Nikola Bakarić, Davor Nikolić

Dataset of stylistic features of Croatian folklore genres

Nabor podatkov o slogovnih značilnostih hrvaških folklornih žanrov

This paper aims to present the creation of a small dataset derived from the computational analysis of Croatian folklore genres. The data was extracted from a collection of charms, counting-out rhymes, tongue twisters, blessings, curses and proverbs using script for automated syllabification and n-gram extraction. The dataset consists of syllable and n-gram ratios and the numerical representation of other stylistic features.

◆ **Keywords**: feature extraction, machine learning, computational stylistics, Croatian folklore genres

Namen prispevka je predstaviti oblikovanje majhnega nabora podatkov, ki izhaja iz računalniške analize hrvaških folklornih žanrov. Podatki so bili pridobljeni iz zbirke zagovorov, izštevank, lomilcev jezika, blagoslovov, kletvic in pregovorov z uporabo skript za samodejno silabizacijo in ekstrakcijo n-gramov. Nabor podatkov sestavljajo razmerja zlogov in n-gramov ter številčna predstavitev drugih slogovnih značilnosti.

◆ **Ključne besede**: ekstrakcija značilnosti, strojno učenje, računalniška stilistika, hrvaški folklorni žanri

1 Introduction

As part of an ongoing research effort which combines the stylistic analysis of Croatian oral literature and natural language processing methods, the authors have constructed a dataset based on a collection of Croatian folklore genre texts. Previous research (Nikolić and Bakarić 2016) has shown that the automated classification of these types of text is possible when observing the text at a sub-word level, the level of the syllable and the phoneme n-gram. Computational stylistic (Rybicki et al. 2016) analysis allows for further examination of other stylistic features such as rhyme (Hoorn et al. 1999) and other figures of style.

The creation of the dataset is part of research for a doctoral thesis, which aims to design a model for classifying Croatian folklore genres based on stylistic features and the use of machine learning classification algorithms.

2 Description of the corpus

The collection of charms, counting-out rhymes, tongue twisters, blessings and curses are part of the Archive of manuscript collections of the Chair for Croatian oral literature at the University of Zagreb's Faculty of Humanities and Social Sciences. An additional collection of modern Croatian language texts containing prose and newspaper articles was added to the collection as a test baseline for classification purposes.[1] The collection of texts was then processed and normalized to create a unified corpus. Since the focus of the research is on short rhetorical genres[2], which can range in length from 15 to 200 characters, other text forms were adjusted to match them in size in order to make them comparable.

The corpus comprises 66 charms, 321 counting-out rhymes, 76 tongue twisters, 125 proverbs, 72 blessings and 342 curses, totalling 8,700 word tokens. Additionally, it includes 25,000 tokens of modern language text.

3 Selection of stylistic features

Even a casual glance at folklore genres detects a certain order and structure in their sound and syllable organization, while a more focused analysis reveals that some genres are more structured than others, especially in terms of sound and syllable repetition. The original purpose of this dataset and the tools used to create it was to measure the levels of sound repetition with the purpose of determining euphony in the poetic expression of poets such as A. G. Matoš (Nikolić and Bakarić 2018).

The first observed stylistic feature group was the repetition coefficient of sounds. The method used to measure the coefficient requires that the segments of the analysed corpora be of similar size for the comparison of results to be meaningful. It is impossible and thankless to talk about constants because the repetition coefficient only makes sense when observed in relation to other elements (due to the mathematical model). It can be further divided into two approaches for measuring repetition – n-gram repetition, which observes sounds (or the approximation of sounds through characters), and syllable repetition. N-gram repetition is independent of any kind of prosodic knowledge and therefore much simpler to measure. Syllable repetition requires that the text is separated into syllables, which is not a trivial task for the Croatian language. The problem of syllabification has several levels but can be described by a rule-based algorithm, which first solves specific and rare syllable border problems and then moves on to more general and common syllable types. Issues that affect the syllabification results are related to the difference between spoken and written text and, even though

[1] Although the folklore texts were recorded in all of Croatia's regions (with relatively high dialectal diversity), the phonological encoding of the texts was comparable to the encoding for the texts written in standard Croatian language. In most cases the original folklore records were not transcribed using narrow phonological transcription because the variations in vowel quality were not relevant for the meaning.

