Frequency of parts of speech in Adolf Heyduk's poetry

Petr Plecháč – Robert Ibrahim

Academy of Sciences of the Czech Republic, Institute of Czech Literature,
Na Florenci 3/1420, 110 00 Prague 1, Czech Republic.
Email: plechac@ucl.cas.cz, ibrahim@ucl.cas.cz

Abstract:

The paper explores the relationship of the frequency of parts of speech with i) the verse length, ii) literary genre (lyrics vs. epics) and iii) metre (trochee vs. iamb). For the purpose of this analysis, approximately 110,000 verse lines from a Czech poet Adolf Heyduk (1835–1923) have been processed, most of which rhyme. The analysis has demonstrated that the frequency of parts of speech depends on the verse length and on the literary genre. The frequencies of parts of speech in the trochee and iamb and in rhymed and unrhymed verse do not differ. The frequency of parts of speech in the beginning and end of line differs from their frequency in the whole text.

Keywords:
Czech poetry of the XIXth Century, Adolf Heyduk, parts of speech, theory of verse, computational linguistics, prosody, digital humanities

The aim of the present paper is to analyse the frequency of parts of speech in Adolf Heyduk's poetry.[1] It is a preliminary study based on extensive material, which may have an impact on the achieved results. Therefore, we wish to point out that our conclusions are not generalizable to all Czech poetry. Let us first briefly outline what will be analysed and on what material.[2]

The Institute of Czech literature AS CR have at their disposal The Czech Electronic Library, which is a publicly accessible database of nineteenth century Czech poetry comprising approximately 1,700 collections (available at www.ucl.cas.cz). The authors of this study used this database in order to compile the Corpus of Czech verse (Korpus českého verše – KČV, henceforth). It contains more than 2 million verse lines, (approximately 15 million words), i.e., a phonetically, morphematically, morphologically and metrically annotated collection of texts. These annotations are carried out automatically.[3] The automatic morphological annotation is based on part-of-speech tagging. The frequency of parts of speech in KČV can be compared for example with the frequency of parts of speech in the SYN2005 corpus, i.e. a corpus of contemporary written Czech, containing 100 million words (tokens) and consisting of three subcorpora (fiction, technical literature, journalism)—see Czech National Corpus (http://ucnk.ff.cuni.cz/syn2005.php).

The present paper is a follow up to our study *Frequency of parts of speech in Czech poetry* (Ibrahim and Plecháč, in press). There we compared the parts-of-speech frequency in KČV and SYN2005 (and its subcorpora), the relation of morphological and syntactic level to the length of verse, and the dependence of the frequency of parts of speech on the school of poetry or the author´s style. Here we look at the relationship of the frequency of parts of speech with i) the verse length, ii) literary genre (lyrics vs. epics) and iii) metre (trochee vs. iamb).

Since the metrical annotation is currently manually controlled, we were forced to select one author, whose texts had been subjected both to automatic computer and manual analysis. The author is a Czech poet Adolf Heyduk (1835–1923). For the purpose of this analysis, we processed approximately 110,000 verse lines from his works, most of which rhyme (approx. 84 per cent). As regards the metre that Heyduk uses, an iamb and a trochee prevail; they account for 97 percent of all the analysed verse lines (approx. 54 per cent of iambs, 44 per cent of trochees). For this reason, we will only deal with iambic and trochaic verse lines.

In the next section, we attempt to answer the following questions:

Q1 Does the frequency of individual parts of speech differ depending on the verse length?

---

1 This study and its translation were supported by the long-term conceptual development of a research institution (68378068) and by a grant from the Czech Science Foundation (GA ČR, 406/11/1825).

2 From works dealing with the problems of sentence and verse length, the relation of parts of speech, clause elements to verse we select: Pszczołowska (1965), Mazáčová (1973), Červenka and Sgallová (1984), and Gasparov and Skulačeva (2004).

