Studia Metrica et Poetica 2.1, 2015, 107–118

The Corpus of Czech Verse

Petr Plecháč, Robert Kolár*

Abstract. The article presents the Corpus of Czech Verse (i.e. a lemmatised, phonetically, morphologically, metrically and strophically annotated corpus of Czech poetry) and the online tools and frequency lists that give access to its data. The following online tools are described: Database of Czech metres – the main tool for working with the corpus data, Gunstick – a web application that serves to investigate the frequency of rhyme pairs and their historical development, Hex – an application which enables to search the Corpus of Czech Verse for texts which contain a keyword specified by the user, or to display all keywords found in the group of texts specified by the user, and Euphonometer – application which enables to quantify the degree of non-randomness of sound repetition in any text.

Keywords: Czech poetry, versification, corpus linguistics, verse theory

1. Introduction

At the end of 2013 we completed the first phase of building the Corpus of Czech Verse at the Institute of Czech Literature, Academy of Sciences of the Czech Republic. The corpus currently contains almost 1,700 poetry collections (almost 80,000 poems, over 2.5 million verse lines) primarily from the 19th and early 20th century. All texts have been lemmatised, phonetically transcribed and morphologically, metrically and strophically annotated.[1] In

* Authors' addresses: Petr Plecháč, Institute of Czech Literature, Academy of Sciences of the Czech Republic, Na Florenci 3/1420, 110 00 Praha 1, Czech Republic. E-mail: plechac@ucl.cas.cz; Robert Kolár, Institute of Czech Literature, Academy of Sciences of the Czech Republic, Na Florenci 3/1420, 110 00 Praha 1, Czech Republic. E-mail: kolar@ucl.cas.cz.

[1] Lemmatisation and morphological annotation were carried out by the researchers at the Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University in Prague (Hana Skoumalová, Milena Hnátková, Tomáš Jelínek and Vladimír Petkevič) in cooperation with the researchers at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague (Jan Hajič, Jaroslava Hlaváčová). Phonetic transcription and metric / strophic annotation was carried out using the computer program Květa developed at the Institute for Czech Literature, Academy of Sciences of the Czech Republic (see Ibrahim, Plecháč 2011). At this moment only syllabotonic verse lines are annotated in terms of metrics. Quantitative, syllabic and free verse lines, which also occur in Czech poetry, are currently classified as „undetermined". However, the annotated syllabotonic verse represents more than 88% of all verse lines in the corpus.

doi: dx.doi.org/10.12697/smp.2015.2.1.05

the present paper we will first describe in detail the structure of individual records in the corpus and introduce the freely available online tools that give access to the data contained in the corpus.

2. The structure of records

Each lexical unit (token) in the corpus is assigned phonetic transcription, lemma (the basic dictionary form) and a morphological tag that contains information about various grammatical categories (part of speech, number, case…)[2]. Each verse line is assigned the following attributes, namely (2.1) type of metre, (2.2) length, (2.3) end of a line, (2.4) metrical pattern, (2.5) rhyme, (2.6) commonly used name of metre, (2.7) rhymed, (2.8) stanzaic and (2.9) fixed form:

2.1. Type of metre

A dactyl with anacrusis (amphibrach)
D dactyl
J iamb
N undetermined (free, syllabic, quantitative, accentual, unrecognised verse)
T trochee
X logaoedic
Y logaoedic with anacrusis

2.2. Length

The number of S in the pattern

2.3. End of a line

m masculine (the pattern ends in S)
z feminine (the pattern ends in Sw)
a acatalectic (the pattern ends in Sww)

[2] For detailed description of morphological tags see Hajič 2004: 32–88.

2.4 Metrical pattern (including substitutions, caesuras etc.)

S strong position
w weak position
X undetermined position (free, syllabic, quantitative, accentual, unrecognised verse)

More metrical patterns can correspond to one metre (i.e. metre type + length + end of a line). In such cases one pattern is always considered as basic, e.g.

D4z SwwSwwSwwSw (basic) V ostravské harendě večer se stavil
D4z SwwSwSwwSw Šel starý Magdón z Ostravy domů

Apart from the symbols S, w and X, the metrical pattern may also contain a hyphen. In ghazals the hyphen separates the so called radif (a word or a group of words recurring at the end of a line), which is not included in the characteristics of the metre, e.g.

