Studia Metrica et Poetica 9.1, 2022, 7–19

Metre and Semantics in the Poetry of Czech Post-Symbolists Accessed via LDA Topic Modelling

Petr Plecháč, Robert Kolár*

Abstract. The article deals with the relationship between semantics and poetic meter in the works of Czech post-symbolist poets and their predecessors. We access the phenomena by means of a machine-driven meter recognition on one hand and LDA topic modelling on the other. We first show how the poetic groups differ in their general preferences for particular topics. Next we analyze the topic distributions in two dominant metres (i.e. iamb and trochee) across the poetic groups.

Keywords: Czech poetry, semantic halo, topic modelling, versification

Introduction

In European (mostly Slavic) versification studies, the semantic halo of metre is a well-established theory which holds that poetic metre is not merely ornamental but rather carries its own semantic associations based on previous usage. First noticed by Russian formalists, described by Kiril Taranovsky (1963) and later systematically analysed by Mikhail Gasparov (1979, 1999), this theory has found supporting evidence in different traditions (cf. e.g. Tarlinskaja, Oganesova 1986; Pszczołowska 1988; Shapir 1991; Piperski 2017; Orekhov 2019). In the context of 19th-century Czech poetry, Miroslav Červenka and Květa Sgallová (cf. Červenka, Sgallová 1988, Červenka 1991, Červenka 1992) have documented this phenomenon in a number of studies. Their findings, when focusing only on two most common meters without distinguishing their variants, may be summarised roughly as follows:

1. Early in the development of accentual-syllabic versification (Dobrovský 1795), a basic opposition arose between the trochee and the iamb where the former (the prevailing form in poetic works) was seen as "local" and "traditional", while the latter was considered "foreign" and "contrived".

* Authors' address: Petr Plecháč / Robert Kolár, Institute of Czech Literature, Czech Academy of sciences, Na Florenci 1420/3, 110 00 Prague, Czech Republic, e-mail: plechac@ucl.cas.cz, kolar@ucl.cas.cz.

https://doi.org/10.12697/smp.2022.9.1.01

2. In the second half of the 19th century, the iamb became the dominant metre. The earlier semantic opposition was largely preserved, but the original negative associations of the iamb disappeared. As a result, what had been "foreign" became "cosmopolitan" while what had been "contrived" became "artistic". In the poetry of the Parnassian generation (dating mainly from the 1870s to the 1880s), this led to the iamb (namely iambic pentameter) becoming essentially a universal metre, while the trochee was associated with specific themes and genres such as historical events, rural settings and poetry for children.

3. In the poetry of the symbolists and decadents of the 1890s, a new form came into play: *vers libr*e. The iamb came to be seen as "traditional", while the trochee basically disappeared.

In fact, after some juvenile experimentation, members of the post-symbolist generation abandoned *vers libre* and incorporated not just the iamb but also the trochee into their poetry. To date, however, there has been no attempt to analyse semantic aspects of their poetic metre. This study aims to fill this gap. Unlike our predecessors, we apply statistical modelling of semantics (as introduced into the study of metrical semantics in Šeļa et al. 2020). Against this backdrop, we analyse not only the poetry of these post-symbolist authors, but also that of preceding generations to determine whether our approach replicates earlier findings.

Materials and methods

We draw on data from the Corpus of Czech Verse (Plecháč, Kolár 2015) – a collection of approximately 80,000 19th- and early 20th-century poems. This is supplemented by 1718 poems that together represent the work of four post-symbolist poets, namely Fráňa Šrámek (1877–1952), František Gellner (1881–1914), Viktor Dyk (1877–1931) and Karel Toman (1877–1946). Each poem has its metre labelled.

To model the semantics, we use Latent Dirichlet Allocation (LDA; Blei, Ng, Jordan 2003), which has become a common practice in large-scale semantic analyses of poetry (Navarro-Colorado 2018; Haider 2019; Plecháč, Haider 2020; Šeļa et al. 2020; Šeļa et al. 2022). To reduce lexical variability, we perform the common pre-processing steps: we rely on lemmatised texts and remove all parts of speech except for nouns, adjectives and verbs. Finally, to capture some common synonym rings and further simplify the vocabulary, we train a word2vec model (Mikolov et al. 2013} to filter words as follows:

1. We retain only the 1000 most common words.

2. For other words, we iterate over their 10 nearest neighbours in the word-2vec model. If the word arrived at belongs to the 1000-most-frequent-words group, it replaces the current word; otherwise, the word is dropped.