3 We have finished the phonetic annotation. Currently, we are working on the morphematic annotation together with the Institute of Formal and Applied Linguistics MFF UK and wish to thank Jaroslava Hlaváčová for her assistance. We acquired the morphological annotation thanks to our cooperation with The Department of Theoretical and Computational Linguistics (we would like to thank the head of the department Vladimír Petkevič and Hana Skoumalová). As for the metrical annotation, we are in the process of verifying the results. The author of the PC programme for phonetic and metrical analysis is Petr Plecháč.

Q2 Does the frequency of individual parts of speech differ depending on the literary genre?
Q3 Does the frequency of individual parts of speech differ depending on whether the verse is trochaic or iambic?
Q4 Does the frequency of individual parts of speech differ depending on whether the verse is rhymed or unrhymed?
Q5 Is there any deviation in the beginning or end of a verse line in view of the frequency of parts of speech in the whole text?[4]
Q6 Does the frequency of parts of speech in the beginning of a line differ depending on whether the verse is trochaic or iambic?
Q7 Does the frequency of parts of speech in the end of line differ depending on whether the verse is rhymed or unrhymed?
Here are the answers:

A1
We are working with the assumption that in Czech syllabotonic verse the correspondence of verse and syntactic segmentation is considered unmarked. In that case, enjambment (the discrepancy between verse and syntactic segmentation) is one of the means of differentiation of verse style. Nevertheless, even in the works of those authors who typically use enjambment, it represents only a relatively small percentage of verse, and hence the statistical significance of correspondence of verse and syntactic boundaries remains strong (see Červenka and Sgallová 1984, 13–14). From short to long verses, the number of positions that need to be occupied by lexical units grows. Thus, if the equation "verse boundary = boundary of a syntactic unit" holds to a large extent true, the poet has three possibilities of "filling" more positions in long verses: 1) "to fill" the longer verse by more syntactic units (clauses), 2) to use longer words and/or 3) "to fill" a longer verse by optional clause elements. In the present study, we focus on the last possibility (the cases 1 and 2 were analysed in Ibrahim and Plecháč; in press). From the parts-of-speech point of view the optional clause elements are realized predominantly by adjectives and adverbs. However, it is not the frequency of these parts of speech that is the main indicator, but the ratio of basic to optional parts of speech that can be expressed by the coefficient NV/AD, i.e., the ratio of the frequency of nouns and verbs to the frequency of adjectives and adverbs. The tendency is that the longer a verse is the lower its NV/AD coefficient, i.e., it contains more optional parts of speech—see Table 1 (we provide only six- and twelve-syllable verse lines, the remaining sample of the *n*-syllable lines being very small).

A2
Table 2 demonstrates that there are no great differences between Heyduk's lyric and epic poetry.[5] However, the analysis has confirmed our expectations: the lyric has a higher frequency of

4 In the present paper, the extent of the beginning, and the end of a verse line have been restricted to the first and the last word, respectively.

5 The values concerning Heyduk's epic are based on the works *Oldřich a Božena*, *Dědův odkaz*, *Dudák*, *Na přástkách*, *Pod Vítkovým kamenem*, *Mohamed II.*, *Dřevorubec*, *Za volnost a víru*, *Běla*, *Na vlnách*; the values concerning Heyduk's lyric are based on the works *Hořec a srdečník*, *V zátiší*, *Šípy a paprsky*, *Lesní kvítí*, *Na potulkách*, *Ptačí motivy*.

nominal groups (nouns and adjectives) than the epic, while the epic has a higher frequency of verbal groups (verbs and adverbs). In the lyric, the AN/DV coefficient is 1.80, in the epic it is 1.42. In the basic parts of speech (nouns, adjectives, pronouns, verbs and adverbs), the values of the epic resemble the average values in KČV, or the values in SYN2005 (the subcorpus fiction).

A3

Table 3 shows the frequency of parts of speech in the trochee and iamb. The data concerning trochaic verse very much resemble the data in the whole KČV. The frequency of parts of speech in iambic verse does not differ from the frequency of trochaic verse, the only difference being the reverse order of conjunctions and adjectives. Below we discuss the question why there are more conjunctions and fewer adjectives in iambic verse (see A6).