T5z SwSwSwSwSw-Sw Užívej, když smutek tebe zkruší, hašiš,
T5z SwSwSwSwSw-Sw vzdechy tiší, slzy rázem suší, hašiš
T5m SwSwSwSwS jesti čaroděje mocný prut,
T5z SwSwSwSwSw-Sw který vazby všednosti v ráz zruší, hašiš...

2.5. Rhyme

A numerical index connecting rhymed verse lines. Zero indicates unrhymed lines.

2.6. Commonly used name of metre

alexandrine	J6 with a caesura (constant word boundary) between the sixth and seventh syllable
blank verse	unrhymed J5
hexameter	X6, pattern: S(w)wS(w)wS(w)wS(w)wS(w)wSw
pentameter	X6, pattern: S(w)wS(w)wS(w)S(w)wS(w)wS

Furthermore, besides the usual bibliographical data (author, the name of the collection, the year of publication etc.) the poem is assigned the following

attributes, namely (2.7) rhymed (rhyme scheme), (2.8) stanzaic (stanza scheme) and (2.9) fixed form:

2.7. Rhymed

0 unrhymed poem
1 rhymed poem

A rhymed poem is a poem in which at least 30% of lines have a nonzero rhyme index. A rhyme scheme shows the distribution of rhymes in stanzas. Thus it is determined only in poems marked as both rhymed and stanzaic. A rhymed scheme is traditionally marked, i.e. [a] for the first rhyme in stanza, [b] for the second rhyme in stanza... [x] for an unrhymed verse. A rhyme scheme is recorded only if it occurs in a poem at least twice. If there are more than three different schemes in a poem, it remains undetermined.

2.8. Stanzaic

0 non-stanzaic poem
1 stanzaic poem

A poem is marked as stanzaic if it consists of sections containing m, or n lines ($2 \leq m, n \leq 14$) with a scheme (1) m.m..., (2) m.n.m.n..., or (3) m.m...n.n... (m.m...n.n...)...

Stanza scheme indicates the distribution of individual metres in the stanza, e.g. the scheme [abab] can (among others) correspond to the following combinations:

T4z T4m T4z T4m . T4z T4m T4z T4m ...
or
T4z D4z T4z D4z . T4z D4z T4z D4z...

Stanza scheme is determined only in poems indicated as stanzaic. Stanza scheme is recorded only if it occurs at least twice.

2.9. Fixed form

At present, the following fixed forms are recognised: Alcaic strophe, arte mayor, Asclepiad IV, Burns stanza, elegiac couplet, ghazal, heroic couplet, huitain, rhyme royal, qaṣīda, limerick, madrigal, Onegin stanza, ritornello, rondel, rondeau, Sapphic stanza, sestina, Sicillian octave, Italian sonnet, English sonnet, Spenserian stanza, ottava rima, terza rima. (For more details, see http://www.versologie.cz/en/kcv_znacky.html).

3. Online tools

The online tools and frequency lists that are continuously being developed give access to the data contained in the Corpus of Czech Verse. At present, the following tools are available: (3.1) Database of Czech metres, (3.2) Gunstick, (3.3) Hex, and (3.4) Euphonometer. All applications are available in Czech, English and Russian translations at <www.versologie.cz>.

3.1. Database of Czech metres

The Database of Czech metres is the main tool for working with the corpus data. The user can both search for and statistically evaluate data on the basis of their own and/or default filters, and browse through individual records in the database.

The filters (which can be freely combined) include: type of metre (2.1), length (2.2), end of a line (2.3), metrical pattern (2.4), commonly used name of metre (2.6), rhymed/unrhymed, rhyme scheme (2.7), stanzaic/non-stanzaic, stanza scheme (2.8), fixed form (2.9). The results of such query are interactive line charts and pie charts displaying the distribution of poems/verse lines complying with the specific parameters, and a tree structure which provide detailed information about the individual poems.

3.2. Gunstick – database of Czech rhymes

Gunstick is a web application that serves to investigate the frequency of rhyme pairs and their historical development. When using the application the user enters a word (token) which will be searched for all rhyme pairs attested in the corpus before 1920. The search can be restricted to a specified author, to a specified time span or a specified end of a line (masculine, feminine, acatalectic, undetermined).

