Oral Tradition, 36/1 (2023):37-62

Type-Token Ratio and Entropy as Measures to Characterize a Forgery of Oral-Formulaic Epics

David L. Cooper, Demetry Ogoltsev, and Michal Ondrejcek

The Queen's Court and Green Mountain Manuscripts (*Rukopisy královédvorský a zelenohorský*, together abbreviated "RKZ" in Czech) present an unusually successful case of literary forgery. These pseudo-medieval Czech manuscripts, presenting folk lyrics, ballads, and epic songs seemingly recorded in the late-thirteenth and in the ninth to tenth centuries, respectively, were taken by most scholars to be genuine monuments of medieval poetry, over the objections of only a few dissenters, for at least sixty years following their discovery in 1817 and 1818. By the late nineteenth century, the manuscripts' imitation of the Old Czech language no longer convinced experts in Czech historical linguistics, who led in their unmasking.[1] Along with the Old Czech language, the epic poems in the manuscripts also imitated the forms of traditional oral epic poetry, familiar at the time from Vuk Karadžić's first publications of Serbian songs (1814 and 1815) and similar traditions (Russian *byliny* and Homeric epics). Their successful imitation of these models, and the limits of that success, have not been sufficiently studied and understood. In this article, we present the results of a project that aimed at characterizing South Slavic and Russian oral-formulaic poetry using informatics measures, such as entropy, and standard natural language processing measures like Type-Token Ratio (TTR) for measuring lexical variance, for comparison with their imitation in the Czech forged manuscripts.

The Manuscripts

Václav Hanka "discovered" the first of the two manuscripts while exploring the vault in a church tower in Dvůr Králové on September 16, 1817. The second was mailed anonymously in November, 1818, to the Highest Burgrave of the Bohemian kingdom for inclusion in the collections of the new national museum. Details emerged later showing it had been uncovered in the castle Zelená Hora. The Queen's Court (Královédvorský) Manuscript contained six epic poems (one fragmentary), two ballads, and six lyric songs, while the Green Mountain (Zelenohorský) Manuscript contained what at first appeared to be fragments of two epic songs but are better understood as one (relatively) complete epic. The leading Czech expert at the time

[1] See Cooper 2018 for a bilingual edition of the manuscripts, which includes an introduction that outlines their reception and influence in Czech culture from the time of their discovery to their unmasking in the mid-1880s.

in Slavic antiquities, Josef Dobrovský, welcomed the discovery of the first manuscript but doubted the second, which was much older than any known fragments of writing in Czech and presented evidence of a highly developed Czech culture in that early period. Soon, he would name his students in Slavistics, Hanka and his roommate Josef Linda, along with Josef Jungmann, as the likely promulgators of the fake (Ivanov 1969:199). Nearly 200 years of subsequent investigation have yielded no direct evidence of the conspiracy linking it to its authors, but the indirect and circumstantial evidence (linguistic, literary, historical, and psychological) indicating Hanka's involvement is overwhelming. The cases for the involvement of other possible collaborators are less decisive.

The form of several of the epic songs of the manuscripts suggests that their models were the South Slavic oral epics in *deseterac* (ten-syllable, unrhymed) form. The single epic of the Green Mountain Manuscript (*Libuše's Judgment*) and two of the epics from the Queen's Court Manuscript (*Oldřich and Boleslav* and *Jaroslav*) are entirely in ten-syllable, unrhymed verse. Two additional epics from the Queen's Court Manuscript, otherwise written in lines of continually varying syllable length, include significant passages of consecutive lines in ten-syllable form (*Čestmír and Vlaslav* and *Záboj, Slavoj, and Luděk*).[2] As we shall see, other features of the poetic discourse also indicate a relationship to the South Slavic oral epic form. The South Slavic tradition of oral epics in the shorter *deseterac* line were being presented at that time by Vuk Karadžić to a European reading public eager for examples of native epic poetry, and his songs quickly became a hit. Václav Hanka was among those at the very forefront of this enthusiastic reception.

In the autumn of 1813, both Hanka and Karadžić arrived in Vienna and came under the tutelage of the Slovenian scholar Jernej Kopitar. Hanka was continuing his law studies at the university and was recommended to Kopitar by Dobrovský, with whom he had recently begun to study Slavistics. Karadžić was more advanced than Hanka, and with Kopitar's encouragement published both his new grammar of Serbian and a first small collection of folksongs in 1814, followed by a second and larger collection of songs in 1815 (Dolanský 1968:15-16). Hanka responded to Karadžić's first volume by calling on the Czechs to collect their own folksongs in a short, anonymous article in a Vienna-based Czech periodical. The article was followed by Hanka's translation of one of the lyric songs from the collection (Máchal 1918:xix-xx; Cooper 2010:96-97). Kopitar reviewed the second volume in 1816 in *Wiener allgemeine Literaturzeitung* and included translations of five songs into German. Hanka translated these same five into Czech, adding two additional lyrics from the volume and one from the first collection, and published it in 1817 as *Prostonárodní srbská muza, do Čech převedená* (*The Serbian Folk Muse, Led Over into Bohemia*) (Dolanský 1968:22-23).[3] Of the eight songs Hanka translated, three

[2] Our analysis ignores single lines and couplets of isolated ten-syllable lines and only includes passages of three lines or longer in ten-syllable form. For *Čestmír* this includes eighty-four of the 229 lines of the poem; for *Záboj* just twenty-one of 255 lines. The remaining two epic poems use, in one case, a native Czech eight-syllable line and, in the other, a ballad-like stanza, both also unrhymed.

[3] The small volume also included translations of two Russian songs.

were in the *deseterac* form of traditional epic.[4] Hanka, following Kopitar, thus preceded with his translation the later and more substantial efforts by the brothers Grimm (in Förster and Tieck 1818) and Sir John Bowring (1827) (Karadžić 1997:11). And, Hanka was completing his translations of Karadžić's epic and lyric songs in the same period as the epic and lyric songs of the forged manuscripts were likely being composed.

The imitations of Slavic oral epic poetry in the Manuscripts proved at the time very convincing to expert judges, both foreign and domestic, and in studies of the poems, the resemblance was remarked and served as a sign of the ancient origin of the poems and the authenticity of the manuscripts. František Palacký, the first historian and honorary "father" of the Czech nation, connected the ten-syllable epics explicitly to Slavic epic forms in 1829, and the lack of rhyme was an important authenticating factor for him as well (Dobiáš et al. 2015:310ff.). The critic Václav Nebeský declared, in an extended discussion of the manuscript published in 1852-53, that "every line of the Queen's Court Manuscript is certainly also a witness to its authenticity," and he cited the recognizability of the poems to Serbs, Russians, and other Slavs as corroborating evidence (Dobiáš et al. 2015:373-74). But how good, in fact, are these Czech imitations of Slavic traditional oral epic poetry? Thanks to the work of Milman Parry and Albert Lord and its extensions by John Miles Foley (1990 and 1991), Patricia Arant (1990), and others, we know far more today about the "oral-formulaic" form and nature of Slavic oral epic poetry, about how it is composed using a limited but flexible repertoire of formulaic patterns at the level of phraseology (formulas), pattern scenes (themes), and story (song) patterns, than any imitators in the early-nineteenth century could have penetrated. The South Slavic tradition, which is the primary model for the epics in the Manuscripts, has been particularly intensively studied and described. From our contemporary perspective, then, how well did the forgers do?

Imitating Folk Epics

The answer is: very well in some respects and not so well in others.[5] For one thing, it is likely impossible to create an oral-formulaic verse form if one does not exist in your language (as we now know to be the case for Czech folk tradition) and if, instead, one is translating and copying a verse pattern from a language that has different meter-forming stress and length characteristics. Foley revises Parry and Lord's description of the formulaic phraseology of oral-traditional epic, seeking a set of general rules that would account not only for formulas but also for the wider spectrum of poetic diction, including formulaic systems and grammatical and syntactic patterns.[6] He proposes a set of traditional rules (which he formulates for ancient Greek, South Slavic, and Old English epics) based on the prosody of the verse form in each language. These rules account for the metrically sound generation of every verse line, while formulaic

[4] These ranged from forty to 201 lines in length in Hanka's translation. See Tureček 2015 on how Hanka's translation represented Serbian songs.

[5] Cooper has prepared an in-depth study on this topic for a book-length study of the Czech Manuscripts, from which the following short summary and examples have been taken.

[6] See note 12, below, for the analysis of a single line and the multiplicity of patterns that converge there.

patterns, sound patterns, syntactic parallelism, and other verse characteristics are second-level shaping processes (Foley 1990:173-75). Such traditional rules are impossible to formulate for the Czech epics in the manuscripts, the verse pattern for which derives from an implicit translation from the *deseterac*. As Roman Jakobson (1935) showed long ago, the verse in the Manuscripts' epics follows a pattern that was already visible in Hanka's translations of Karadžić's epics. Because Czech word stress is immobile, fixed on the first syllable of the word, and vowel length only plays a secondary role in meter—both unlike in Bosnian-Croatian-Serbian—the verse characteristics change in the translation. Unlike in South Slavic verse, and *also unlike in native Czech ten-syllable verse* (which existed only in genres other than epic), in Hanka's translations and in the verse of the Manuscripts, there is a higher incidence of stress on the seventh syllable than on what should be the metrically marked ninth syllable. This translated and imitated verse, then, is revealed as non-native in its metrical features (lacking what Foley calls the Indo-European principle of right justification),[7] a sure sign for Jakobson that the verse is formed on an implicit orientation toward the *deseterac* (Jakobson 1935:48-49). Not only, then, does the Czech verse not follow Foley's traditional rules for South Slavic verse (it cannot, the language having different metrical characteristics); it also exhibits features that would greatly challenge the formulation of its own traditional rules (the higher incidence of stress on the seventh syllable is highly unlikely to be traditional and native, given its deviation from the pattern in native Czech ten-syllable verse; the caesura is also mobile in a manner that varies unpredictably from poem to poem, in contrast to the strictly fixed caesura of the *deseterac*).

On the other hand, the forgers were quite successful in imitating the rich texture of the phraseology of the epic discourse, the formulaic patterns, sound patterns, syntactic parallelism, terracing, and thematic focus that further shape the phraseology, according to Foley, within the context of the traditional rules. The opening lines of the first song in the Queen's Court Manuscript, *Oldřich and Boleslav*, establish the discourse patterns well (ll. 1-9, 49-61):[8]

. . . sě v *črn les*	. . . into the dark forest
tamo, kam[o] sě vládyky sněchu,	To the place where the lords had assembled,
sedm sich vládyk s udatnými sbory.	Seven such lords with stalwart companies.
Výhoň Dub tamo s niem [s] snahú chváta	Výhoň Dub speeds there eagerly
se vsjú chasú svojú temnem nočniem.	With his entire band in the dark of night,
Sě chasa mu bieše na sto chlapóv,	A band of about a hundred men,
vsěch sto jmieše v nožnách *břietné meče*,	All hundred had a sharp sword sheathed,
k mečém vsěch sto jmieše *mocná paže*,	And a strong arm each one for the sword,
k Výhoňu v útrobách statnú vieru.	And stout faith in Výhoň in their guts.