A4

The frequency of parts of speech in rhymed and unrhymed verse does not differ (see Table 3).

A5

The frequency of parts of speech in the beginning and end of a line differs from the frequency in the whole text (see Table 3).

A6

The beginning of a trochaic verse line

Let us start with the beginning of a trochaic verse line (see Table 3). We can see that there is a considerable increase in prepositions and conjunctions as well as a slight increase in the frequency of verbs and adverbs, while the frequency of nouns and pronouns decreased.

Miroslav Červenka (2006, 94) included among the correspondence rules of the Czech syllabotonic verse a rule that the first strong position of trochaic verse corresponds to the stressed syllable. This rule may explain the higher distribution of prepositions and the lower distribution of pronouns in the trochaic beginning of a line: in the prosody of Czech verse prepositions are always stressed, while monosyllabic pronouns are unstressed. However, this rule cannot explain the higher frequency of conjunctions (in Czech monosyllabic conjunctions are unstressed). Miroslav Červenka even quotes Heyduk's example "A všed v jizbu, pravil: 'bude krásně!'" pointing out that such trochaic verses (i.e., verses with unstressed first syllable) are rare. Petr Plecháč demonstrated (2012, 402n) that the above-mentioned correspondence rule is not justified in the description of the metrical norm of Czech syllabotonic verse. This is because the distribution of unstressed beginning of a line in the trochaic verse exceeds language probability (i.e. the number of unstressed trochaic beginnings of a line in Heyduk and in the whole KČV is higher than we would expect based on probability). The reasons for the realization of the trochaic beginning of a line by an unstressed syllable (e.g., a conjunction) may differ depending on the author or text. The reasons are several: the effort for rhythmic diversification, the influence of syllabic verse or rhythmic habits (see Červenka 2006, 94nn). The higher frequency of conjunctions in the beginning of a line is also associated with the fact that the beginning of a line is identical with the beginning of a syntactic unit and that it always follows the prosodic boundary. The anacruses (some of them are of course units beginning with a monosyllabic conjunction) cluster (see Červenka 2006, 88) after the prosodic boundary.

The beginning of an iambic verse line

Even greater deviation from the values counted for the whole text can be found at the beginning

of a line in iambic verse. Here conjunctions, pronouns and adverbs prevail (there is also a higher frequency of particles). On the other hand, the frequency of nouns, verbs and particularly adjectives is lower than their frequency in the whole iambic or trochaic verse (or the trochaic beginning of a line). The iambic beginning of a line in the Czech theory of verse has received much attention (see Jakobson 1979). Miroslav Červenka proposes five types of the beginning of a line in the Czech iamb. Adolf Heyduk belongs to the strictest type, i.e., the first position of a line is always occupied by a monosyllable, the frequency of unstressed monosyllables being higher than the frequency of stressed ones (the strictness or orthodoxy of this type stems from the radical interference in the rhythmic lexicon), which is represented in KČV only by a few authors (see Červenka 2006, 89). This explains the noticeable deviation in the distribution of individual parts of speech at the beginning of lines in Heyduk's iambs (we could obtain different results with another author; here it is the type of beginning of a line that plays a role).

If the author needs monosyllabic words, a part of speech that meets this requirement most is the conjunctions (89 per cent of Heyduk's conjunctions are monosyllabic), pronouns (82 per cent of Heyduk's pronouns are monosyllabic), particles (81 per cent), or adverbs (48 per cent). The parts of speech that meets this requirement the least are for instance adjectives (only 3 percent of Heyduk's adjectives are monosyllabic).[6] These values have been counted on the basis of Heyduk's verse and they can be compared with the data in SYN2005 and SYN2005 (the subcorpus fiction). In SYN2005, or more precisely in SYN2005 (the subcorpus fiction) there are 71.7 per cent, or more precisely 70.4 percent of monosyllabic conjunctions, 70.8 per cent, or more precisely 75.4 per cent of monosyllabic pronouns, 40 per cent, or more precisely 45.7 per cent of monosyllabic particles, and 30.7 per cent, or more precisely 34.4 per cent of monosyllabic adverbs. Hence, in verse (at least in Heyduk's verse) there are more monosyllabic words than in SYN2005 and SYN2005 (the subcorpus fiction), which is to a great extent caused by iambic verse.[7]