[2] *Rhetorical genres* is a term used by many Croatian folklorists, especially those with a philological background, to refer to a group of relatively small genres, which are characterized by elements of persuasion and verbal playfulness (Čubelić 1970; Kekez 1996, 1998; Botica 1995, 2013). Verbal charms, toasts, counting-out rhymes, tongue-twisters, blessings and curses are considered to be typical rhetorical genres, which use their aesthetic qualities to produce specific communicative effects (Nikolić 2019: 70).

Croatian spelling is mostly phonetic, this should be taken into consideration when preparing the input text. Repetition of sounds contributes to euphony, in the form of alliteration, assonance and rhyme (Bishop 1985: 345), and is a distinguishing feature when comparing and classifying text.

An additional stylistic feature used to detect levels of euphony is the pronunciation difficulty coefficient. This feature looks at two phenomena, which contribute to the difficulty of pronunciation within a given text. The authors focused on the syllabic r and consonant clusters within syllables. Reporting the occurrence of these phenomena also requires syllabification and the additional analysis of syllable quality, which is an integral part of the syllabification algorithm. Although essentially quantitative, the pronunciation difficulty coefficient still includes some qualitative elements. This concerns the counting of specific elements previously recognized as non-euphonic (syllabic r and consonant groups), but still neglects the vocal composition within the consonant groups. The coefficient does not differentiate between sets of voiceless and voiced consonants. A thorough analysis of tongue twisters confirmed a high proportion of sets with voiceless consonants (Nikolić 2013: 119–120, 2019: 200), which is an indisputable indicator of pronunciation difficulty and affects the overall reduced impression of euphony.

Computational stylistics is described as a data-driven field, which promises new insights into texts through digital methods (Herrmann et al. 2021). However, to make its contribution to analysing euphony more interesting and useful to literary historians and prosodists, it is necessary to link the results of objective measurements at the phonetic level of the text with the content level. The results of such research could confirm (or perhaps refute) euphony as a distinguishing feature between genres, or even between individual authors of written literature. Using a dataset such as the one presented here as a distant reading tool can assist in detecting phenomena, which can then be focused on and further analysed.

4 Feature extraction

Corpus preparation included changing characters to lowercase and transcribing text with the goal of having one character represent one phoneme. This is important for accurate syllabification, one of the steps of feature extraction. Simple regular expressions were used to replace the digraphs lj, nj and dž with ḷ, ń and ǯ respectively. A similar method was used to mark the syllabic r as any r between two consonants. An additional attempt was made to solve the difference in spelling and pronunciation of accentuated and non-accentuated words by removing the spaces between the words in question. However, it is still unclear to what extent this impacts the final results.

The corpus was then processed in order to extract the stylistic features of each entry. Syllabification was carried out using the custom syllabification algorithm written in Python (Bakarić and Nikolić 2017), mentioned in the previous paragraph. Using the algorithm, the authors were able to create a statistical overview of the genres and extract features from individual entries. They were able to observe the frequency of syllable types for each of the genres, as can be seen in Table 1. C represents consonants and V vowels. The syllabification algorithm differentiates between consonant groups and uses them to define some of the syllabification rules.

	Charms	Counting-out rhymes	Curses	Blessings	Proverbs	Tongue twisters
CV	0.60	0.65	0.65	0.61	0.67	0.61
V	0.15	0.06	0.10	0.09	0.07	0.05
CCV	0.14	0.11	0.17	0.18	0.16	0.15
CVC	0.08	0.13	0.06	0.09	0.06	0.13
CCVC	0.01	0.02	0.01	0.01	0.01	0.04
VC	0.01	0.02	0.01	0.01	0.01	0.00

Table 1: Frequency of syllable types across genres.

The syllabification script was extended to include an algorithm, which calculates the type/token ratio for unigrams, bigrams, trigrams, syllables, meta-syllables, consonant clusters and syllabic /r/ coefficients and rhyming coefficients. The script can be further developed to include other features, such as syllable quality (e. g. open/closed syllable ratio) or their frequency, which can be seen in Table 2.