Two-thirds of the lines (1, 3, 5, 7, 8, 9) end in adjective-noun combinations that could easily be formulaic (those italicized are quasi-formulaic in the Manuscripts, being repeated in two cases

[7] Right justification names the tendency of verses to be more flexible in their metrical and formulaic aspect at the beginning of the verse line (on the left) and gradually more inflexible, strictly aligned with the meter, and formulaic as one moves toward the end of the line (on the right).

[8] This epic is fragmentary, as the partial first line begins the first full page of the manuscript but is preceded by two pages that have been cut away, leaving text only on the inner margins. So, this is really the beginning of the end of the epic. Text and translations are from Cooper 2018. The same discourse patterns could be demonstrated in the opening lines of *Libuše's Judgment* for the Green Mountain Manuscript.

with synonymous rather than identical adjectives).[9] The *črn les* ("black forest") of the first line is part of a thematic focus on a gathering in a dark forest to conspire against an illegitimate or foreign power and recurs in that function in the opening of *Záboj, Slavoj, and Luděk*. Many of the remaining adjective-noun combinations relate to another thematic focus in this opening: a descriptive theme that fleshes out the accoutrements of the *udatní sbory* ("stalwart companies"), with their *břietné meče*, *mocná paže*, and *statnú vieru* ("sharp swords," "strong arms," and "stout faith"). The passage is also characterized by repetitions occurring at multiple levels, from assonance and alliteration and other sound patterning (*tamo-kamo*, *sněchu-snahú-statnú*, *chváta-chasa-chlapóv*, *bieše-jmieše-meče-paže*, *nožnách-útrobách*); to preposition repetitions (*s*, "with," in the fourth line, with the second instance added as necessary by the editors)[10] and word repetitions, sometimes repeated through grammatical changes (*tamo*, *sto*, *chasú-chasa*, *meče-mečém*); to syntactical forms. The final three lines of the passage all repeat the same syntactical units in a kind of parataxis, even if they are slightly rearranged in order. If A = *všech sto jmieše* ("all hundred had"), B = *v nožnách* ("in sheaths," or in other containers), C = *brietné meče* ("sharp swords," or another adjective-noun pair), and D = *k mečém* ("for the sword," or for some other dative object), these lines are then of the form ABC, DAC, DBC. The variation in the line openings returns to strict parallels in line ends, which we expect from the principled right justification of the verse (which is visible and well imitated at this level, if not at the metrical level).

What is notable about several of the other quasi-formulas in the Manuscripts is that they are derived from traditional Russian epithet-noun phrases. Their repetition in the Manuscripts, then, also echoed their repetition in Russian sources, for those that were familiar with them, confirming their traditional status and suggesting again their antiquity from a shared Slavic heroic past. One appears in the address to prince Oldřich by Výhoň Dub early in that epic, urging him on to the fight (ll. 16-22):

„Hoj, poslyš, ty *veleslavný kněže*!	"Hark you now, o glorious prince!
Bóh ti bujarost da u vsě údy,	God gave you vim in all your limbs,
bóh ti da věhlasy v *bujnú hlavu*;	God gave you wit in your *brash head*.
ty ny vedi proti zlým Polanóm!	Lead us now against the evil Polans!
Po tvém slově pójdem v pravo, v levo,	By your word we'll go right or left,
buď v před, buď v zad, u vsě *pótky lúté*.	Forward or backward in all fierce battles.
Vzhóru! Vzmušte chrabrost bujných srdec!"	Arise! Wake valor in vigorous hearts!"

The brash or reckless head is a commonplace in Russian traditional oral epic. A typical example, which also includes a *formula of address* similar to the one here, can be found in the epic *Ilya*

[9] We should recall here Parry's definition of the formula: "a group of words which is regularly employed under the same metrical conditions to express a given essential idea" (Lord 1960:30). In his work on the South Slavic material, Lord analyzes repetitions within or across the two cola (four syllables + six syllables) that make up the line to account for the metrical conditions. Because the amount of material is so small for the ten-syllable Czech epic imitations, we take any repetition of two or more meaningful words as a quasi-formula.

[10] Repetition of prepositions is rare in the South Slavic *deseterac* form but more common in the longer Bugarštica form and definitely common in the longer Russian *bylina* line, so this also has Slavic oral-formulaic epic antecedents. The Old Czech text of the Manuscripts in Cooper 2018 is taken from Dobiáš 2010, the most recent scholarly edition.

Muromets and Kalin Tsar recorded by Gilferding from the singer T. G. Riabinin: "Ай же ты, Владимир-князь да стольнокиевский! / Не сруби-тко мне да буйной головы" ("Hail to you, Vladimir, Prince of Capital Kiev! / Don't cut off my reckless head!") (Илья Муромец и Калин-Царь n.d.).[11] This adjective-noun phrase was not, however, traditional in Old Czech poetry, although the adjective existed both in the sense of "vigorous" and in the sense of "unruly." It is embedded here, though, in a speech that abounds in traditional and quasi-formulaic phraseology. It ends a pair of lines formed by parallel syntax (*bóh ti da* X *v*/*u* Y) that also feature significant sound repetition (*bujarost—bujnú hlavu*). In this case, we also have the repetition of the adjective in a compound neologism, combining the sense of *bujný* with *jarý* ("vigorous" or "fervent"), which is suggestive of Greek epic compounds. The adjective repeats again in the final line in a combination that does not recur in the Manuscripts. The repetitions in this passage already begin to establish the would-be traditional, quasi-formulaic nature of the epithet even before it repeats in the epic *Jaroslav*. Other quasi-formulas appear in the passage as well: *pótky lúté* ("fierce battles"), which repeats in *Jaroslav* (l. 277), and the phrase *veleslavný kněže* ("glorious prince"), which is echoed in the epic in eight-syllable lines, *Ludiše and Lubor* (l. 4). Finally, the everyday formulaic language of right or left, forward or backward, rounds out the traditional and quasi-traditional phraseology of this passage. The skillful forgers, as we see here, were able to enrich the forms of South Slavic oral epic with traditional language and features from other prestigious or related traditions (Russian and Greek) in their Czech imitation.

The analysis of these two short passages suggests how qualitatively well the Czech imitators were able to simulate traditional oral epic phraseology in its many-layered features. Not all passages yield as well to such analysis, though, and the number of repeated, quasi-formulaic phrases seems small even relative to the tiny size of the Manuscripts' corpus (the epics entirely in ten-syllable lines combined with passages in ten-syllable form from the two others amount to a total of 576 lines, just under 3,100 words): for example, just twenty-five epithet-noun combinations repeat within or across songs in that corpus, to which we can add only thirteen more if we allow confirmation of combinations in those lines from other parts of the Manuscripts (as is the case for the phrase "glorious prince" noted above). How might that compare quantitatively to the repetitions in the South Slavic or the highly repetitive Russian *bylina* traditions? Is there a quantitative measure that could be used to characterize oral-formulaic verse, or does the multiformity of different formulaic systems preclude easy quantitative characterization and necessitate a tradition-dependent analysis? In our case, is there a measure that would give an indication of how close the known imitation is to its models?

Formulaic Density as a Measure and Associated Digital Methods

The notion of "formulaic density" has been used, and critiqued, as a measure that can characterize oral-formulaic verse and distinguish it from other types. The method derived from Parry's and Lord's demonstrations, on sample passages from the *Iliad*, the *Odyssey*, and the *Song*

[11] Russian text from the website Русские Былины (https://www.byliny.ru/content/text/ilya-muromets-i-kalin, accessed September 9, 2019). Translation from Bailey and Ivanova 1998:68. Flajšhans suggests it comes from a commonplace in the *bylina* collection by Kirsha Danilov (Vojtěch and Flajšhans 1930:7).

of Bagdad from the Parry collection, of the pervasiveness of formulaic phrasing in texts, the formulaic nature of which had been well established (Bynum 1978:6-7; Lord 1960:45ff.). It involved underlining with a solid line those phrases (or lines and half-lines) that repeated exactly in the corpus of the singer (Homer or Salih Ugljanin) and with a dotted line those phrases that belonged to a formulaic pattern, repeating rhythm and syntax and at least one word, and calculating the percent of the sample that, by evidence of repetition, was formulaic (solid lines) or belonged to formulaic systems (dotted). The method was used, though, as a kind of measure to try to establish the oral origins of texts whose oral provenance had not been established, not only in studies of Old English literature (by Francis P. Magoun and others) but also by Lord and his students on a variety of medieval texts from different traditions (Bynum 1978:8-11; Lord 1986:479). The results were highly disputed, not only regarding the methods of conducting the calculations but also concerning whether such a test was at all diagnostic of the oral origins of the texts.

David E. Bynum, in his critique of this practice, objects to the mechanical application of the method as a test to texts that have not had their formulaic nature established, and in particular to the highly problematic definition of formulas in such tests as any phrase that repeats within the same text or elsewhere, noting that Parry himself cautioned against such a reduction (1978:6). After all, non-formulaic poetic texts also make use of various kinds of repetitions for thematic and emphatic purposes, and repetition in formulaic texts can also be of those varieties. Lord himself retrospectively concurred with this criticism (1986:491-93). Foley reviewed the problems with the use of "formulaic density" as a test as part of his argument for establishing the principles of genre-dependence and tradition-dependence for analysis (quantitative or otherwise) of the traditional, formulaic nature of any texts. That is, formulaic composition belongs in particular only to certain genres of oral poetry, not all, and how that formulaic nature is realized differs from one language tradition to another (including in the aspect of how pervasive formulaic repetition is) (Foley 1990:3-4). When Foley then replicates the method of analysis by underlining and calculating percentages of repetition of two sample texts in his volume (167-70), he insists "that the present study aims not at a determination of orality via formulaic density (if indeed this were possible) or other quantitative goal, but rather at a general demonstration of formulaic structure and the more fundamental role of traditional rules" (129 n. 9), perhaps thus returning the method to its original purpose. In the process, though, he also aims to show, through his in-depth analysis of the phraseology, how Lord's treatment of formula impoverishes our understanding of it by glossing over "the natural heterogeneity of traditional phraseology" by testing only for identical units (172-73). Foley's extended discussion of the formulas and systems disclosed by this analysis demonstrates that repeated phrases often are not simple formulas in themselves but can belong to more than one formulaic system of different lengths, varying from

a half line to multiple lines, and of different sorts.[12] To these fundamental criticisms of the method one might also add the glaring problem of the use of a representative short sample text to conduct the calculation, where any other fourteen- to twenty-five-line passage from the same poem might yield quite divergent results.

Digital methods offered the possibility of improvements in calculations of formulaic density, greatly facilitating the search for repeated words and phrases (even through simple searching of digital text or computer concordances) through corpora of unwieldy length for manual searching (such as all of Homeric epic or 12,000 lines of South Slavic epic). But the major problems with the method, like inattention to genre-dependence and the misleading definition of formulas as verbatim repetitions, often remained. Vikis-Freibergs and Freibergs addressed the "sampling problem" with their method for calculating the formulaic density of an entire corpus of texts (1978:333). They also addressed the variety of different kinds of formulas, defining three types ("syntagmatic"—any verbatim repetition; "paradigmatic"—repetition with possible substitution; and "syntactic"—minimal verbal repetition within a repeated syntactical shape (331-32)), but given the inadequate language processing possibilities at that time (1978), limited themselves to only the first, exact repetition (and that without regard to its metrical shape or place in the poetic line).[13] Furthermore, they chose as the corpus for their calculation a set of short lyric songs (Lithuanian sun-songs), which are of undoubted folk origin but not at all the sort of song that the oral-formulaic theory was formulated to account for: longer epic or narrative songs recomposed in performance in a manner facilitated by a highly structured, traditional phraseology. They conclude, very problematically, that "formulaic structure is typical of oral literature" in general, and that "Furthermore, this characteristic seems independent of the genre of literature in question, since our short, lyrical songs seem to be as formulaic as the long narrative epics analyzed earlier" (338). They certainly found repetitions in their corpus, but are these really formulas at all, or the kind of repetition and variation one might find in any set of folksong variants on a narrow topic?