The end of a line

In the case of the end of a verse line, we do not distinguish between the iambic or trochaic verse, or between the masculine and feminine ending. Table 3 shows that in the end of lines nouns and verbs absolutely prevail, while the number of pronouns considerably decreases.[8] The values of adjectives and adverbs are similar to the values in the whole text. The frequency of the other parts of speech does not exceed 1 per cent. The explanation relates to the position at the end of a line: some parts of speech (e.g., prepositions and conjunctions) do not occur before the syntactic boundary, or their occurrence would be marked (an adjective in the end of a line would be either an instance of inversion or enjambment). The end of a line is thus a place where not only the syntactic, prosodic and sound accentuation but also the semantic reinforcement occur (there is an accumulation of lexical parts of speech).

---

6 Note that lots of monosyllabic words can also be found among interjections and prepositions. Prepositions are always stressed in Czech. The combination preposition + noun are considered a polysyllabic word, and thus they do not comply with the strictest type of the iambic beginning of a line.

7 The statistics of n-syllabic words with selected Czech poets and in prose clearly confirm the increase in the frequency of disyllabic words in trochee and monosyllabic words in iamb (Červenka and Sgallová 1978).

8 A high prevalence of nouns and verbs at the end of a line was also observed in Russian verse (based on works of A. S. Puškin, K. N. Batjuškov a J. A. Baratynskij). See Shaw (1993); Gasparov and Skulačeva (2004, 67).

The frequency of parts of speech in the last position of a line influences the frequency of parts of speech in the penultimate position of a line. If the most frequent parts of speech in the end of a line are nouns, it is likely that there will be an increase in the occurrence of adjectives in the penultimate position and Heyduk's verse confirms this.[9] Thus, at the end of a line we observe a frequent occurrence of the syntagma adjective + noun. Gasparov and Skulačeva (2004, 271–272) point out that this strong syntactic bond is typical for the end of a line, as it underlines—in contrast with the weak line-internal bonds—the end of a line and it enables to single out the verse as an independent, by senses perceivable, unit (in prose, the present authors have not encountered such tendency). More verbs in the end of a line cause more pronouns in the penultimate position.[10]

A7
At the end of a line, in rhymed and unrhymed verse, there is not a great difference in the frequency of parts of speech.[11] On the other hand, at the end of rhymed verse lines, there is a slightly higher number of nouns and verbs and a slightly lower number of adjectives and pronouns than at the end of unrhymed verse lines. This might be caused by grammatical rhymes. As can be seen from Table 4, approximately three fifths of rhymed pairs are realized by the combination of a noun and a noun, a verb and a verb, and a noun and a verb.

Conclusions
The analysis of Adolf Heyduk's poetry (approximately 110, 000 verse lines) has demonstrated that:

1. The frequency of parts of speech depends on the verse length. The longer the verse is, the more optional parts of speech (i.e. more adjectives and adverbs) it contains.
2. The frequency of parts of speech depends on the literary genre. We can confirm the assumption that the lyric has a higher frequency of nominal group (nouns and adjectives) than the epic, and the epic has a higher frequency of verbal group (verbs and adverbs).
3. The frequencies of parts of speech in the trochee and iamb do not differ.
4. The frequency of parts of speech in rhymed and unrhymed verse does not differ.
5. The frequency of parts of speech in the beginning and end of line differs from their frequency in the whole text.