This method results in replacements such as *peřej* (rapids) > *vlna* (wave), *pološero* (semigloom) > *šero* (gloom) and *východní* (eastern) > *východ* (east)[1] along with the dropping of many low-frequency words.

An LDA model is then trained on the entire (simplified) corpus to identify 100 topics. Since for the LDA, a single topic represents a probability distribution over the entire vocabulary, we follow standard practice and label each topic based on its 5 highest-scoring words and an arbitrarily assigned index ranging from 1 to 100. This generates topics such as (1) *year, time, period/age, day, long* and (2) *woman, man, partner, young, beautiful*. (For a full list of topics in Czech, see the Acknowledgements below.)

Among the authors in the corpus, we focus our analysis on four schools based on standard literary periodisations. Besides the post-symbolists, who are our main concern, these schools are symbolist and decadent authors (grouped together as "symbolists") and two older groups generally known to contemporary literary historians under the umbrella term "Parnassians": *Lumír* (a cosmopolitan group) and *Ruch* (a national group).

From the works of these authors, we exclude both very short (less-than-4-line) and very long (more-than-100-line) texts. For the remaining works, LDA model is used to infer the topic probabilities in particular poems. Finally, each poem is represented as a vector of 100 topic probabilities and labelled with its (1) poetic metre, (2) author and (3) the group to which that author belongs.

Topic distribution across three generations of authors

We first address the relationship between poetic movements and topic distributions. Since poetic schools and movements are generally thought to favour specific themes and motifs, we expect our groups to differ in their overall affinities to particular topics.

To test this hypothesis, we represent each author by the average of all their poem-vectors and transform these values into *z*-scores. This gives us a simple way to measure the distances between particular authors so that topics that are generally less common have the same weight as the more frequent ones.

---

[1] Some high-frequency synonyms such as *matka*, *matička* and *máť* (all of which denote mother) are, however, retained in our data after this step.

Figure 1 shows that there are indeed affinities between poetic schools and topic preferences. Four clusters (labelled A, B, C, D) can be discerned that correspond roughly with the four schools. Even apparent misclassifications tend to accord with literary historians' findings: Both Borecký and Auředníček appear in cluster B along with the *Lumír* authors. Although the two are labelled here as symbolists, their work is generally thought to lie on the boundary between two generations (i.e. the Lumír/Ruch generation on the one hand and the symbolists on the other) (Červenka 1991: 16, 27). Sládek's poetry, on the other hand, is seen as rather civil and less rhetorical than that of his *Lumír* contemporaries (Červenka, Sgallová 1988). It is sometimes even explicitly compared to the work of the post-symbolist Toman (Novák 1994: 255; Červenka 1966: 153).