Nikolayev applies his new algorithm to a corpus of texts that have been independently shown to be formulaic, Russian *byliny*, but follows Vikis-Freibergs and Freibergs in his definition of the formula, for the purposes of formula-searching and calculation of formulaic density, as verbatim repetition (Nikolayev 2016:112). Where Vikis-Freibergs and Freibergs

---

[12] Foley's analysis, even of the first line of his first sample passage, demonstrates this sufficiently. The line, *Kad je Pero knjigu načinijo* ("When Pero formulated the letter"), has solid underlines for the first three words (first colon formula) and for the next two words (second colon formula), as well as a solid line under the first four words and dotted under the final word (whole-line formulaic system with possible substitution of the final verb). But, the first colon formula also belongs to a substitution system where any other two-syllable name or noun can be inserted for Pero; and the second colon formula also exists as a substitution system with different verbs following different line openings. Moreover, the first colon substitution system ("When X") is of an entirely different type than the second colon one ("composed/formulated/decorated the letter"), which has limited substitution of the verb, because it hardly contains the "essential idea" of a formula at all, but sets up, with a tremendous variety of two-syllable subject words, a limitless number of possibilities for finishing the line. Foley further demonstrates how the line is a product of traditional rules at a more fundamental level than formulaic repetition (1990:178-80). The method of "formulaic density," which considers primarily exact repetitions, reduces this complexity and simply calculates this line as 100% formulaic.

[13] We are perhaps only now becoming potentially capable of addressing the complicated issue of syntactic repetition with the kind of natural language processing possible for some languages.

supply different density numbers for differently sized units of repetition (word pairs, word triplets, lines, and couplets), which makes visible the multiform aspect of the formula, Nikolayev prefers to compute a "unified formulaic density" for a text or corpus (112), once again making different lengths of formulas invisible. The problem of treating any repetition as a formula is further magnified in Nikolayev's approach when he calculates what he calls the "internal formulaic densities" of each poetic text, that is, the portion of the text that repeats within the same poem, and averages across the corpus of poems (124). Repetitions within poems can be for emphasis, for delineation of theme, or for any of a number of poetic purposes outside of formulaic composition, while repetitions across poems (involving different themes and characters) are more likely to be formulaic.[14] Nikolayev does, in the end, calculate a formulaic density across the entire corpus, which he finds to be 56% for his large corpus of *byliny* (125).

Nikolayev expresses surprise at the lack of follow-up on Vikis-Freibergs and Freibergs' calculations of formulaic density for a whole corpus of texts (111), but it seems likely that this is attributable to the fundamental criticisms of formulaic density that emptied the notion of its diagnostic power and analytical utility. If formulaic density cannot be used to determine if a text is composed using the phraseology of an oral-formulaic tradition or not, then what does that 56% figure for Russian *byliny* tells us that is of analytical importance? One might compare the closely related Russian "historical songs" and ballads, whose phraseology suggests they are also oral-formulaic, and see if they differ in density;[15] or one might compare, across languages and traditions, to the much lower densities calculated for Old English heroic verse. In both cases, though, any understanding of differences in figures would necessitate a deeper understanding of the different formulaic systems in their multiformity and how they are deployed, a task beyond computer analysis. Could one also compare to literary epic poetry in Russian or to the great stock of Russian lyric poetry? One could apply the same algorithm to such a corpus, but what would the figure produced mean? Could lyric poetry be even 5% formulaic? Since those texts are not traditional or composed in an oral-formulaic manner, strictly speaking, their density of formula would have to be zero. Here one would have to give up the fiction that one is calculating formulaic density and acknowledge that one is calculating density of phrasal repetition in a particular manner.

A Different Approach

Why not start instead with a method of calculation that could be applied to any type of text and that might be diagnostic when comparing texts derived from traditional oral-formulaic phraseology to those composed in other manners? That is what we did in our attempt to compare

[14] Lord (1960:45-46) avows that he chose his sample passage as one that avoided commonly recurring themes, like letter-writing, that occur even across poems, and that he did not include passages from other versions of the same song by the singer, in order to make the formula calculation more valid.

[15] Bailey and Ivanova (1998) include two historical songs and one ballad in their translation of Russian folk epics, focused on the *byliny*.

the pseudo-folk epics of the Czech Manuscripts to their formulaic models and to other Czech literary and folk texts.

The measures we selected to test all proceed from calculations based on the repetition of words or pairs of words—in this, they resemble measures of formulaic density. But they are not based on any definition of the formula and can be applied to any kind of text.[16] In a sense, we were testing for a symptom of formulaic composition, which results, we hypothesized, in increased textual repetitiveness, rather than trying to define a meaningful, quantifiable measure of the special forms of repetitiveness that characterize the formula. One of the most basic measures of the diversity or repetitiveness of a text's vocabulary, included in our tests, is *type-token ratio* (TTR), which is the ratio of the total number of unique words in a text (types) to its total number of words (tokens). Information theory offers, in measures of *entropy*, a number of more subtle measures of the relative uncertainty or, from the opposite perspective, predictability of texts based on their repetitions. These calculations are based not on a binary result (the word repeats or does not repeat) but on the word's probability within the text (the number of occurrences divided by the total number of words in the text). They thus employ a more nuanced quantification of repetitions. One can calculate the entropy for the words of a text (*unigram entropy*) or for successive pairs of words (*bigram entropy*). (See the section on methodology, below, for definitions of these measures and details on our implementation.) The measure of *conditional entropy* also initially looked promising for characterizing oral-formulaic poetry. This can be used to calculate the average uncertainty for the second word, given the first word, in successive pairs of words for a text. It is, in a certain sense, a measure of the predictability of the continuation of a phrase, given a word in the phrase. Intuitively, this looked like a good measure for a formulaic text, because someone familiar with the formulaic phraseology of a tradition can readily predict, given a word from a common formulaic phrase, the following word or words, even to the fairly probable completion of whole poetic lines.[17] Given that, for all these measures, higher probabilities of repetition translate into lower total entropy, we hypothesized that the unigram entropy, bigram entropy, and conditional entropy of word pairs should be relatively low (low uncertainty) on average for oral-formulaic texts, in comparison to other genres of texts. The repetitive vocabulary of the formulaic texts should also result in lower TTR than for other texts.

Methodology

In information theory, entropy is a basic measure of information that is defined by the probability space of possible events and quantifies the uncertainty or amount of information that

---

[16] The authors warmly thank our colleague Ted Underwood for an early consultation on methods, in which he shared possibilities he saw, including calculations of entropy, based on his broad expertise and experience in digital humanities, allowing us to get beyond the notion of formulaic density. We also would like to thank the National Center for Supercomputing Applications at the University of Illinois for funding for the project through a Faculty Fellowship for David Cooper and Michal Ondrejcek.

[17] We casually tested this with a Russian-speaking colleague familiar with *byliny*, feeding her parts of phrases and adjectives that she readily completed with likely phrase continuations and formulaic adjective-noun combinations.

space represents. The Shannon entropy for a sample space X (in the case of this study, for example, some text in Czech) and probability space P where $p(x_i)$ is the probability of outcome x_i (the occurrence of a particular word or letter in the text), is given by Lubbe (1997:8):

$$H(\mathrm{X}) = -\sum_{x=1}^{n} p(x_i)\log_2 p(x_i)$$

This can be used to determine the entropy of the words (unigram entropy) or pairs of successive words (bigram entropy) in a text. For unigram entropy, for example, if a text repeats the same word throughout, there is no uncertainty, and the entropy would be zero; if every word is unique, the uncertainty and amount of information in the text would be maximum for a text of that length. Another measure, conditional entropy, calculates the remaining uncertainty for one variable if the outcome of another variable is known. This can be used for pairs of words, based on the probabilities of the two words occurring together and individually, and allows us to calculate the uncertainty across the text for the second word, given the first word, in successive pairs of words for a text. The conditional entropy for Y given X for sample spaces X and Y and probability space P where $p(x_i,y_j)$ is the joint probability of outcomes x_i and y_j is defined as (Lubbe 1997:18):

$$H(\mathrm{Y}\,|\,\mathrm{X}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i, y_j\right)\log_2 \frac{p\left(x_i, y_j\right)}{p\left(x_i\right)}$$

Our Python code calculates TTR, unigram entropy, bigram entropy, and conditional entropy for texts.[18] A significant challenge for our study, as for any study involving these measures, was normalization for text size. All of these measures change with the size of the text: entropy increases as the text gets larger (more information), while TTR tends to decrease with increasingly larger texts. The discrepancy between the size of our test text (the ten-syllable portions of the Czech Manuscripts, just under 3,100 words) and those texts and corpora of texts we wanted to compare it with (collections of *byliny*, South Slavic oral epics, and corpora of lyric poetry, epic poetry, and prose in all three languages, all tens to hundreds of thousands of words in size) was significant, making whole-text calculations, the norm for these measures, fundamentally impossible to compare. We initially adopted a known solution for this by breaking our larger texts into chunks of text the size of the small test text (3,093 words) and calculating average TTR, unigram entropy, bigram entropy, and conditional entropy across the chunks, with standard deviation. We report computed average values for chunks, including an estimate of the level of confidence reflected in the uncertainty bars (+/- one standard deviation). These results, based on precisely the same text length, are then directly comparable. For our final data set, we ran the calculations on a large set of texts, beginning with a small chunk size and increasing the chunk size with each iteration until the texts, one by one, fell out of the calculation as the chunk size exceeded their whole text size. Each text could be compared to any other that way at several identical chunk sizes.

[18] Our final code and some test texts are available at https://github.com/ncsa-mo/oral-formulaic-poetry.