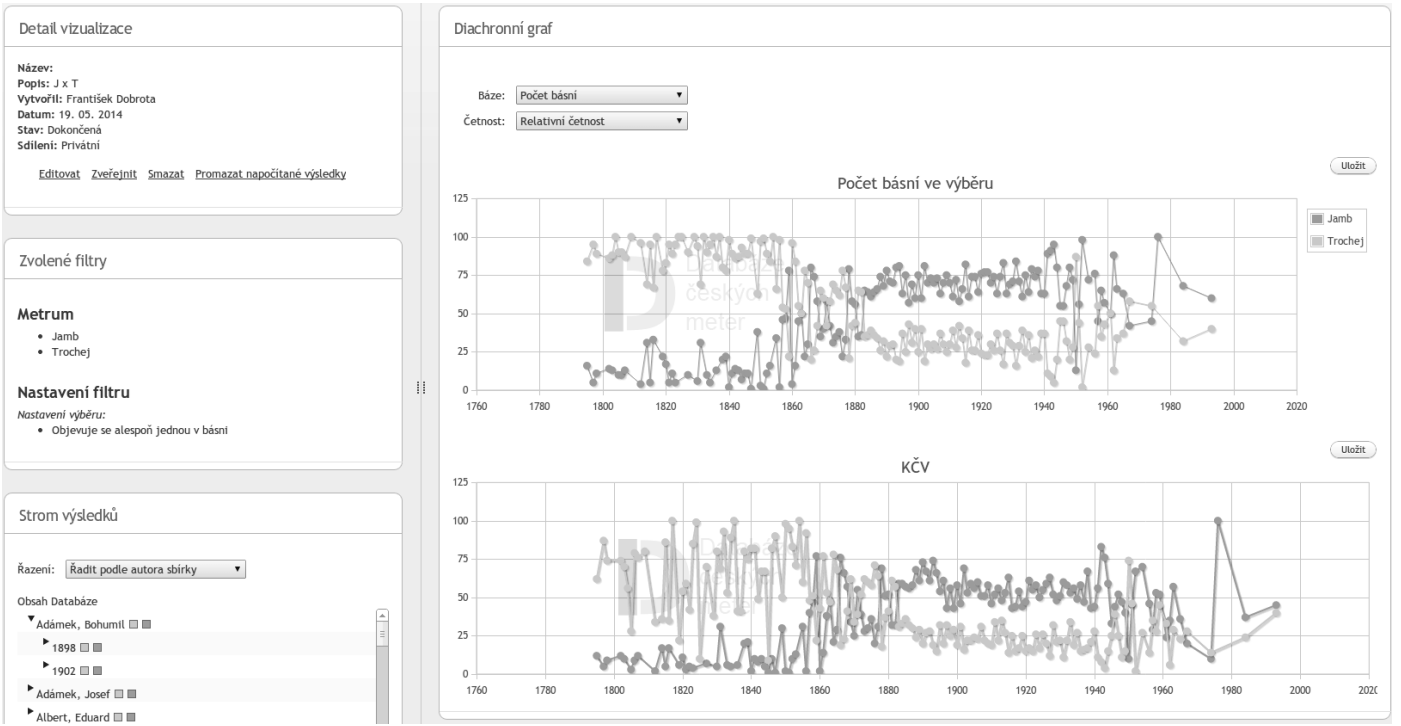


Chart 1: Database of Czech metres. Distribution of iambic and trochaic lines (relative frequency measured by number of poems).

After entering the query a pie chart is displayed illustrating the frequency of occurrences of individual rhymes with the searched word. The left-click may be used to select a sector and thus display the selected data in the area chart, which shows the number of occurrences of the selected rhyme pair for each year, and in the table at the bottom of the screen, which displays among others the full text of both rhyming verse lines and a reference to the full text of the particular collection.