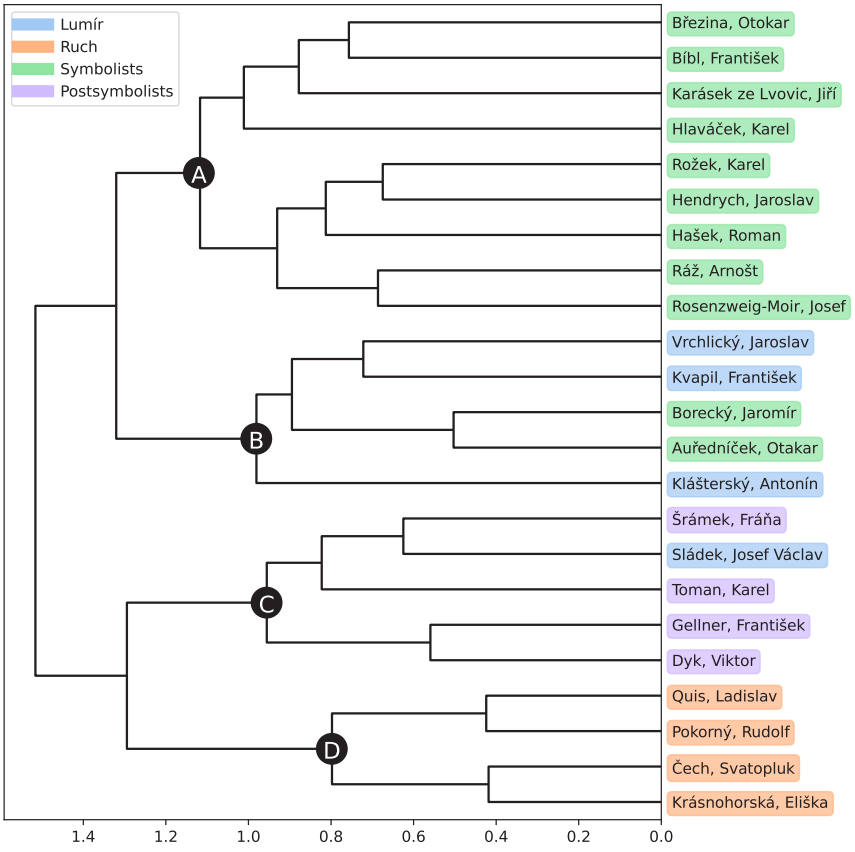

Figure 1. Dendrogram of authors' works in a topic-defined vector space (cosine distance, complete linkage)

To determine which topics are most typical for particular groups, we employ supervised machine learning. We first evaluate the accuracy of these models using cross-validation and then record the most significant features for particular classes. This experiment is set up as follows:

- We balance our dataset by reducing the symbolist group to the 4 most prolific authors (Borecký, Bíbl, Březina, Karásek).
- 5 random samples are taken from each author.
- The topic probabilities for each poem are transformed into *z*-scores.
- Each sample is represented by the average of the z-scores across its 10 poems.
- Leave-one-out cross validation is performed with a support vector machine (SVM) (linear kernel, $C = 1$).
- The procedure is repeated 10,000 times for each of the sample sizes $n \in \{1, 2, 3, 4, 5, 10, 15, 20\}$.

Figure 2 shows that even a classification of standalone poems significantly outperforms the random baseline (mean accuracy = 0.4, S.D. = 0.07; random baseline = 0.25). Nevertheless, these results fall far short of the values reported in Haider's 2019 large-scale study of German poetry; in that case, however, the classification was performed with isolated stanzas (which reportedly improved its accuracy) and the classes covered longer time spans. In our model, accuracy increases steadily as the sample size grows until at $n = 20$ (i.e. the largest possible sample size when 5 samples are taken from each author), it is as high as 0.92 (S.D. = 0.04).

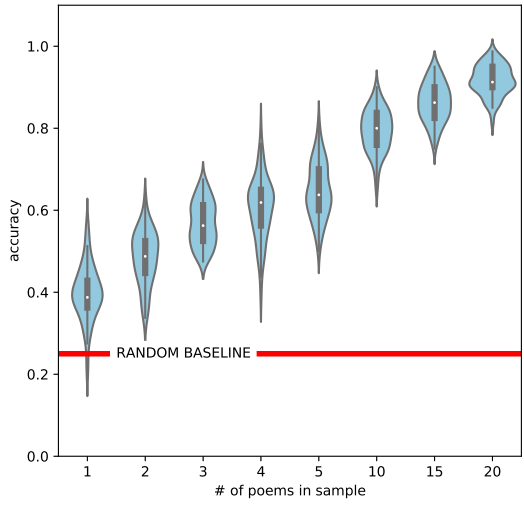

Figure 2. Cross-validation of SVM models trained on samples of different sizes

Table 1 lists the five most significant features for each group (this is based on the mean of the respective values for normal vectors across 100 models where $n = 20$). These findings largely conform with expectations. In the *Lumír* group, topics with bright and optimistic connotations are common, while in the *symbolist* group, gloomy and obscure topics (36, 50, 90) prevail. In the *Ruch* group, human collectives are a key topic, and this may involve humanity in general (51), the nation (78) or the family (40). In contrast, among the *post-symbolists*, poems often centre on the human individual and their activities, feelings and inter-personal relations. (Notably poets in this group also use a significantly higher proportion of verbs than their counterparts in other schools[2]).