	Charms	Counting-out rhymes	Curses	Blessings	Proverbs	Tongue twisters
1.	i	ka	da	ti	je	po
2.	u	je	o	i	ne	pe
3.	ko	ti	ti	bo	na	je
4.	o	li	la	go	i	ko
5.	su	ci	bo	o	u	ka
6.	ne	i	te	vi	ni	o
7.	na	ko	gda	da	ti	pa
8.	ka	na	i	ne	ma	na
9.	a	se	u	bla	ka	pi
10.	si	po	se	bog	se	la

Table 2: Top 10 syllables across genres.

Phonostylistics, a sub-discipline of stylistics, delves into stylistic effects at the phonetic and phonological levels (Trubetzkoy 1971). It analyses literary texts with a primary focus on the repetition of sound patterns on multiple levels: the repetition of metrical patterns, syntactic parallelism, the repetition of motifs or key words and the repetition of certain sounds or sound sequences (alliteration, assonance, poetic homophones), as can be seen in Figure 1. As mentioned earlier, n-gram and syllable repetition represent two approaches to detecting stylistic features caused by sound sequences. While n-gram analysis does not require language knowledge, it is unable to observe metrical patterns and sound sequences on the same level as the syllable analysis approach. It allows a comprehensive qualitative exploration of syllables, including analysis of consonant clusters, the syllabic r, open and closed syllable ratios, to mention a few.

The design of the syllabification algorithm allows for further qualitative analysis with minimal intervention. The syllabification rules are based on dividing consonants into groups according to their production and sonority. This approach offers a finely-tuned qualitative method that might yield further insights into the nature of phonostylistic phenomena.

Ni u moru mjere, ni u mački vjere. (alliteration)
Pusti koku na policu, ona će i na stolicu. (assonance)
Od zbora do tvora ima prostora. (poetic homophones)

Figure 1: Example of sound sequences in proverbs.

5 Dataset structure

The main idea behind the creation of this dataset was to describe stylistic features as numerical values, which can then be analysed, compared and used as a testing ground for data driven analysis at the sub-word level.

The dataset is formatted as a table with the following fields (from left to right):

Text entry – original text entry prior to processing (type: text)

The input text entry is prepared for processing in the previous step. Preparation includes transcription of digraphs, switching to lowercase and marking the syllabic r. The input text is fed to the processing algorithm as a plain text unicode UTF-8 standard file. Individual inputs are separated by new lines and each genre is contained in a separate file.

Unigram – type/token ratio of unigrams or single characters subtracted from 1 (type: number, interval 0-1)

The algorithm determines the frequency of single characters, which, at this point, represent sounds. The output is the type/token ratio which is then subtracted from 1 in order to represent the repetition coefficient.

$$K_{Unigram} = 1 - \frac{Type_{Unigram}}{Token_{Unigram}}$$

Bigram - type/token ratio of bigrams, two consecutive characters in a word, subtracted from 1 (type: number, interval 0-1).

Here we determine the type/token ratio of pairs of sounds in the input text. The pairs are created by looking at consecutive pairs of characters with overlap. For example, an input text consisting of six characters (without spaces) would produce five distinct bigrams. The number of bigram tokens is therefore n-1, where n is the number of characters in the input text. Subtracting the type/token ratio from 1 yields the bigram repetition coefficient.

$$K_{Bigram} = 1 - \frac{Type_{Bigram}}{Token_{Bigram}}$$

Trigram – type/token ratio of trigrams or three consecutive characters in a word (type: number, interval 0-1).

