa Literary Magazine of the Literary Society of Mutual Adoration 'Li-TWA'),* number 24. http://www.litwa.czest.pl/archiwum-galerii/. Access 17 June 2013.
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.
Gombrowicz, W. (1937). Ferdydurke (a novel).
Greene, E., Bodrumlu, T. and Knight, K. (2010). *Automatic analysis of rhythmic poetry with applications to generation and translation.* In proceedings of 2010 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.
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. (1832). *Dziady część III (Forefathers' Eve part III)*, chapter Wielka Improwizacja (Great improvisation, a poem).

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).

Piersiak, T. (2008). *Z „pana Częstochowy" nie możemy być dumni ("Mr Częstochowa" brings us no pride*, newspaper article). Gazeta Wyborcza, 23 October 2008.

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 .

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.

Woliński, M. (2003). *System znaczników morfosyntaktycznych w korpusie IPI PAN (System of morphological taggers in IPI PAN Corpus).* Polonica, XXII–XXIII: 39–55.

Appendix A: Rhyme Detection

There is little literature related to rhyme detection, probably due to the fact that CAP 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. To simplify the study, only the line-final rhymes were detected.

The last word in each verse was compared with the last words of the preceding four verses. 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).

Appendix B: Derivation and Properties of the 'Częstochowa' Test

The reference poem, *Pan Tadeusz* by Adam Mickiewicz, 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 44 rhymes. Distribution of *Cz*-scores resulting from imposing such a constraint was investigated by drawing 10^6 random subsamples, whereas 10^6 or 'Milijon' is another constant suggested in *Great improvisation* (Mickiewicz, 1832):

> *Nazywam sie Milijon – bo za milijony kocham i cierpię katusze.*
> *My name is Milijon – because for millions I love and I suffer agonies.*

In this way I obtained a distribution of *Cz*-scores in 44-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 'Częstochowa' 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 $\alpha = 44\%$.

It appears that the 'Częstochowa' test is nearly equivalent to the common *t* test. According to the Central Limit Theorem, the distribution of the mean of *Cz*-scores in 44-rhyme samples tends to normality. To see if 44 is a sufficiently large number, I checked 'Milijon' (10^6) 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 44-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

$$\sigma = \sqrt{\frac{p_0 \cdot 0^2 + p_{50} \cdot 50^2 + p_{100} \cdot 100^2 - \mu}{44}} = \sqrt{\frac{p_{50} \cdot 50^2 + p_{100} \cdot 100^2 - \mu}{44}} \qquad (1)$$

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

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

The square root of 44 in the denominator of equation (1) reflects the calculation of a mean from a sample of 44 independently chosen rhymes.