Table 1. Five most contributing features for each group

Lumír	*Ruch*
(84) to see, sight, cheek, face, dream (45) bird, to fly, wing, nest, butterfly (23) dream, youth, to dream, fairytale, soul (69) child, small, childish, kid, big (96) laughter, joy, ball, cheery, to laugh	(52) eye, cheek, word, dark, sight (78) motherland, Czech, Czech, nation, Bohemia (13) beautiful, flower, grove, sweet, sun (51) age, glory, big, humanity, famous (40) mother, mother, child, mother, father
Symbolists	*Postsymbolists*
(83) flame, fire, hot, blood, heat (36) shadow, soul, evening, fog, gloom (90) soul, sadness, pain, sad, heart (50) space, mysterious, eternity, Earth, secret (95) pearl, golden, nymph, breast, flower	(35) to come, to wait, to be coming, to enter, to long for (27) to believe, friend, companion, to remember, name (2) woman, man, partner, young, beautiful (11) to play, to laugh, to dance, game, circle (56) to want, to say, to give, to get, to do

[2] Cf. „Just like the symbolists, the [post-symbolists] have a dream, but this dream is grounded in reality. Poets dream of a better life, and to make this beautiful dream come true, they require courage and daring, action, commitment, a sense of greatness and the ability to convey human truths and desires through ordinary intelligible signs." (Vodička 2001: 34, translation: pp).

Association between topics and poetic metres

We proceed to the association between topic distribution and poetic metre. Our aim is to determine whether – despite the apparent differences in overall topic probabilities – some continuity exists between the *Lumír* and *Ruch* groups and the post-symbolists in terms of how particular topics affect the choice of metre.

To ensure sufficient data, we confine our analysis to the two most common metres (the iamb and the trochee) and do not distinguish among their variants. (Iambic trimeter, iambic tetrameter and iambic pentameter are, thus, all treated as members of the same iambic class.). In addition, we focus exclusively on the *Lumír*, *Ruch* and post-symbolist groups since most symbolist poems are written in iambic metre or *vers libre* and contain very few trochees.

Figure 3 shows the results of clustering performed in the same manner as in Figure 1, only here each author's works are divided into iambic and trochaic poems[3] and *z*-scores are calculated not for the entire dataset but for each author's works separately.[4] The dendrogram clearly shows that the same metres tend to cluster together irrespective of the group that the author belongs to: there is a purely iambic cluster at the top, a purely trochaic cluster at the bottom and a mixed cluster in the middle comprising three samples only.

[3] Only monometric poems, i.e. those written exclusively in iambs or trochees are taken into account.

[4] In other words, our concern is not whether author *A* writes iambic poems about topic *T* more often than author *B*, but rather whether topic *T* is more likely to appear in an iambic poem than in a trochaic poem within *A*'s works.

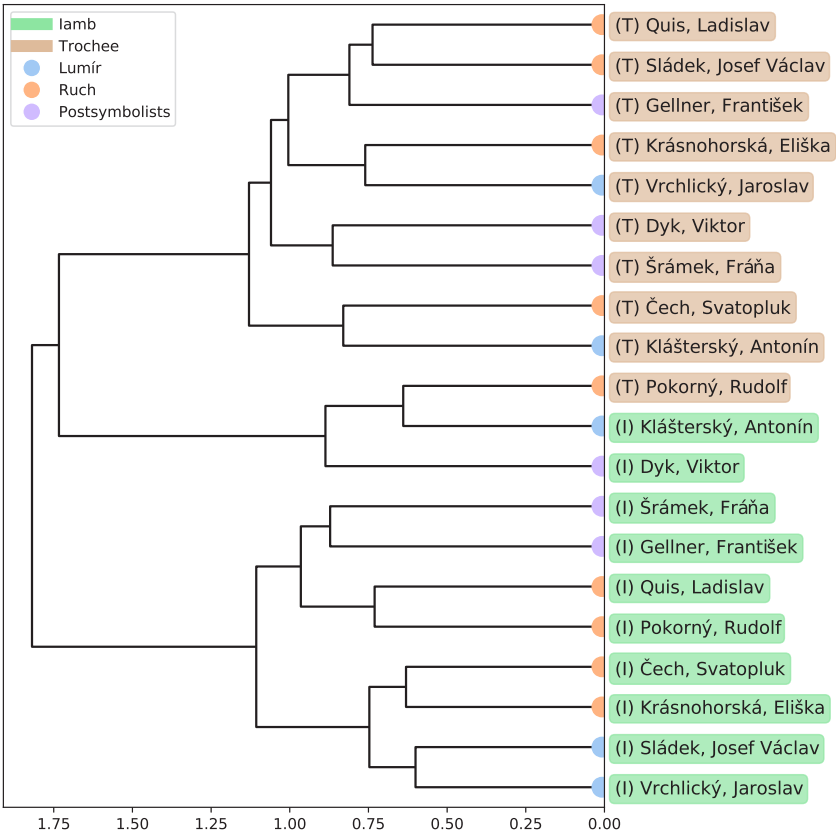

Figure 3. Dendrogram of authors' iambic and trochaic works in a topic-defined vector space (cosine distance, complete linkage)