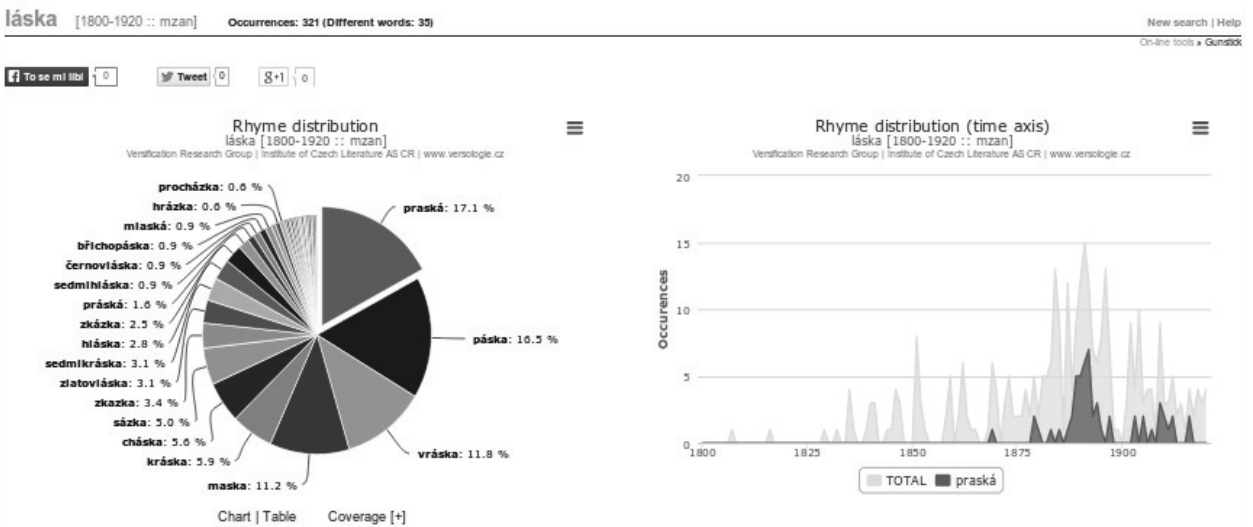
Clicking the button "coverage [+]" below the pie chart enables to display charts illustrating the coverage and data volume. The chart "Data volume" shows the number of occurrences of all rhymes (i.e. all occurrences in the database) for each year (when specifying the filter "author" it shows the number of occurrences complying with the given conditions). The chart "Coverage" indicates the percentage of rhymes with the searched word within all occurrences in each year.

3.3. Hex – key words in Czech poetry

The Hex application enables to search the Corpus of Czech Verse for texts which contain a keyword specified by the user, or to display all keywords found in the group of texts specified by the user[3]. In both cases the user can narrow down the selection by using the filter "the name of the author" and defining the time span. In addition, when searching a specified group of texts the user can use the filters "name of the collection" and "name of the poem". Keywords are those lemmata whose frequency in the given poem is significantly higher than the frequency in the whole corpus. The statistical significance is verified by the χ^2 (with Yates's correction) and log-likelihood tests. The user can specify whether the tests will be performed at the significance level $\alpha = 0.001$ (i.e. the 0.1% risk that the lemma whose higher frequency in the poem is only a coincidence will be incorrectly marked as a keyword), or $\alpha = 0.01$ (i.e. 1% risk). Along with this, the user can specify which parts of speech should be excluded from the analysis (by default only nouns, adjectives and verbs are allowed) and determine the minimum number of occurrences of a lemma in the poem required for its inclusion among the keywords.

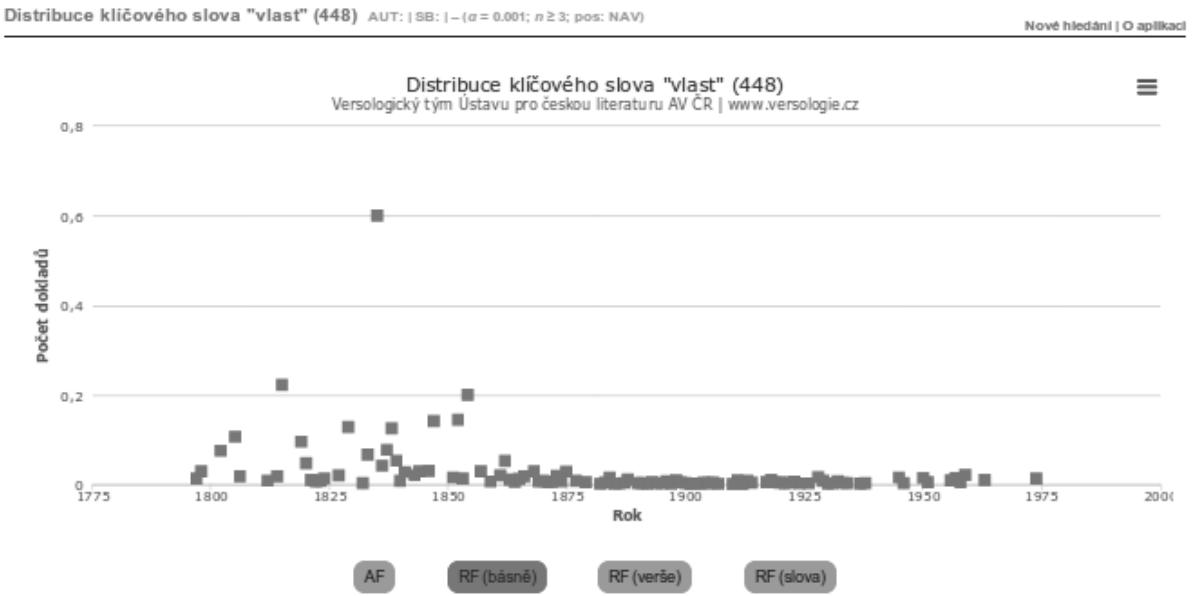
When searching for a specific key word, after entering a query an interactive chart is displayed showing the frequency of occurrences in each year,