The trigram determines the type/token ratio of sequences of three sounds in the input text. Character trigrams are created by observing consecutive triplets of characters with overlap. Therefore, the number of trigram tokens is n-2, where n is the number of characters in the input text.

$$K_{Trigram} = 1 - \frac{Type_{Trigram}}{Token_{Trigram}}$$

Syllable – type/token ratio of all syllables in an entry subtracted from 1 (type: number, interval 0-1)

The syllable repetition coefficient represents the type/token ratio of syllables subtracted from 1. The algorithm calculates the type/token ratio after the syllabification procedure.

$$K_{Syllable} = 1 - \frac{Type_{Syllable}}{Token_{Syllable}}$$

Meta-syllable – type/token ratio of all syllables transcribed as only vowels and consonants (type: number, interval 0-1)

The meta-syllable coefficient is calculated after the syllabification and transcription of syllable elements (consonants are transcribed as C, vowels as V). Here all syllables are reduced to categories such as CV, CCV, V, etc. After that, the type/token ratio is calculated and subtracted from 1.

$$K_{Meta-syllable} = 1 - \frac{Type_{Meta-syllable}}{Token_{Meta-syllable}}$$

CC-R – ratio of syllables containing consonant clusters or syllabic /r/ (type: number, interval 0-1)

The pronunciation difficulty coefficient is calculated by counting the syllables, which contain consonant clusters (a group of more than one consecutive consonant in a syllable, such as CCV) and syllables which contain the syllabic r. We then calculate the type/token ratio, where type is the total number of syllables containing either consonant clusters or the syllabic r and the token is the total number of syllables.

$$CCR = \frac{Type_{CCR}}{Token_{CCR}}$$

Rhyme coefficient – ratio between all the words and the rhyming words in an entry (type: number, interval 0-1)

The rhyme coefficient was added to the algorithm later and is not part of the original script (available on Github (Bakarić and Nikolić 2017)). It calculates the ratio between words that end with the same syllable and the total number of words in a text entry.

Length – length of text entry under 1 (type: number, positive integer)

The length of text entry was calculated after processing using Excel's len() function. It includes interpunction and spaces as contained in the original entry. This auxiliary field is used for filtering entries of similar or identical lengths, which is important for separating the dataset into comparable subsets.

Class – genre of text entry (type: text, list of names [charm, counting-out rhyme,…, modern language text])

The last field contains the entry class or genre. Ultimately, using the data described in the previous fields, we will train a machine learning model that will be able to classify the entry by genre.

The structure of the dataset allows for statistical analysis and preparation for further processing. It should be taken into consideration that machine learning classification algorithms are very sensitive to data distribution (Hastie et al. 2009) and most of the values in this dataset are skewed in some way. The analysis of the dataset can show which parameters of machine learning algorithms should be considered and how the data should be approached.

Table 3 is a representation of the algorithm output for a single entry, in this case a tongue twister. The algorithm uses this output to create the dataset, but it can be modified to produce several different outputs beyond the dataset described in this paper.

petar petru plete petļu	
unigram	e:5 p:4 t:4 r:2 u:2 a:1 l:1 ļ:1
bigram	et:4 pe:3 ta:1 ar:1 tr:1 ru:1 pl:1 le:1 te:1 tļ:1 ļu:1
trigram	pet:3 eta:1 tar:1 etr:1 tru:1 ple:1 let:1 ete:1 etļ:1 tļu:1
syllable	pe:3 tar:1 tru:1 ple:1 te:1 tļu:1
meta syllable	kv:4 kkv:3 kvk:1

Table 3: Example of the extracted data for a tongue twister.

The authors have made the dataset publicly available for download in CSV and XLSX formats using a data repository (PUH 2022) (Bakarić and Nikolić 2023). The syllabification Python script is publicly available on GitHub (Bakarić and Nikolić 2017). There are plans to extend the dataset with lyrical and epic poetry, as well as poetry written by named authors.

5 Conclusion

The dataset, as seen in Table 4, was primarily designed for the machine classification of Croatian folklore texts. However, the described methods and the dataset itself can be used as a basis for drawing general conclusions on stylistic features in other texts, a direction the authors plan to take.

The presented process depicts the steps taken to build a dataset of numerically described stylistic features of selected Croatian folklore texts. Building a dataset is the crucial first step of many natural language processing research activities. Its design and structure can have a great impact on research results. The dataset presented here is limited in scope but can serve as a starting point for future efforts in the computational stylistic analysis of Croatian language texts.