One of the methods used in natural language processing to reduce the "noise" of textual data is the removal of stop words. Currently, most researchers use manually curated stop word lists, which cannot be easily generalized across languages or corpora. Nikolayev, for example, uses a "stop-list of the most frequent function words," such as ай, без, бы, в, and вам, in his analysis of Russian *byliny* (2016:113). The removal of function words is certainly an aid in the retrieval of formulas; however, the task of characterizing the whole text using unified measures, as we attempt to do, must take function words into account. The strict *deseterac* of the South Slavic oral-epic tradition, which influenced the work of the Czech forgers, favors the liberal use of conjunctions, especially at the beginning of line-initial formulas (Lord 1960:41-42). Similarly, the use of ай or да at the beginning of verse lines in Russian *byliny*, which are less strict metrically than the South Slavic epics, is an important component of rhythm and repetition. As an illustration, the 601-line *bylina* from the Putilov collection "Садко" ("Sadko") contains 353 lines, or almost 59%, that begin with Ай. Such a high frequency of function words will have a substantial impact on both TTR and statistical calculations. Not only are stop words, then, an important component of the structure of the whole text, but considering that our aim is to determine the usefulness of informatics and statistical measures for the characterization of the whole text, not limited to the often variable definitions of formulas employed in other quantitative approaches, the most common, repeating, small words that combine with a myriad of other words are an essential part of the probabilistic space of any text for the calculations.[19]

Text Compilation and Preprocessing

Building our corpus of comparative texts presented challenges, not only in the finding or creation of electronic texts but also in their preparation for use in the calculation. Given the generally agreed upon principle that formulas maintain their identity across grammatical and morphological variation, and in general to capture the repetition of words, for these highly inflected Slavic languages, it was important to lemmatize the texts. For all three languages, we were able to find reliable lemmatizers we could use to produce accurate strings of lemmas from

[19] Our calculations are not aimed at all, unlike those of Vikis-Freibergs and Freibergs or Nikolayev, at finding formulas in the text, but at a global measure of the text's lexical variability and word-by-word uncertainty. An entirely separate aspect of our project did involve searching for formulas, for which we found the collocation tools in the Natural Language Toolkit (NLTK) to be a flexible and effective proxy, as formulas share some characteristics with natural language collocations. The method of "fuzzy" phrase matching offered by Broadwell, Leonard, and Tangherlini (2017) also looks promising for finding the kind of partial repetitions that define formulaic systems, though we did not attempt it.

the texts.[20] Some texts, though, presented special problems. Nikolayev (2016:113-14) noted the issue of the dialects in Russian *byliny*. It was necessary to edit many unusual and frequently occurring morphological features (adjective and verb endings, in particular) to standard forms—this reduced the uncertainties in the results to an acceptably low level.[21] The South Slavic epics also had some orthographic features (especially the use of abbreviations) that troubled the lemmatizer, necessitating some pre-editing of the text.[22] The pseudo-Old-Czech text of the Manuscripts presented unique problems, as the verb system, in particular, has changed radically in modern Czech, so the lemmatizer could not be used. Author Cooper made use of a concordance to the texts of the Manuscripts to lemmatize manually (Flajšhans 1897).

For our primary set of texts for comparison to the Czech Manuscripts, we wanted to include the texts that most likely served as models for Hanka and his collaborators, which meant the *deseterac* epics published by Karadžić in 1814 and 1815, as well as Russian *byliny* from the Kirsha Danilov collection, published in 1804.[23] For more representative samples of these oral-formulaic traditions, we also compiled larger selections of *byliny* and South Slavic epics from

---

[20] A 2018 study of four Russian morphological parsers showed that our chosen lemmatizer—PyMorphy2 (https://github.com/kmike/pymorphy2)—had a consistent accuracy of lemmatization above 93%, reaching 98.29% for OpenCorpora (Kotelnikov, Razova, and Fishcheva 2018). It was additionally chosen because of its permissive licensing under MIT License and ease of integration with our code. Our Czech lemmatizer, MorphoDiTa, Morphological Dictionary and Tagger (https://ufal.mff.cuni.cz/morphodita), uses the MorfFlex CZ 161115 language model as dictionary and tagger. It is discussed in Straková, Straka, and Hajič 2014. For South Slavic texts we used ReLDIanno, a text annotation service which is part of ReLDI, a SNSF-funded SCOPES project that enables processing of three South Slavic languages using srLex and hrLex inflectional lexicons of Serbian and Croatian languages. The implementation pipeline for ReLDI Tagger and Lemmatiser, a Slovene, Croatian, and Serbian lemmatizer with reported accuracy above 94% for both standard and non-standard texts, is described in Ljubešić et al. 2016. We are aware that there are challenges with accuracy metrics.

[21] We did not alter unusual or phonetic spellings (акиян vs. океан), as we found our lemmatizer treated such forms consistently across occurrences. We tested Nikolayev's practice of using 4-grams as a substitute for lemmatized text, running a TTR calculation on a variety of texts in all three languages. Where he found that the practice exaggerated formulaic density calculations by just 1.5% (2016:114), we found a much larger exaggeration for TTR. We calculated TTR for the original text, 4-grams, and lemmatized text and compared how much the 4-grams and lemmatization reduced the number of unique words for the text. (Both methods address the difference in morphological endings for the same word in these inflected Slavic languages, enabling recognition of identical words by a short, 4-letter stem or assignment of the word to its lemma.) Overall, across a variety of texts in all three languages, the 4-gram method reduced the number of unique words by 40% *more* than lemmatization. For Russian texts it overreduced the number of unique words by 49% and for Serbian texts by 47%. The percentage for Czech texts was smaller, just 22%, because, for the Old Czech text of the Manuscripts, the 4-gram method actually resulted in a 17% smaller reduction in the number of unique words than lemmatization. For the modern Czech texts, the 4-gram method overreduced unique words by 30%. For our calculations, then, the 4-gram method results in an unacceptable level of error, reducing far too many different words to identity in their 4-letter stems.

[22] The most common orthographic variation, even within a single corpus or text, was the variable use of *al'* or *ali.* All the abbreviated forms were changed to *ali.* Similarly, *da l'* was often used at the beginning of interrogative sentences or *l'* after the main verb, which had to be normalized to *da li* and *li*, respectively.

[23] We took digital versions of the Karadžić texts from the website Викизворник (https://sr.wikisource.org/wiki/Додатак:Епске_народне_песме, accessed November 18, 2019). These included the two epics Hanka translated, four from Karadžić's 1814 volume, and all the epic songs from the 1815 volume. From the Kirsha Danilov volume, we selected thirty-four texts that were entirely in verse form from the website of the Fundamental Digital Library of Russian Literature and Folklore (http://feb-web.ru/feb/byliny/default.asp?/feb/byliny/texts/kir/kir.html, accessed August 5, 2021).

both Christian and Muslim singers.[24] For literary texts, our primary comparison would be to corpora we selected of literary epic poetry and lyric poetry in each language, mostly published in the last two decades of the eighteenth century and in the first third of the nineteenth century, that is, in the period leading up to and following the discovery of the manuscripts.[25] We also added collections of folk ballads in each language in order to include authentic Czech narrative folk poetry (which uses rhyme in a relatively short form and thus is likely a memorized, not oral-formulaic, genre) as a point of comparison, as well.[26]

Results

Our initial results, calculated while we were still compiling and preprocessing many of our key sets of texts for comparison, showed that several measures were promising, with one surprising exception. We did not expect to see, as we did (Figure 1), that when translated into each of the three languages, Milton's *Paradise Lost*, a literary epic, had a *lower* conditional entropy than the oral-formulaic texts and the Manuscripts.[27] How could such a literary epic be less entropic, word-by-word, than oral-formulaic epic texts? The answer was that the kind of predictability oral-formulaic texts have in relation to other texts is not the same kind of predictability that the conditional entropy calculation was measuring. Because the calculation does not take all the possibilities presented by the language as potential options for the second word, once we know the first word, but only those possibilities *that are realized within the text or corpus used in the calculation*, a text with less repetition and more unique combinations of

[24] The expanded *bylina* corpus was taken from the volume edited by Putilov, available online in the Fundamental Digital Library of Russian Literature and Folklore (http://feb-web.ru/feb/byliny/default.asp?/feb/byliny/texts/bpu/bpu.html, accessed August 12, 2021). The Muslim traditional epics were taken from Marjanović 1898—the third volume of a Matica Hrvatska publication, focused specifically on Muslim singers. The Karadžić texts from the 1814 and 1815 publication were expanded with the texts from Karadžić 1987.

[25] For Czech poetry, we were fortunate to be given access to the excellent "Corpus of Czech Verse" (Plecháč et al. n.d.) (many thanks to Petr Plecháč). The corpus includes nearly every volume of poetry published in the late eighteenth and nineteenth centuries, and texts are lemmatized. For Russian poetry, we selected texts from major and minor authors on the website Библиотека Максима Мошкова (http://lib.ru/, accessed July 27, 2021). For Serbian poetry, we selected nineteenth-century texts from the websites Антологија српске књижевности (http://www.antologijasrpskeknjizevnosti.rs/, accessed September 7, 2021) and Пројекат Растко (https://rastko.rs/, accessed September 7, 2021), as well as Popović 1968.

[26] Czech texts were from Erben's folksongs volume (1886). Russian ballads were from the electronic edition of Kirdan 2001, available at https://www.booksite.ru/fulltext/bal/lad/yry/index.htm (accessed July 27, 2021). Serbian narrative folk poetry texts were taken from Krstanović 1990, with additions from the websites Антологија српске књижевности and Пројекат Растко (see previous note).

[27] In Figure 1, all texts are chunked to the size of the ten-syllable parts of the Czech Manuscripts (RKZ); the graph gives the average conditional entropy across chunks, with uncertainty bars for one standard deviation. The value for the manuscripts is not an average, but calculated for the entire text, thus the lack of uncertainty bars for that value. Translations of Cervantes' *Don Quixote* into each language are included here as well.

words has much higher probabilities for those following words, and thus lower entropy.[28] What conditional entropy measured did not conform to our initial, intuitive understanding of it (beware, digital humanities scholars!). Conditional entropy was *not* a good measure, then, as we continued to see, for characterizing the oral-formulaic texts in contrast to other related kinds of texts.

Fortunately, some of the other measures we were also calculating looked (and eventually proved) much more promising. The evident lexical repetitiveness of the oral-formulaic texts shows up quite clearly in the most basic measure of lexical diversity, TTR. In these traditions, oral-formulaic texts make use of a relatively narrow, traditional, and even ritualized vocabulary. The somewhat more subtle but related informatics measure of unigram (word) entropy (which also accounts for frequency of repetition in the higher probabilities of common words) also proved useful for distinguishing the oral-formulaic texts from some similar types. Bigram entropy frequently demonstrated similar distinctions, but at a much smaller relative scale of differentiation.

The Czech Manuscripts' proximity in TTR to South Slavic and Russian oral-formulaic texts is shown in Figure 2, which includes results for just the ten-syllable parts of the Manuscripts, as well as for the Manuscripts in their entirety. At small chunk sizes, all the measurements for average TTR have large standard deviation uncertainty bars, but as we approach the full size of the texts or corpora, we get an accurate measure, and the final point for each curve represents a whole text/corpus measurement without any averaging across chunks.[29] (We are showing the deviation here only at three chunk sizes, for clarity.) The ten-syllable epic parts of the Manuscripts have the highest TTR figures but remain quite close to their likely

[28] A little thought experiment made this clear to us. Imagine an alternate world in which a contemporary oral-formulaic epic poetry exists in English, and that this heroic tradition is, for whatever reason, formulaically "happy." In a 2,000-word text from that epic poetry, we find the following formulaic adjective-noun combinations: "happy day" (three times), "happy horse" (two times), "happy home" (four times), "happy slaughter" (two times), and "happy wife" (sadly, just once). In real life, though, this culture is as unhappy as any other, so in a 2,000-word prose text, we find only these combinations: "happy day" (two times) and "happy birthday" (once). In the formulaic text, the probability of the word "happy" is 12/2,000 (it occurs twelve times); in the prose text, just 3/2,000. The conditional probability of "day" given "happy" in the formulaic text is just 3/12 (three out of twelve times, it follows "happy") or 1/4; but in the prose text, its conditional probability is a whopping 2/3 (it occurs two out of three times after "happy"). Higher probabilities mean lower entropies, and a quick calculation shows that the bigram "happy day" contributes five times more to the entropy sum in the formulaic text than it does in the prose text. The prose text is, indeed, by this manner of calculation far more predictable and thus lower in entropy.