This does indeed suggest some sort of continuity of topic–metre associations. To track this further, we train a new set of Support Vector Machine models to recognise poetic metre with each group separately. This experiment is set up as follows:

- From each of the 3 groups, 15 random samples are taken; each sample comprises 20 poems for each poetic metre.
- The model is used to classify samples from the 2 remaining groups.
- The procedure is repeated 10,000 times.

Figure 4 shows the results. When models are trained with the group that the samples came from (cross-validation), the average accuracy rate is between 0.71 and 0.8. The accuracy for cross-group metre recognition is lower (between 0.6 and 0.77 on average), but it still markedly outperforms the random baseline level (0.5).

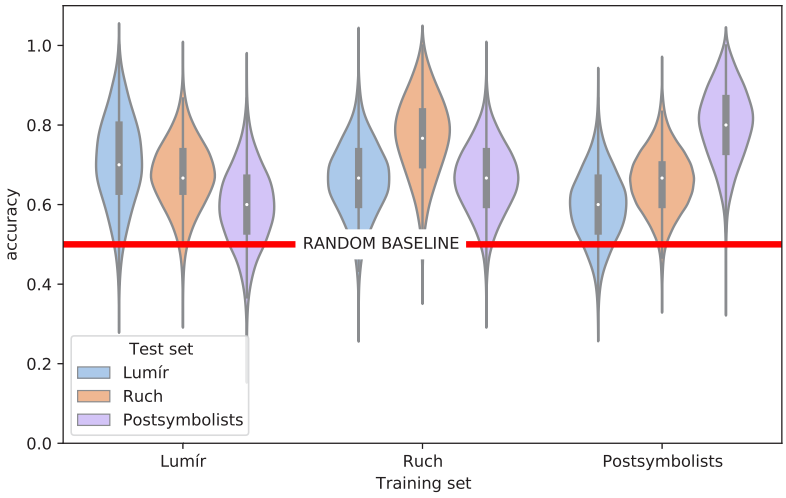

Figure 4. Accuracy of metre recognition

A general picture emerges when we consider the most important features for each of the three sets of models (Tab. 2). For the *Lumír* group, this list seems to confirm the observations of Červenka and Sgallová that are summarised in our introduction (iambic pentameter is the universal metre while the trochee is reserved for specific themes and genres such as historical events, rural settings and poems for children). In the iambic group, we, thus, find the signature topics of art (74) and humanity (51) along with other somewhat unrelated and obscure themes; in contrast, in the trochaic list, the prevailing topics are linked to historical events (68, 10) and youth (34, 53). The poems in the *Ruch* school's iambic list also concern typical topics for this group such as humanity (51) and emancipation (89), to which also the theme of war (10) may be added. This may seem an opportunistic inclusion since in the *Lumír* group, we linked this topic to historical events. A closer look, however, reveals that the thematic context is quite different for the two groups. While in poems from the *Lumír* group, topic (10) is associated most often (based on Pearson's *r* measure) with historical references such as (68) *emperor, Rome, to give, god, to go* and (64) *king, throne, empire, proud, crown*, in the *Ruch* group, it rather relates to emancipation and co-occurs with topics such as (51) *age, glory, big, humanity, famous*, (65) *power, labour, strong, work, hand* and (77) *shackles, free, slave, wild, freedom*. Overall trochaic topics appear more likely to be rooted in folklore.