The deviation from the values in the whole text is associated with both syntax and rhythm.
The beginning of a line is often identical with the beginning of a syntactic unit and it always follows the prosodic boundary. There is a higher concentration of certain parts speech (e.g., conjunctions and pronouns) on the prosodic boundary and, at the beginning of a line, this tendency is reinforced by rhythmical aspects. At the beginning of iambic lines, we observe a greater deviation than at the beginning of trochaic lines. At the beginning of Heyduk's iambic

9 On the increase in the frequency of nouns in the last position of an utterance (the samples are from journalism, technical literature and fiction), or the increase in the frequency of adjectives in the penultimate position—see Průcha (1967).
10 According to SYN2005 (subcorpus fiction) the noun is most frequently preceded by an adjective, the verb by a pronoun—see Bartoň et al. (2009).
11 The same holds true also for Puškin's rhymed and unrhymed verse (Shaw 1993, 17).

verse there is a strong preference for an unstressed word or a stressed monosyllabic word (this accounts for the increasing frequency of those parts of speech which are typically monosyllabic in Heyduk's poetry, i.e. conjunctions, pronouns, particles and adverbs). At the beginning of Heyduk's trochaic verse, on the other hand, there is a strong preference for a stressed word (e.g., verbs, nouns and prepositions, the stress of which always falls on the first syllable). However, the beginning of Heyduk's trochaic lines also shows the increasing frequency of conjunctions, which may be connected with the effort for rhythmic diversification of the beginning of trochaic lines.
The end of a line is often identical with the end of a syntactic unit. At the end of a syntactic unit, the occurrence of certain parts of speech (conjunctions, prepositions) is excluded. At the end of a line, we also observe a higher concentration of nouns and verbs, and thus a semantic accentuation. The increasing frequency of nouns and verbs influences the frequency of parts of speech in the preceding position, i.e. in the penultimate position of a line. There we observe an increasing frequency of adjectives (in the corpus of Czech texts (fiction) SYN2005 a noun is most frequently preceded by an adjective) and pronouns (in the corpus of Czech texts (fiction) SYN2005 a verb is most frequently preceded by a pronoun).
At the end of rhymed verse lines there is, apart from the syntactic requirement, also the requirement of sound repetition, which might cause a slightly increasing frequency of nouns and verbs (related to the use of grammatical rhyme).

Translated by Gabriela Brůhová

Table 1: 6–12syllable verse lines.

number of syllables per line	6	7	8	9	10	11	12
N	36.92	35.67	33.82	33.51	34.08	33.59	33.58
A	9.24	8.76	8.5	8.84	9.32	9.48	10.52
P	13.72	13.32	13.48	13.66	14.21	13.96	14.4
C	0.66	0.67	0.68	0.57	0.73	0.57	0.69
V	19.99	19.58	20.47	20.02	19.03	19.95	18.58
D	5.99	6.73	7.4	7.23	6.54	6.05	5.86
R	8.96	9.87	9.55	10.66	10.15	10.65	10.69
J	3.53	4.26	4.56	4.28	4.62	4.48	4.51
T	0.86	0.96	1.26	1.07	1.19	1.1	1.07
I	0.14	0.17	0.3	0.18	0.12	0.16	0.11
NV/AD	3.74	3.57	3.41	3.33	3.35	3.45	3.18

Table 2: Lyric vs. epic poetry.

	Lyric	Epic	Syn2005 (fiction)	KČV
N	32.93	30.33	24.3	30.19
A	9.96	8.50	8.9	9.91
P	12.98	13.53	14.9	13.69
C	0.44	0.70	1.6	0.64
V	17.18	19.42	21.2	18.38
D	6.67	7.96	8.4	7.75
R	10.65	11.18	9.8	10.02
J	7.42	6.50	8.9	7.13
T	1.29	1.52	1.8	1.39
I	0.47	0.37	0.11	0.89
AN/DV	1.8	1.42	1.12	1.52

Table 3: Trochee vs. iamb; Beginning of a trochaic verse line vs. beginning of an iambic verse line; Unrhymed vs. rhymed verse lines; End of unrhymed verse vs. end of rhymed verse.