[29] What do the uncertainty bars here represent? We are using statistical measures for whole texts but breaking our texts and corpora into chunks of different size in order to be able to compare measurements for texts of very different sizes. The data points, then, are averages across the chunks, which, depending on chunk size, could include less than any full single text; a handful of lyric poems by a single author or approximately a whole epic poem; lyric and epic poems by several authors or singers; up to a representative collection of literary or oral epic and lyric poems in a language from the period. The uncertainty bars thus show the variation in the statistical measure for different texts within the genre at a given sampling size.

models: the epics published by Karadžić and those of Danilov.[30] This closeness to their models translates into a distinct distance from contemporaneous Czech literary epic and lyric poetry, as is shown in Figure 3.[31] The Manuscripts do not resemble, by this measure or by unigram and bigram entropy (Figures 4 and 5), the poetry that was being written and published in Czech in the decades surrounding their discovery (having a much less diverse vocabulary), but are closer to their models.[32] Figure 3 also includes a collection of native Czech narrative folk poetry, from Karel Jaromír Erben's collection of folksongs (1886). These are not oral-formulaic narrative songs, but for these measures, they fall neatly into the range of the Russian and South Slavic oral-formulaic traditions. The Manuscripts fall in between native folk traditions, then, and literary traditions, but are much closer to the folk traditions they imitate. In fact, the Manuscripts' imitations very much resemble, by these measures, the much revered and canonical imitations of Russian folk poetry by Czech writer František Ladislav Čelakovský, as shown in Figure 6, a resemblance that belies the very different treatment of the Manuscripts' authors (as reviled forgers).

A distinct difference between oral poetic texts and literary poetic texts can also be seen in the Russian tradition, as Figure 7 shows. The distinction is not as pronounced in the case of Serbian literary texts (Figure 8), but this is not surprising when one remembers how important Karadžić's publications and folk traditions, in general, were for the formation of Serbian national literature in the nineteenth century. Many of the Serbian literary texts imitate or incorporate aspects of folk traditions. Foley's principle of tradition-dependence is worth keeping in mind here for the literary texts as well as the oral texts. These graphs also include results for ballads or narrative folksong texts, which, as in the Czech case, fall into the same range as the oral-formulaic texts (the Russian ballads may themselves be formulaic) and cannot be distinguished from them by these measures. Figure 9 shows the relationships as measured by TTR between all these texts together, with languages represented by shapes and text genres by color: oral texts in green, literary epic in blue, literary lyric in red, and imitations of oral texts, including the Manuscripts, in black. Even across languages, the oral poetic texts here are distinct from the literary ones, and the imitations fall in between. As for conditional entropy, Figure 10 shows how this more complex measure gave ambiguous results, with Russian literary texts showing lower

[30] Does the highly repetitive vocabulary of the Manuscripts resemble that of its models? Among the fifty most common words in the Manuscripts, Karadžić's epics, and the Danilov *byliny*, respectively, are the following related terms: *meč*, *sablja*, — ("sword"/"saber"); *kněz*, —, *kniaz* ("prince"); *slovo*, *slovo*, *slovo* ("word"); Tatařín, Turčin, — ("Tartar," "Turk"); *hrad*, *grad*, *gorod* ("fortress"/"city"); —, *car*, *tsar'* ("czar"); and —, *konj*, *kon'* ("horse"). The absence of "czar" from the Czech text is due to the different form of Czech kingship, but the absence of "horse" reflects an interesting reduction in the formulaic language and themes concerning horses from the South Slavic and Russian models.

[31] For our study, the first important chunk size is 3,000 words, approximately the full size of the ten-syllable portions of the Manuscripts. Critically, there is little overlap between literary epic and lyric genres and oral genres, even if we were to double the uncertainty bars to two standard deviations, at chunk sizes that begin to represent meaningful measures of genres. Only the most marginal phenomena of the literary and oral genres overlap. The Manuscripts position themselves in the vicinity of their oral models and at a significant distance from the literary genres of their day.

[32] In our results, unigram entropy shows very similar relationships between texts and genres to that shown by TTR. Bigram entropy also shows similar relationships, but less distinctly—Figures 4 and 5 zoom in a bit closer to a portion of the data to make the differences more visible.

conditional entropy than oral texts at smaller chunk sizes, and higher entropy at sizes representing larger collections of texts.

We can conclude, then, that TTR and unigram entropy are good measures for comparing the Czech Manuscripts' epic songs to the oral-formulaic traditions they imitated, and they show that the Manuscripts clearly distinguished themselves from their contemporary Czech literary-poetic practices, moving far in the direction of the oral texts, but not, in the end, quite as far in the direction of reduced and repetitive or ritualized lexicon.[33] TTR and unigram entropy, however, did not distinguish between formulaic and non-formulaic forms of oral narrative songs in our test corpora. The advantage of these measures from informatics and natural language processing over the notion of formulaic density is that they can be applied to any kind of text. TTR and unigram entropy look like good measures for distinguishing oral-formulaic epic poetry from literary poetry of all types in these language traditions, and thus as measures of texts of unknown provenance or imitations of oral-formulaic epics, to see how well they fit. Whether they will be useful for distinguishing oral-formulaic poetry from other types of texts in other languages and traditions, though, is something that must be tested in each case and on large corpora of texts, given how tradition-dependent the features of oral-formulaic texts can be.

University of Illinois Urbana-Champaign

[33] Recall from the first passage examined above the existence of quasi-formulas in the Manuscripts where synonymous adjectives are applied to the same noun (*břietný*/*ostrý meč* ("sharp sword"), *mocná*/*silna paže* ("powerful/strong arms")). Traditional oral diction would more likely select one formulaic combination, rather than allowing two possibilities that transmit the same meaning under identical metrical conditions (all the adjectives here are two-syllable). This is one way in which the Manuscripts are less formulaically repetitive than their models.

Figures

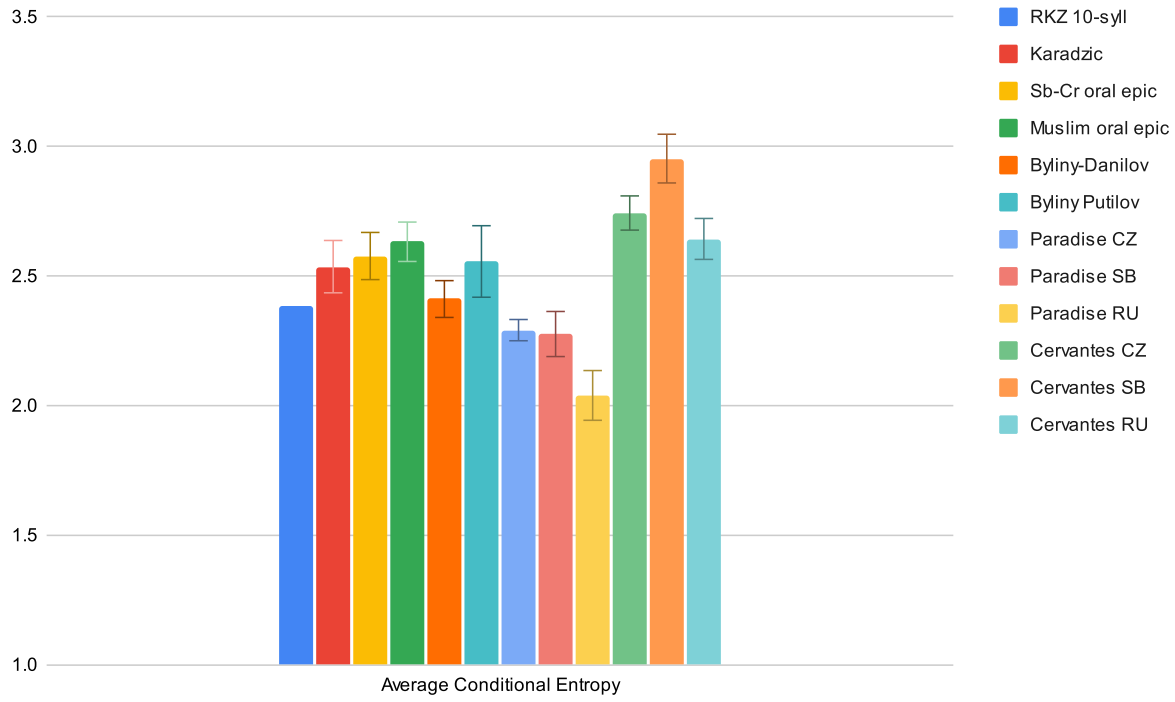

Fig. 1. Average conditional entropy at a chunk size equal to the length of the Czech manuscripts (RKZ) for corpora of oral-formulaic poetry, as well as translations of Milton's *Paradise Lost* and Cervantes' *Don Quixote.*

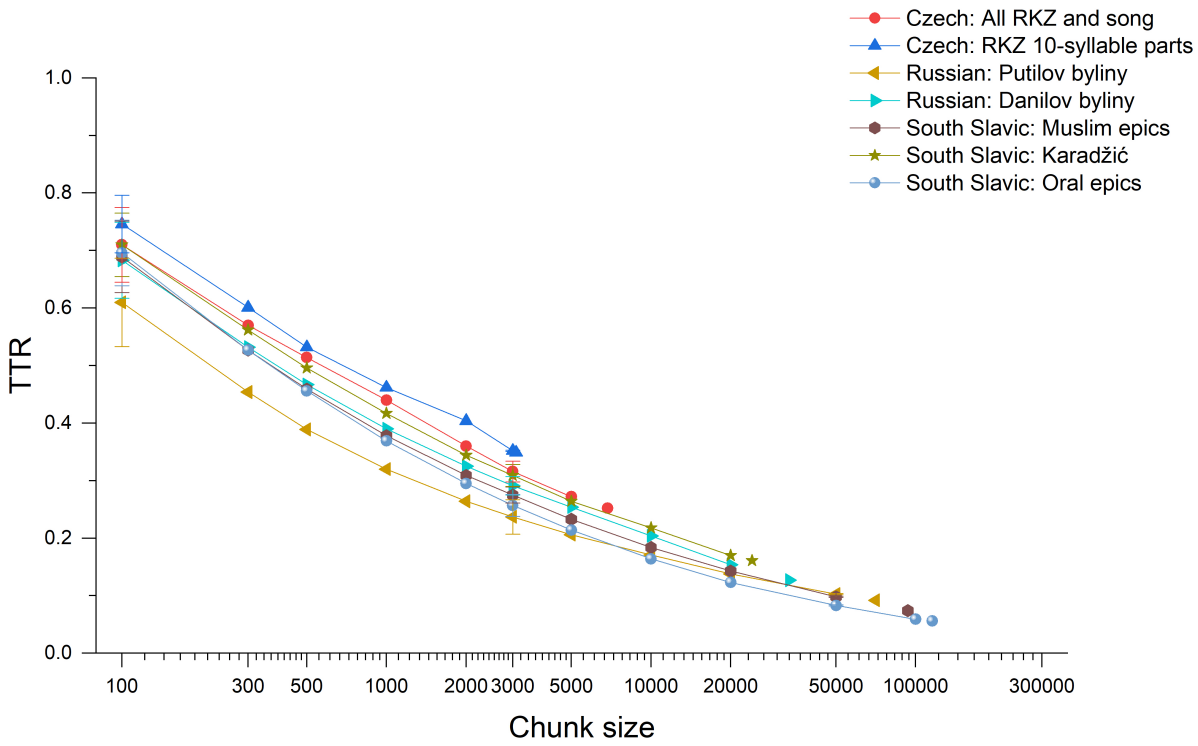

Fig. 2. Type-token ratio across different chunk sizes (on a logarithmic scale) for the Czech manuscripts in comparison with corpora of South Slavic and Russian oral-formulaic epics.