[3] For analysis of keywords in your own texts we recommend the application KWords (Cvrček, Vondřička 2013) developed by the Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague, that we drew inspiration from when developing Hex.



Rhyme	Line 1	Line 2	Author	Poem	Collection	Year	End of a line
praská	Těm dívkám zpívejte, jež otrávila láska	jež s vlasem spuštěným a v hedbávu, jenž praská,	Borecký, Jaromír	Slavíkům hřbitova.	Rosa mystica	1892	z
praská	buď velebena věčná Boha Láska,	nechť svět se řítí, země hyne, praská,	Bouška, Sigismund	Ó kniho Bohem psaná...	Pietas	1897	z
praská	Opatrnost přerůstá vám láska.	vidím, jak to syčí, jak to praská.	Dyk, Viktor	XI.	Zápas Jiřího Macků	1916	z
praská	nadšení i sama láska –	– půda kolem hřmí a praská;	Geisslová, Irma	NA VODÁCH.	Z Podkrkonoší	1889	z
praská	což je rekům ženy láska!...	nechať žíly, mozek praská,	Geisslová, Irma	HRUBOSKALSKÁ PANNA.	Z Podkrkonoší	1889	z

Chart 2: Gunstick – database of Czech rhymes. Rhymes of the word "láska".



Doklady klíčového slova "vlast" (448) AUT: | SB: | – (α = 0.001; n ≥ 3; pos: NAV)

Autor	Báseň	Sbírka	Rok	AF	RF	φ
Balcárek, Aleš	Pád říše Velko-Moravské. »	Pozůstalé básně »	1862	9	12064	0.0033
Balcárek, Aleš	Moravanův sláva. »	Pozůstalé básně »	1862	4	9523	0.0018
Balcárek, Aleš	Slávy vesna. »	Pozůstalé básně »	1862	8	18735	0.004
Balcárek, Aleš	Píseň Čechoslovanův. »	Pozůstalé básně »	1862	3	26315	0.0026

Chart 3: Hex – keywords in Czech poetry. Distribution of the keyword "vlast" (relative frequency measured by the number of poems).

either the absolute frequency (the total number of occurrences in each year) or relative frequency with respect to (a) the number of poems (i.e. absolute frequency divided by the number of all poems published in the given year), (b) the number of verse lines (i.e. absolute frequency divided by the number of all verse lines contained in the poems published in the given year), (c) the number of words (i.e. absolute frequency divided by the number of all words contained in the poems published in the given year). In addition, a table is displayed containing, among other things, the name of the poem in which the keyword was found. At the same time, the name of the poem serves as a link to the list of all key words found in the poem according to the parameters entered, and the name of the collection which the poem comes from serves as a link to the full text of the collection.

When searching a specific group of texts the user can choose whether the output should be a list of poems with keywords, or a frequency list of the selection.

3.4. Euphonometer

The application Euphonometer enables to quantify the degree of non-randomness of sound repetition in any text (the so-called euphonic coefficient). The application draws upon the method based on the binomic test, which was proposed by Gabriel Altmann (Altmann 1966a; 1966b; Čech et al. 2011) and later slightly modified (Plecháč, Říha 2014). The results of the analysis are values of the euphonic coefficient of each line of the searched text and the total (average) euphonic coefficient which can be compared with the values counted for each poem in the corpus.

3.5. Frequency lists of Czech poetry

Frequency lists of Czech poetry contain information about the frequency of words in the works of poetry included in the Corpus of Czech Verse. The lists provide information about both the frequency of lemmata and frequency of word forms (tokens), not only in the individual poetry collections but also in the author's subcorpora and the entire Corpus of Czech Verse.