Table 2. Five most contributing features by poetic metre for each group

Lumír	
Iamb	Trochee
(84) to see, sight, cheek, face, dream (51) age, glory, big, humanity, famous (74) beauty, art, shape, charming (30) blood, fear, death, anger, murderer (95) pearl, golden, nymph, breast, flower	(68) emperor, Rome, to give, god, to go (34) boy, girl, girl, young, girl (9) sir, man, thing, wise, advice (10) fight, army, sword, weapon, hero (53) old, new, young, nice, school
Ruch	
Iamb	Trochee
(51) age, glory, big, humanity, famous (84) to see, sight, cheek, face, dream (10) fight, army, sword, weapon, hero (89) people, nation, freedom, flag, liberty (52) eye, cheek, word, dark, sight	(32) boyfriend, dear, to cry, girlfriend, to meet (71) head, hand, hair, white, golden (1) year, time, period/age, day, long (9) sir, man, thing, wise, advice (60) window, door, cottage, village, room
Post-symbolists	
Iamb	Trochee
(94) rose, to bloom, red, flower, bush (36) shadow, soul, evening, fog, gloom (83) flame, fire, hot, blood, heat (90) soul, sadness, pain, sad, heart (50) space, mysterious, eternity, Earth, secret	(4) to say, to ask, to know, to tell, to tell (42) God, heaven, to protect, to give, to create (66) sea, boat, wave, shore/bank, to sail (16) horse, to drive/hurry, to run, jump, to drive (35) to come, to wait, to be coming, to enter, to long for

To confirm that these results are not based on cherry-picking of the data (i.e. over-reliance on top-scoring words), we model the action/description ratio for each poem using the Busemann Coefficient (cf. e.g. Andreev, Místecký 2018). This is defined as

$$Q = \frac{V}{V + A}$$

where V is the number of verbs and A is the number of adjectives.

The results strongly corroborate our hypothesis. In the post-symbolist group, the values for Q are significantly higher in trochaic poems (mean = 0.73, S.D. = 0.12) than they are in iambic poems (mean = 0.69, S.D. = 0.11); t = 6.75, $P > 10^{-10}$, Cohen's d = 0.44. This does not hold for the *Lumír* group where the difference is far more negligible (mean(trochee) = 0.684, mean(iamb) = 0.679, S.D.(trochee) = 0.13, S.D.(iamb) = 0.12; t = 1.38, P = 0.17, Cohen's

$d = 0.03$) or for the *Ruch* group where the difference hovers around the 0.05 alpha level (mean(trochee) = 0.64, mean(iamb) = 0.63, S.D.(trochee) = 0.16, S.D.(iamb) = 0.12; $t = 2.01$, $P = 0.05$, Cohen's $d = 0.1$).

Conclusions

Despite marked differences in the topic preferences of the poetic groups that we analysed, our findings show that the post-symbolist association of two dominant metres (the iamb and the trochee) with specific topic distributions stems from 19th-century poetics. A closer look also reveals the emergence of a new organising principle, which relates to the distinction between action and description.

Acknowledgements

For both LDA and word2vec, we used the implementation in Gensim Python library (Rehurek, Sojka 2011). Lemmatisation and POS-tagging were performed using the MorphoDiTa tagger (Straková et al. 2014). Metrical recognition was achieved via the metrical tagger KVĚTA (Plecháč 2016).

Data (including a full list of topics in Czech) and the code to reproduce the analysis are available at https://github.com/versotym/lda-czech.[5]

References

Andreev, Sergey; Místecký, Michal 2018. Activity in Czech and Russian nineteenth-century sonnets: A contrastive study. In: *Glottotheory* 9(1), 89–104. https://doi.org/10.1515/glot-2018-0004

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. 2003. Latent dirichlet allocation. In: *Journal of Machine Learning Research* 3, 993–1022. https://jmlr.csail.mit.edu/papers/v3/blei03a.html

Červenka, Miroslav 1966. *Symboly, písně a mýty*. Praha: Československý spisovatel.

[5] This study was supported by Czech Science Foundation (GAČR 20-15650S).

Červenka, Miroslav 1991. *Z večerní školy versologie II: Sémantika a funkce veršových útvarů*. Prague: Akcent.

Červenka, Miroslav 1992. Lumírovec: sémantika verše v Zeyerově epice. In: *Slovo a Slovesnost* 53, 241–247.

Červenka, Miroslav; Sgallová, Květa 1988. Český verš. Sémantika metra v poezii lumírovců. In: Pszczołowska, Lucylla (ed.), *Słowiańska metryka porównawcza* 3: *Semantyka form wierszowych*. Wrocław: Zaklad Narodowy Imienia Ossolińskich, 55–104.

Dobrovský, Josef 1795. Böhmische Prosodie. In: Pelcl, František Martin, *Grundsätze der Böhmishen Grammatik*. Praha: František Jeřábek, 209–246.