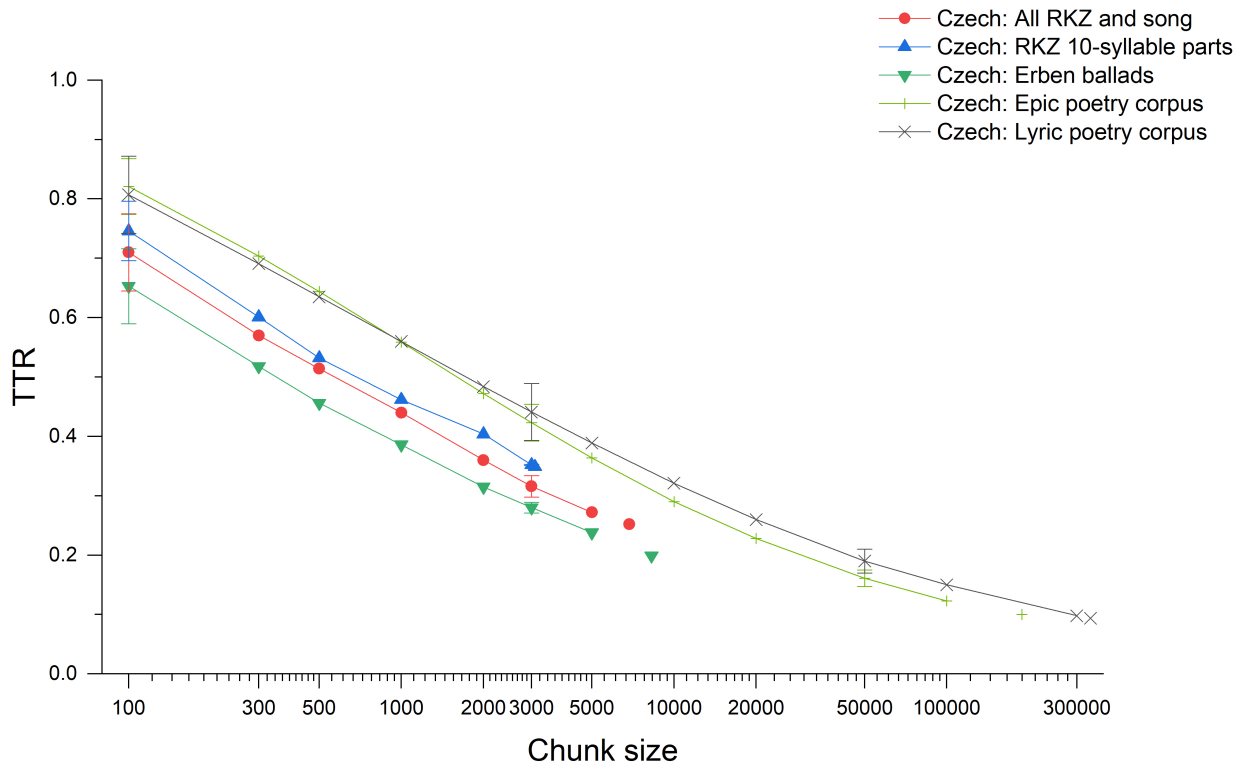

Fig. 3. Type-token ratio across different chunk sizes (on a logarithmic scale) for the Czech manuscripts in comparison with corpora of Czech literary epic and lyric poetry as well as Czech folk ballads.

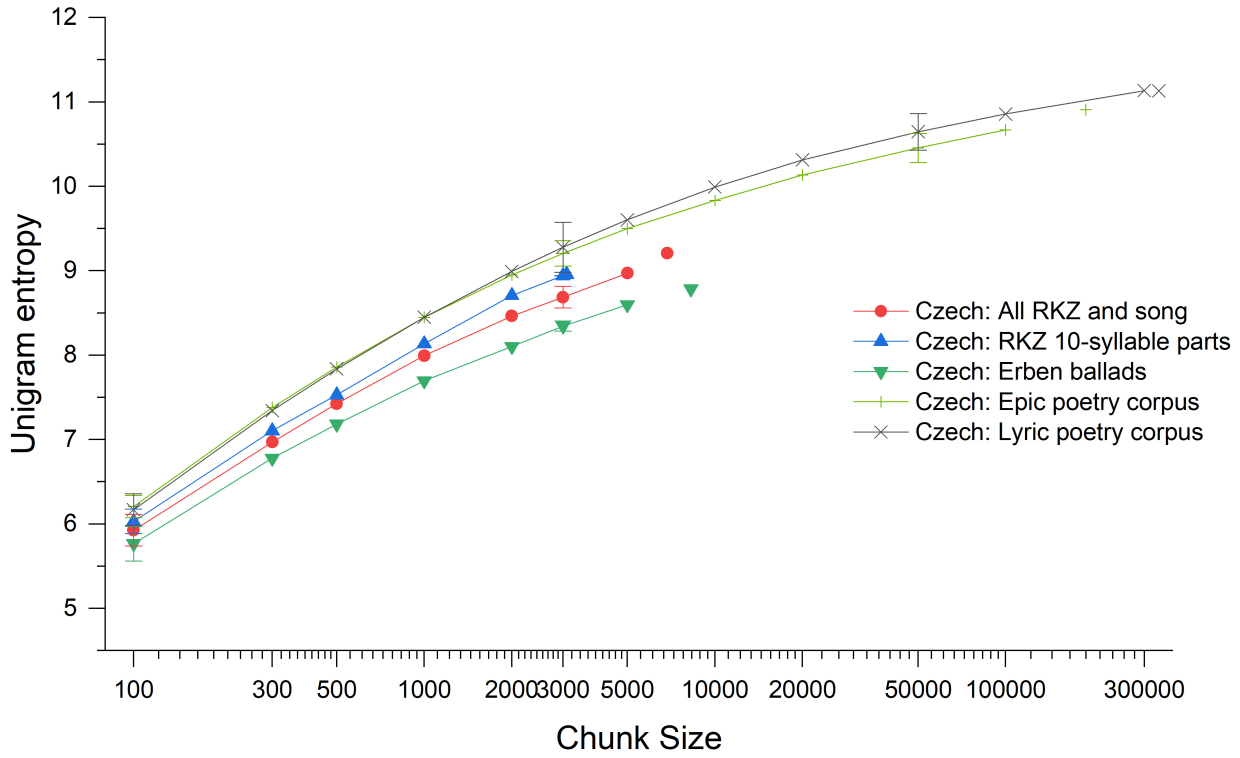

Fig. 4. Unigram entropy across different chunk sizes (on a logarithmic scale) for the Czech manuscripts in comparison with corpora of Czech literary epic and lyric poetry, as well as Czech folk ballads.

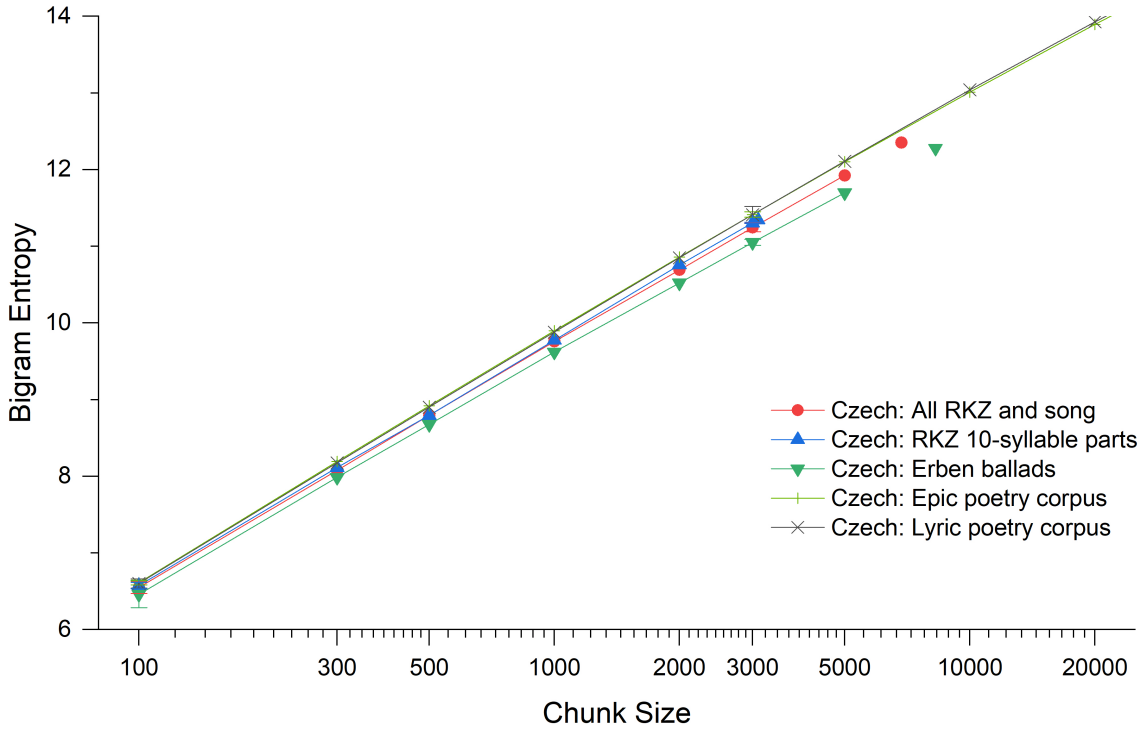

Fig. 5. Bigram entropy across different chunk sizes (on a logarithmic scale) for the Czech manuscripts in comparison with corpora of Czech literary epic and lyric poetry, as well as Czech folk ballads. The scale here is larger than on other graphs to make visible the smaller differences.