Entry	Unigram	Bigram	Trigram	Syllable	Meta-syll.	CC and R	Length	Class
Enci benci nakamenci šija bija kompanija van!	0,66	0.35	0.25	0.24	0.82	0.00	45	counting-out rhyme
Četristočetrdesetčetiri čavke čuče načamcu ičopaju čičke.	0,65	0.22	0.05	0.09	0.83	0.22	57	tongue twister
Vrakte drpo išćipo cjelu noć.	0,25	0.05	0.00	0.10	0.60	0.50	29	curse
Višeti Bog dao negunjega želila.	0,42	0.05	0.00	0.00	0.77	0.00	32	blessing

Table 4: Example of the dataset.

References

Bakarić, Nikola ; Nikolić, Davor, 2017 : Croatian syllabification script. Online: https://github.com/nbakaric/Croatian-syllabification (retrieved 3.12.2023).

Bakarić, Nikola; Nikolić, Davor, 2023: Stylistic features of Croatian folklore genres. Online: https://puh.srce.hr/s/ygbcraCJCnprPtT (retrieved 3. 12. 2023).

Bishop, Lloyd, 1985: Euphony: A New Method of Analysis. *Language and Style* 18/4, 343–362.

Botica, Stipe, 1995: *Hrvatska usmenoknjiževna čitanka*. Zagreb: Školska knjiga.

Botica, Stipe, 2013: *Povijest hrvatske usmene književnosti*. Zagreb: Školska knjiga.

Čubelić, Tvrtko, 1970: *Usmena narodna retorika i teatrologija*. Zagreb: s. n.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome, 2009: Model Assessment and Selection. In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. 219–257.

Herrmann, J. Berenike; Jacobs, Arthur M.; Piper, Andrew, 2021: Computational Stylistics. In: Kuiken, Don; Jacobs, Arthur M. (eds.), *Handbook of Empirical Literary Studies*. Berlin, Boston: De Gruyter. 451–486.

Hoorn, J. F.; Frank, S. L.; Kowalczyk, W.; van der Ham, F., 1999: Neural network identification of poets using letter sequences. *Literary and Linguistic Computing* 14/3, 311–338. DOI: https://doi.org/10.1093/llc/14.3.311.

Kekez, Josip (ed.), 1996: *Poslovice, zagonetke i govornički oblici*. Zagreb: Matica hrvatska.

Kekez, Josip, 1998: Usmena književnost. In: Škreb, Zdenko; Stamać, Ante (eds.), *Uvod u književnost: Teorija, metodologija*. Zagreb: Nakladni zavod Globus. 133–192.

Nikolić, Davor, 2013: *Fonostilistički opis hrvatske usmenoknjiževne retorike*. Doctoral thesis. Zagreb: University of Zagreb.

Nikolić, Davor, 2019: *Između zvuka i značenja: fonostilistički pristup hrvatskim usmenoretoričkim žanrovima*. Zagreb: Disput.

Nikolić, Davor; Bakarić, Nikola, 2016: What Makes Our Tongues Twist?: Computational Analysis of Croatian Tongue-Twisters. *Journal of American Folklore* 129/511, 43–54. DOI: https://doi.org/10.5406/jamerfolk.129.511.0043.

Nikolić, Davor; Bakarić, Nikola, 2018: Korelati eufonije u Matoševim sonetima. In: Botica, Stipe et al. (eds.), *Šesti hrvatski slavistički kongres. Zbornik radova. Vol. 2*. Zagreb: Hrvatsko filološko društvo, 801–811.

PUH data storage and management, 2022. University computing centre-SRCE. Online: https://www.srce.unizg.hr/en/puh (3.12.2023).

Rybicki, Jan; Hoover, David L.; Eder, Maciej, 2016: Computational Stylistics and Text Analysis. In: Crompton, Constance; Lane, Richard J.; Siemens, Ray (eds.), *Doing Digital Humanities: practice, training, research*. London, New York: Routledge, 123–144.

Trubetzkoy, Nikolay Sergeyevich, 1971: *Principles of phonology*. Berkley, Los Angeles: University of California Press.