Gasparov, Mikhail Leonovich 1979. Semanticheskij oreol metra. K semantike russkogo trekhstopnogo jamba. In: *Lingvistika i poetika*. Moskva: Nauka, 282–308.

Gasparov, Mikhail Leonovich 1999. *Metr i smysl. Ob odnom iz mekhanizmov kul'turnoj pamjati*. Moskva: RGGU.

Haider, Thomas Nikolaus 2019. Diachronic topics in new high German poetry. In: *DH2019 Book of Abstracts*. Utrecht University. https://dev.clariah.nl/files/dh2019/boa/1031.html

Mikolov, Tomas; Chen, Kai, Corrado, Greg; Dean, Jeffrey 2013. Efficient estimation of word representations in vector space. In: *arXiv*. https://arxiv.org/abs/1301.3781

Navarro-Colorado, Borja 2018. On poetic topic modeling: Extracting themes and motifs from a corpus of Spanish poetry. In: *Frontiers in Digital Humanities* 5, 15. https://doi.org/10.3389/fdigh.2018.00015

Novák, Arne 1994. *Dějiny českého písemnictví*. Praha: Brána.

Orekhov, Boris 2019. *Bashkirskij stikh XX veka. Korpusnoe issledovanie*. St. Petersburg: Aletejia.

Piperski, Aleksandr 2017. Semantic halo of a meter: a keyword-based approach. In: *Komp'juternaja lingvistka i intellektual'nye tekhnoloigii / Computational Linguistics and Intellectual Technologies* 16(2), 342–354. https://www.dialog-21.ru/media/3936/piperskiach.pdf

Plecháč, Petr 2016. Czech verse processing system KVĚTA: Phonetic and metrical components. In: *Glottotheory* 7(2), 159–174. https://doi.org/10.1515/glot-2016-0013

Plecháč, Petr; Haider, Thomas Nikolaus 2020. Mapping topic evolution across poetic traditions. In: *DH2020: Book of Abstracts*. https://dh2020.adho.org/wp-content/uploads/2020/07/600_MappingTopicEvolutionAcrossPoeticTraditions.html

Plecháč, Petr; Kolár, Robert 2015. The Corpus of Czech Verse. In: *Studia Metrica et Poetica* 2(1), 107–118. https://doi.org/10.12697/smp.2015.2.1.05

Pszczołowska, Lucylla (ed.) 1988. *Slowiańska metryka porównawcza* 3: *Semantyka form wierszowych*. Wroclaw: Zaklad Narodowy Imienia Ossolińskich.

Rehurek, Radim; Sojka, Petr 2011. Gensim–python framework for vector space modelling. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3(2).

Šeļa, Artjoms; Orekhov, Boris; Leibov, Roman 2020. Weak genres: Modeling association between poetic meter and meaning in Russian poetry. In: *CHR 2020: Workshop on Computational Humanities Research*. Amsterdam: CEUR-WS, 12–31. http://ceur-ws.org/Vol-2723/long35.pdf.

Šeļa, Artjoms; Plecháč, Petr; Lassche, Alie 2022. Semantics of European poetry is shaped by conservative forces: The relationship between poetic meter and meaning in accentual-syllabic verse. In: *PLOS One* 17(4). https://doi.org/10.1371/journal.pone.0266556

Shapir, Maksim 1991. "Semanticheskij oreol metra": Termin i ponjatie (Istoriko-stikhovedskaja retrospektsija). In: *Literaturnoe obozrenie* 12, 36–40.

Straková, Jana; Straka, Milan; Hajič, Jan 2014. 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.* Baltimore, Maryland: Association for Computational Linguistics, 13–18. https://doi.org/10.3115/v1/P14-5003

Taranovsky, Kiril 1963. O vzaimootnoshenii stikhotvornogo ritma i tematiki. In: *American Contributions to the Fifth International Congress of Slavists.* The Hague: Mouton & Co., 287–332.

Tarlinskaja, Marina; Oganesova, Naira 1986. Meter and meaning: The semantic halo of verse form in English romantic lyrical poems (iambic and trochaic tetrameter). In: *The American Journal of Semiotics* 4(3/4), 85–106. https://doi.org/10.5840/ajs198643/422

Vodička, Felix 2001. Literatura na počátku 20. století. In: Janáčková, Jaroslava; Hrabáková, Jaroslava (eds.), *Česká literatura na předělu století*. Jinočany: H&H, 125–148.