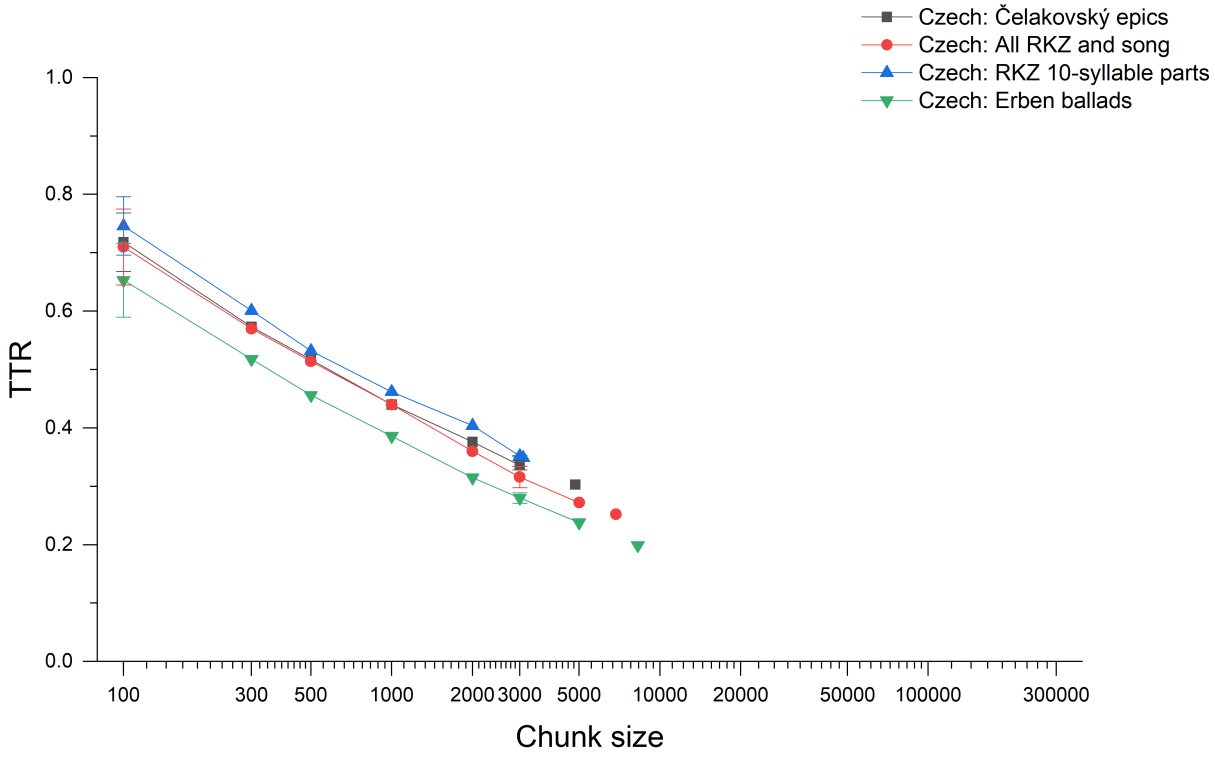

Fig. 6. Type-token ratio across different chunk sizes (on a logarithmic scale) for the Czech manuscripts in comparison with Čelakovský's *Echoes of Russian Songs* and Erben's collection of folk ballads.

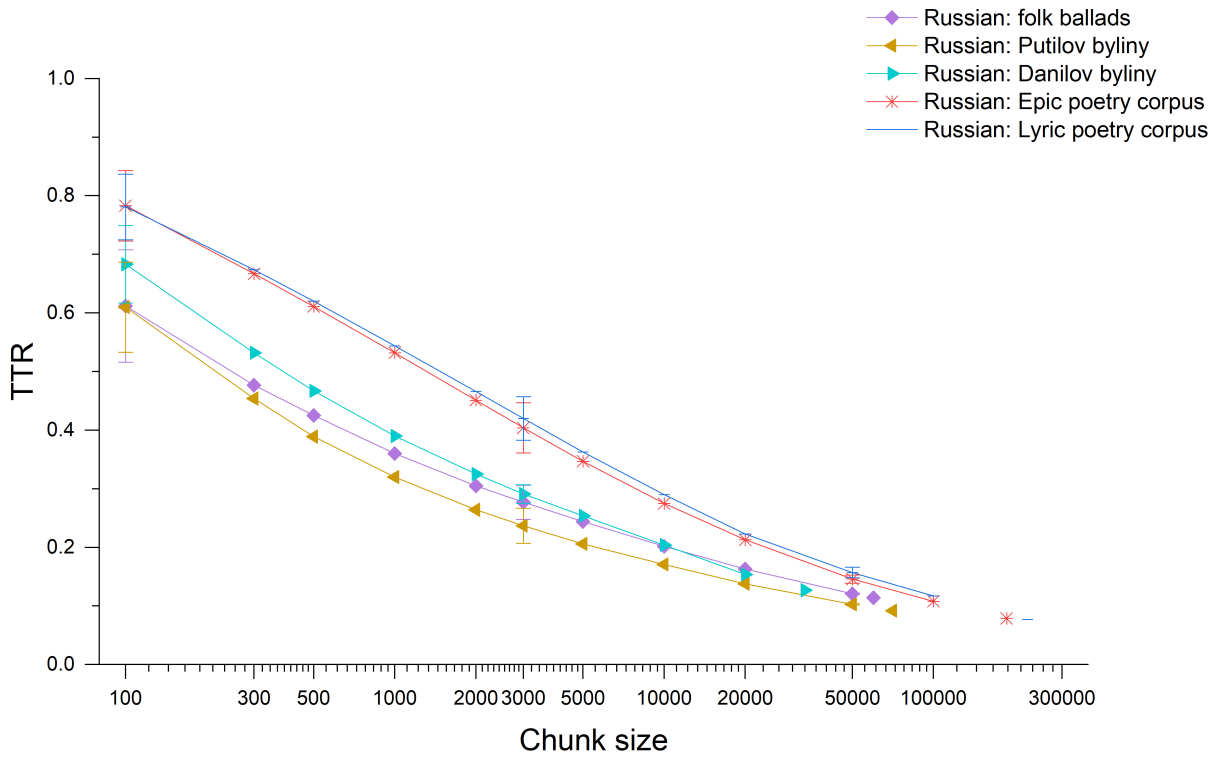

Fig. 7. Type-token ratio across different chunk sizes (on a logarithmic scale) for Russian literary epic and lyric poetry in comparison with Russian oral-formulaic epics and ballads.

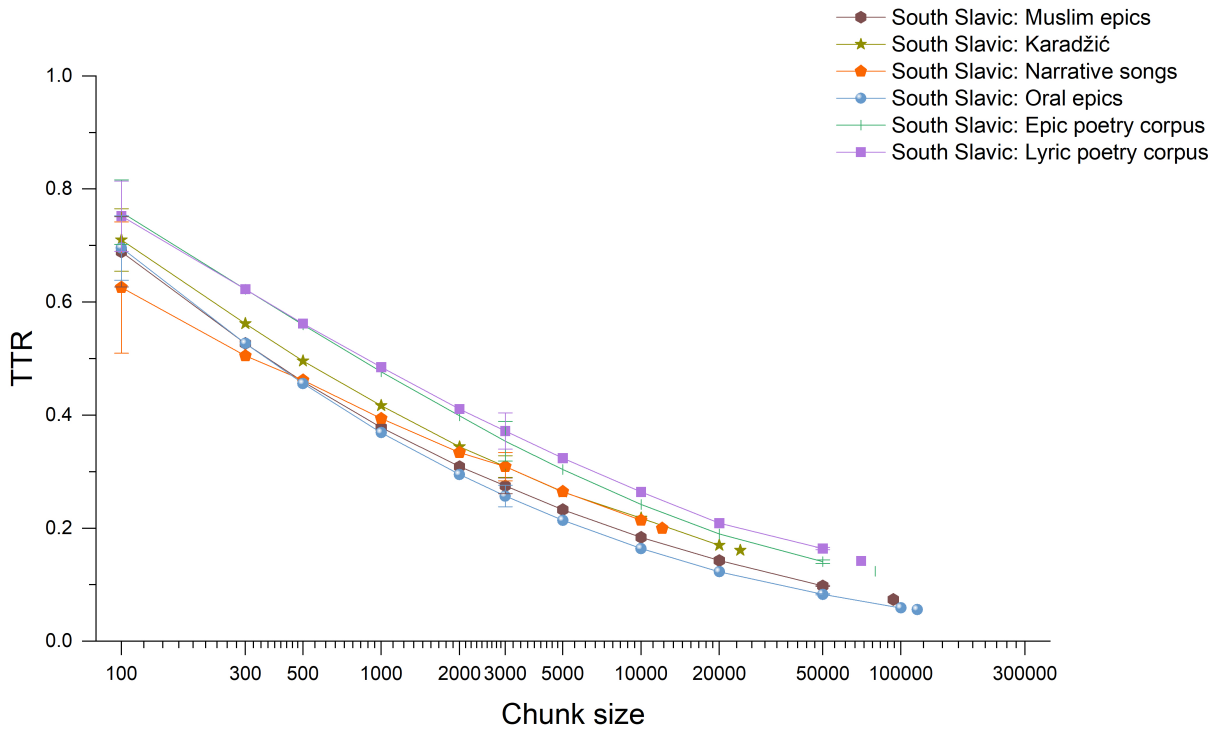

Fig. 8. Type-token ratio across different chunk sizes (on a logarithmic scale) for South Slavic literary epic and lyric poetry in comparison with South Slavic oral-formulaic epics and ballads (narrative songs).

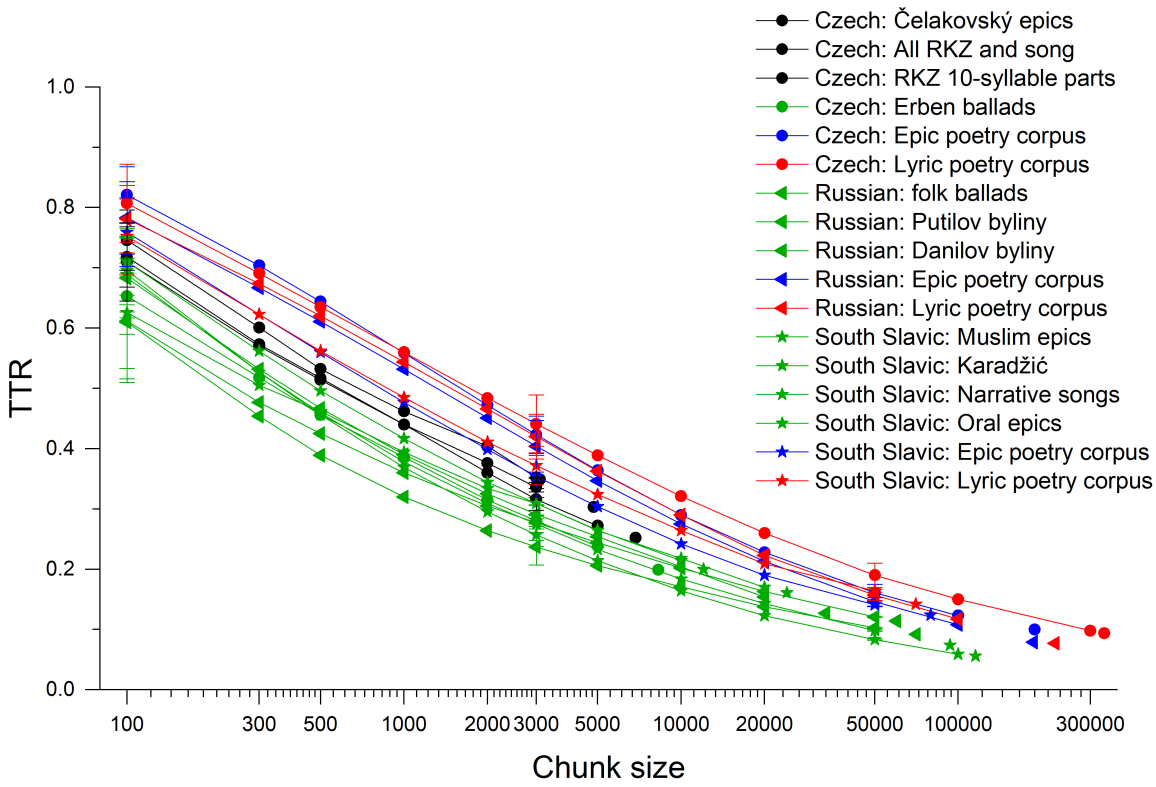

Fig. 9. Type-token ratio across different chunk sizes (on a logarithmic scale) for all study texts and corpora. Language is shown by shape, genre by color.