	Beginning of a trochaic verse line	Trochee	Beginning of an imabic verse line	Iamb	End of unrhymed verse	Unrhymed verse	End of rhymed verse	Rhymed verse	KČV
N	17.59	31.4	10.31	29.54	48.04	30.26	52.44	30.34	30.19
A	9.41	9.12	1.13	7.61	10.05	8.41	6.69	8.2	9.91
P	6.83	13.15	17.5	13.84	6.88	13.38	3.13	13.6	13.69
C	0.97	0.78	0.41	0.58	0.48	0.87	0.5	0.62	0.64
V	21.58	19.09	10.11	18.86	25.81	19.23	29.76	18.96	18.38
D	9.4	7.09	13.53	8.1	7.15	7.61	6.34	7.67	7.75
R	17.61	11.33	7.3	9.69	0.02	10.04	0.05	10.44	10.02
J	12.88	6.27	32.21	9.33	0.47	7.69	0.27	8.06	7.13
T	2.47	1.4	4.95	1.82	1.03	1.74	0.71	1.62	1.39
I	1.26	0.37	2.55	0.64	0.07	0.77	0.11	0.49	0.89

Table 4: Rhymed pairs (we provide only pairs whose frequency exceeded the limit of 1 per cent).

NN	34.56
VV	17.04
NV	9.83
VN	8.14
DN	4.53
ND	3.38
AN	2.75
NA	2.47
AV	2.21
PN	1.70
AA	1.63
VA	1.54
NP	1.42
DV	1.12

N – Noun
A – Adjective
P – Pronoun
C – Numeral
V – Verb
D – Adverb
R – Preposition
J – Conjunction
T – Particle
I – Interjection

Works Cited

Bartoň, Tomáš et al. 2009. *Statistiky češtiny*. Prague: Nakladatelství lidové noviny and Ústav Českého národního korpusu.

Červenka, Miroslav. 2006. *Kapitoly o českém verši*, edited by Květa Sgallová and Jiří Holý. Prague: Karolinum.

Červenka, Miroslav and Květa Sgallová. 1978. "Český verš." In *Słowiańska metryka porównawcza I. Słownik rytmiczny i sposoby jego wykorzystania*, edited by Zdzisława Kopczyńska and Lucylla Pszczołowska, 45–93. Wrocław: PAN.

Červenka, Miroslav and Květa Sgallová. 1984. "Český verš." In *Słowiańska metryka porównawcza II. Organizacja składniowa*, edited by Zdzisława Kopczyńska and Lucylla Pszczołowska, 11–61. Wrocław: PAN.

Gasparov, Michail and Tatiana Skulačeva. 2004. *Статьи о лингвистике стиха*. Москва: Языки славянской культуры.

Ibrahim, Robert and Petr Plecháč. In press. “Частотность частей речи в чешской поэзии.“ In *Славянский стих.*

Jakobson, Roman. 1979. “Toward a description of Mácha’s verse.” In *Selected Works V*, edited by Stephen Rudy and Martha Taylor, 433–485. The Hague: Mouton.

Mazáčová, Stanislava. 1973. “Sylabická délka věty v českém trocheji a jambu.” In *Semiotyka i struktura tekstu*, edited by Maria Renata Mayenowa, 259–273. Wrocław: PAN.

Plecháč, Petr. 2012. “Miroslav Červenka a generativní model metrické normy českého sylabotónického verše.” *Česká literatura* 60 (3): 398–408.

Průcha, Jan. 1967. “On word-class distribution in Czech utterances.” *Prague Studies in Mathematical Linguistics* 2: 65–76. Prague: Czechoslovak Academy of Sciences.

Pszczołowska, Lucylla. 1965. “Długość wersu a budowa zdania.” In *Poetyka i matematyka*, edited by Maria Renata Mayenowa, 79–96. Warszawa: PIW.

Shaw, J. Thomas. 1993. “Parts of speech in Puškin’s rhymewords and nonrhymed endwords.” *Slavic and East European Journal* 37 (1): 1–22.