The data in the lists are classified as follows:

column 1: rank
column 2: lemma/token

column 3: part of speech[4]
column 4: the absolute frequency of lemma/token
column 5: the relative frequency of lemma/token[5]

Each list is published in two formats: (1) xls (Microsoft Excel, OpenOffice Calc, LibreOffice Calc) and (2) txt with UTF-8 encoding, where individual columns are separated by a tabulator (the latter is convenient for further processing). The lists can be downloaded as compressed archive containing the frequency list of the entire author's subcorpus (00_dilo) and frequency lists of individual collections of poems of an author ([year of publication] _ [name of collection]).[6]

4. Conclusion

The online tools and frequency lists that have been presented in this paper are of course limited in their functions and cannot make use of the full potential of the Corpus of Czech Verse. One can easily imagine that mere frequency lists may not be sufficient for every user, and that their research project may require for example not only information about the frequency of lemmata in a given author's work, but rather more specific frequency lists generated for each metre used by this author separately. Other users could prefer – in order to be methodologically coherent – the analysis of thematic concentration to the keyword analysis (cf. Popescu 2007; Popescu, Altmann 2011).

One possibility could be a direct online access to the entire database via SQL queries. Thus, the user could enter any query without being limited by the functions of the tools. However, only a small number of potential users would know the query language, and the results of such queries would in most cases have to be further processed using a statistical software. The optimal approach therefore appears to be a compromise – to build the interface for direct SQL

---

[4] The part of speech is indicated by the first position of a morphological tag (A – adjective, C – numeral, D – adverb, I – interjection, J – conjunction, N – noun, P – pronoun, R – preposition, T – particle, V – verb, X – unknown, indeterminable part of speech, see Hajič 2004: 32–88). This allows us to distinguish between homonyms like "bez" (noun/preposition) and to further filter out the data obtained from the lists (e.g. to evaluate the frequency of nouns only).

[5] Given in ppm (10 000 ppm ~ 1 %) and rounded to whole numbers.

[6] For information about the frequency of lemmata/tokens in prose, which the data included in the frequency lists of Czech poetry can be compared with, see Křen 2010.

queries (see the new application Babel at http://www.versologie.cz/babel/) as well to continue with the development and update of the tools according to users' requirements.[7]

References

Altmann, Gabriel 1966a. The Measurement of Euphony. In: Levý, Jiří (ed.), *Teorie verše* I: *Sborník brněnské versologické konference 13.–16. května 1964* (Spisy Filozofické fakulty Univerzity J. E. Purkyně v Brně 107). Brno: Universita J. E. Purkyně, 263–264.

Altmann, Gabriel 1966b. Binomial Index of Euphony for Indonesian Poetry. In: *Asian and African Studies* 2, 62–67.

Cvrček, Václav; Vondřička, Pavel 2013. *KWords*. Praha: Ústav Českého národního korpusu FF UK. URL: http://kwords.korpus.cz (accessed June 5, 2014).

Čech, Radek; Popescu, Ioan-Iovitz; Altmann, Gabriel 2011. Euphony in Slovak Lyric Poetry. In: *Glottometrics* 22, 5–16.

Hajič, Jan 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Praha: Karolinum.

Ibrahim, Robert; Plecháč, Petr 2011. Toward Automatic Analysis of Czech Verse. In: Scherr, Barry et al. (eds.), *Formal Methods in Poetics*. Lüdenscheid: RAM, 295–305.

Křen, Michal 2010. *Srovnávací frekvenční seznamy*. Praha: Ústav Českého národního korpusu FF UK. URL: http://ucnk.ff.cuni.cz/srovnani10.php (accessed June 5, 2014).

Plecháč, Petr; Říha, Jakub 2014. Measuring Euphony. In: Vekshin, Georgy (ed.), *Metodologija i praktika russkogo formalizma* (Brikovskij sbornik 2). Moskva: Azbukovnik, 194–199.

Popescu, Ioan-Iovitz 2007. Text Ranking by the Weight of Highly Frequent Words. In: Grzybek, Peter, Köhler, Reinhard (eds.), *Exact Methods in the Study of Language and Text*. Berlin, New York: Mouton de Gruyter, 557–567.

Popescu, Ioan-Iovitz; Altmann, Gabriel 2011. Thematic Concentration in Texts. In: Kelih, Emmerih et al. (eds.), *Issues in Quantitative Linguistics* 2. Lüdenscheid: RAM, 110–116.

---

[7] The paper was translated by Gabriela Brůhová. This paper and its translation were supported by Czech Science Foundation (P406/11/1825) and by the long-term conceptual development of a research institution 68378068.