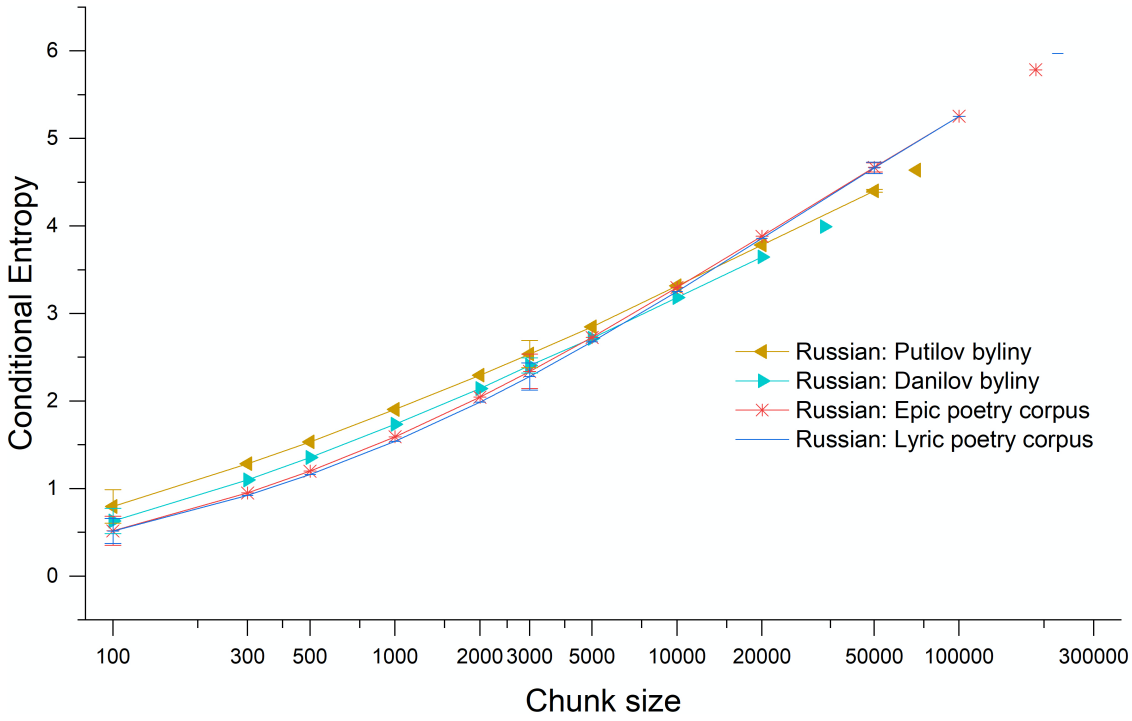

Fig. 10. Conditional entropy across different chunk sizes (on a logarithmic scale) for Russian oral epics in comparison to literary epic and lyric poetry.

References

Arant 1990 — Patricia M. Arant. *Compositional Techniques of the Russian Oral Epic, the Bylina*. New York: Garland.

Bailey and Ivanova 1998 — James Bailey, and Tatyana Ivanova, trans. *An Anthology of Russian Folk Epics*. Farmonk, NY: M. E. Sharpe.

Bowring 1827 — John Bowring. *Servian Popular Poetry, Translated by John Bowring*. London: the author.

Broadwell, Leonard, and Tangherlini 2017 — Peter M. Broadwell, Peter Leonard, and Timothy R. Tangherlini. "'Hvad der om dagen, blev revet ned om natten . . .': Word Sequence Repetition in Danish Legend Tradition." *Svenska landsmål och svenskt folkliv: Tidskrift för talspråksforskning, folkloristik och kulturhistoria*, 140:9-25.

Bynum 1978 — David E. Bynum. *The Dæmon in the Wood: A Study of Oral Narrative Patterns*. Publications of the Milman Parry Collection: Monograph Series, 1. Cambridge, MA: Center for Study of Oral Literature, Harvard University.

Cooper 2010 — David L. Cooper. *Creating the Nation: Identity and Aesthetics in Early Nineteenth-Century Russia and Bohemia*. DeKalb: Northern Illinois University Press.

Cooper 2018 — _____. *The Queen's Court and Green Mountain Manuscripts with Other Forgeries of the Czech Revival*. Czech Translations, 6. Ann Arbor: Michigan Slavic Publications.

Dobiáš 2010 — Dalibor Dobiáš. *Rukopis královédvorský; rukopis zelenohorský*. Brno: Host.

Dobiáš et al. 2015 — Dalibor Dobiáš, Michal Fránek, Martin Hrdina, Iva Krejčová, and Kateřina Piorecká. *Rukopisy královédvorský a zelenohorský a česká věda (1817-1885)*. Prague: Academia.

Dolanský 1968 — Julius Dolanský. *Neznámý jihoslovanský pramen Rukopisů královédvorského a zelenohorského.* Prague: Academia.

Erben 1886 — Karel Jaromír Erben. *Prostonárodní české písně a říkadle: s přílohou nápěvů*. Prague: Alois Hynek. https://catalog.hathitrust.org/Record/008958429

Flajšhans 1897 — Václav Flajšhans. *Podrobný seznam slov Rukopisu kralodvorského, se zvláštním zřetelem ke kritice čtení a výkladu.* Archiv pro lexikografii a dialektologii, 2. Prague: Nákladem české akademie Císaře Františka Josefa pro vědy, slovesnost a umění.

Foley 1990 — John Miles Foley. *Traditional Oral Epic: The Odyssey, Beowulf, and the Serbo-Croatian Return Song*. Berkeley: University of California Press.

Foley 1991 — _____. *Immanent Art: From Structure to Meaning in Traditional Oral Epic*. Bloomington: Indiana University Press.

Förster and Tieck 1818 — Friedrich Christoph Förster and Ludwig Tieck. *Die Sängerfahrt: Für Freunde der Dichtkunst und Mahleren*. Berlin: Maurer.

Hanka 1817 — Václav Hanka. *Prostonárodnj srbská muza*. Prague: J. F. Vetterl z Wildenbrunnu.

Ivanov 1969 — Miroslav Ivanov. *Tajemství RKZ*. Prague: Mladá Fronta.

Jakobson 1935 — Roman Jakobson. "K časovým otázkám nauky o českém verši." *Slovo a slovesnost*, 1:46-53.

Karadžić 1814 — Vuk Stefanović Karadžić. *Mala prostonarodnja Slaveno-Serbska pjesnarica*. Vienna: Schnierer.

Karadžić 1815 — _____. *Narodna srbska pjesnarica*. Vienna: Schnierer.

Karadžić 1987 — _____, ed. *Srpske narodne pjesme*. Belgrade: Prosveta.

Karadžić 1997 — _____. *Songs of the Serbian People: From the Collections of Vuk Karadžić*. Ed. by Milne Holton and Vasa D. Mihailovich. Pittsburgh, PA: University of Pittsburgh Press.

Kirdan 2001 — B. P. Kirdan, ed. *Ballady*. Biblioteka russkogo fol'klora, v. 6. Moscow: Russkaia kniga.

Kopitar 1816 — Jernej Kopitar. "Serbische Literatur." *Wiener allgemeine Literaturzeitung*, 20-21:314-33.

Kotelnikov et al. 2018 — Evgeny Kotelnikov, Elena Razova, and Irina Fishcheva. "A Close Look at Russian Morphological Parsers: Which One Is the Best?" In *Artificial Intelligence and Natural Language: 6th Conference, AINL 2017, St. Petersburg, Russia, September 20-23, 2017, Revised Selected Papers.* Ed. by Andrey Filchenkov, Lidia Pivovarova, and Jan Žižka. Communications in Computer and Information Science, 789. Cham, Switzerland: Springer International Publishing. pp. 131-42.

Krstanović 1990 — Zdravko Krstanović. *Zlatna pjena od mora: Narodne pjesme Srba u Hrvatskoj*. Belgrade: Rad.

Ljubešić et al. 2016 — Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, Tanja Samardžić, Maja Miličević, Filip Klubička, and Filip Petkovski. "Easily Accessible Language Technologies for Slovene, Croatian and Serbian." In *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija*. Ed. by Tomaž Erjavec and Darja Fišer. Ljubljana: Znanstvena založba Filozofske fakultete. pp. 120-24. http://nl.ijs.si/isjt16/proceedings-sl.html

Lord 1960 — Albert Bates Lord. *The Singer of Tales*. Harvard Studies in Comparative Literature, 24. Cambridge, MA: Harvard University Press.

Lord 1986 — _____. "Perspectives on Recent Work on the Oral Traditional Formula." *Oral Tradition*, 1.3:467-503. https://journal.oraltradition.org/wp-content/uploads/files/articles/1iii/Lord.pdf

Lubbe 1997 — J. C. A. van der Lubbe. *Information Theory*. Cambridge: Cambridge University Press.

Máchal 1918 — Jan Máchal. "Úvod." In *Hankovy písně a prostonárodní srbská muza, do Čech převedená*. Novočeská knihovna, 3. Prague: Nákladem české akademie císaře Františka Josefa pro vědy, slovesnost a umění. pp. ix-xliv.

Marjanović 1898 — Luka Marjanović, ed. *Hrvatske narodne pjesme, knjiga treća: junačke pjesme (Muhamedovske)*. Zagreb: Matica hrvatska. https://catalog.hathitrust.org/Record/102474245

Nikolayev 2016 — Dmitry Nikolayev. "A New Algorithm for Extracting Formulas from Poetic Texts and the Formulaic Density of Russian *Bylinas*." *Oral Tradition*, 30.1:111-36. https://journal.oraltradition.org/wp-content/uploads/files/articles/30i/06_30.1.pdf

Plecháč et al. n.d. — Petr Plecháč, Robert Kolár, Jakub Říha, and Dalibor Dobiáš. *Korpus českého verše*. Versologický tým. Institute of Czech Literature, Czech Academy of Sciences. https://versologie.cz/v2/web_content/corpus.php?lang=en

Popović 1968 — Bogdan Popović. *Antologija novije srpske lirike*. 12th ed. Belgrade: Srpska književna zadruga.

Straková et al. 2014 — Jana Straková, Milan Straka, and Jan Hajič. "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition." In

Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg, PA: Association for Computational Linguistics. pp. 13-18.

Tureček 2015 — Dalibor Tureček. “Hankova verze jihoslovanské hrdinské epiky.” *Bohemica litteraria*, 18.1:40-55.

Vikis-Freibergs and Freibergs 1978 — Vaira Vikis-Freibergs and Imants Freibergs. “Formulaic Analysis of the Computer-Accessible Corpus of Latvian Sun-Songs.” *Computers and the Humanities*, 12.4:329-39.

Vojtěch and Flajšhans 1930 — Viktorin Vojtěch and Václav Flajšhans, eds. *Rukopisy královédvorský a zelenohorský*. Prague: Česká grafická unie.