CLOSED CONCEPTUAL DOMAINS IN HUNGARIAN CANONICAL POETRY: A CORPUS LINGUISTIC APPROACH

Péter Horváth

ELTE Eötvös Loránd University
horvath.peeteer@gmail.com
https://orcid.org/0000-0002-3517-5623

Abstract

The paper shows how the distribution of different concepts within a closed conceptual domain can be used for quantitative analysis of poetic corpora. The method is presented in three case studies based on the complete poems of 23 canonical Hungarian poets. The first case study analyzes the frequency of the concept NIGHT, which is part of the conceptual domain TIMES OF DAY. The second case study analyzes the frequencies of the four season concepts, and the third case study assesses the frequencies of color concepts. The change in the frequencies of the concepts analyzed seems to reflect the well-known poetic change in Hungarian poetry at the beginning of the 20th century. The paper also demonstrates that in canonical Hungarian authors' poetry, there is a strong positive correlation between the frequencies of the three conceptual domains, which may mean that referring to these domains is part of the same poetic toolkit aiming to highlight sensory impressions of the physical setting of poems. Finally, the paper shows which concepts from the three different conceptual domains co-occur in the same poems with a higher mutual information score.

Keywords: closed conceptual domain, poetry, corpus linguistics, correlation, MI-score, time of day, season, color

1. Introduction

The paper presents a quantitative, corpus linguistic approach of Hungarian poetry, based on the distribution of concepts of closed conceptual domains. The method can contribute to distant reading techniques applied to poetry (on distant reading see Moretti 2000, 2005). Until now, relatively little quantitative corpus linguistic research has been done on Hungarian poetry. These studies usually focus either on the poems of a single Hungarian poet (see Jékel–Papp 1974; Jékel–Szuromi 1980; Lesi 2008; Labádi 2018) or on a few poems by several poets (see Fónagy 1959; Zsilka 1974; Simon 2020). Technological advances in recent years have made it possible to study corpora containing all the poems of many Hungarian poets. The corpus analysis presented below aims to demonstrate the new possibilities inherent in the quantitative investigation of Hungarian poetry, by using a much larger corpus than previous studies.

I apply the term *closed conceptual domain* for well-delimited conceptual fields consisting of a finite number of concepts. For instance, the domain of COLORS has a finite number of concepts expressed by the color terms. By measuring the occurrences of the concepts of a closed conceptual domain, we can gain frequency data about the appearance of the investigated closed conceptual domain in a corpus of poetry and we can compare different subcorpora with each other on the basis of these frequency data. In corpus linguistics, the quantitative analysis of the occurrence of concepts is not new. Rayson (2008), for example, used the keyword method to analyze semantic domains in texts (see also McIntyre 2013). The distribution of the concepts of a closed conceptual domain in a set of poems can direct our attention to features which cannot be revealed by methods

DOI: 10.54888/slh.2021.33.66.78 ISSN 2732-1142

of close reading. These features are typically related to general features of the physical, social and mental spheres of the constructed fictive worlds of poems (cf. Tátrai 2011: 171–189, 2015a)[1].

The paper presents the method by means of frequency analysis of three closed conceptual domains in the complete poems of 23 Hungarian canonical poets. The three closed conceptual domains are the following: TIMES OF DAY, SEASONS and COLORS. These domains pertain to the physical sphere of the poems, they contribute to the constitution of the physical setting of the narrated scene or the fictive lyrical speech situation in which a fictive speaker directs the attention of a fictive addressee to a fictive scene (on the apostrophic nature of lyrical discourses, see Culler 1981: 135–154; Tátrai 2015b). By the investigation of these conceptual domains, we can learn about the physical setting of a given set of poems, and by comparing different sets of poems on the basis of these domains, typical and atypical patterns can be identified.

Section 2 presents the main properties of the research corpus, and the tools of frequency analyses. Section 3 shows the frequency data of the concept NIGHT, which is related to the closed conceptual domain TIMES OF DAY, section 4 presents the frequency data of the domain SEASONS, and section 5 shows the frequency data of the domain COLORS. Section 6 demonstrates the existence of a strong positive correlation between the frequencies of the three conceptual domains. Section 7 presents the results of the co-occurrence analysis of the concepts from different conceptual domains. Section 8 briefly summarizes the research with some final remarks.

2. Corpus and tools

The research corpus for the frequency analyses presented here consists of the complete poems of 23 Hungarian poets. This corpus was extracted from the ELTE Poetry Corpus,[2] which is a database consisting of the complete poems of 49 canonical Hungarian poets. I used the poems of those poets who have more than 100 poems in the ELTE Poetry Corpus and who were not born before Csokonai. Besides lyrical poems, the research corpus also included the authors' longer narrative poems. The format of the corpus is TEI XML (TEI Consortium 2019). The TEI XML files contain not only the text of the poems but among other types of annotations, the lemma, the part of speech and the morphosyntactic features of words as well. These grammatical annotations have been created by the program e-magyar, an NLP tool for the automatic analysis of the grammatical features of Hungarian texts (Váradi et al. 2018; Indig et al. 2019). The research corpus containing the texts of 23 Hungarian poets has 11,262 poems and 2,120,996 words. Table 1 presents the 23 authors with their dates of birth and death, and the number of poems and words, respectively. The authors are shown in birth order. The subsequent tables also sort the frequency data in this order.

For the frequency analysis, the programming language Python was used, with the lxml library,[3] which makes the query of XML files simple. The gained frequency data were loaded into a spreadsheet program, where further frequency analyses were accomplished. The search terms referring to the concepts analyzed were collected manually, on the basis of thesauruses and my own research intuition.

Table 1. Content of the research corpus

Author	Birth and death	Number of poems	Number of words
Csokonai Vitéz, Mihály	1773–1805	394	125,421
Berzsenyi, Dániel	1776–1836	136	20,763
Kisfaludy, Károly	1788–1830	123	24,001
Kölcsey, Ferenc	1790–1838	149	18,833
Vörösmarty, Mihály	1800 – 1855	662	210,244

[1] Tátrai (2011, 2015a) applied his tripartite model to narrative fiction, but it can be applied to lyrical fiction as well.

[2] https://github.com/ELTE-DH/poetry-corpus

[3] https://lxml.de

Arany, János	1817–1882	417	276,092
Tompa, Mihály	1817–1868	491	172,565
Petőfi, Sándor	1823–1849	839	148,466
Madách, Imre	1823–1864	319	67,287
Gyulai, Pál	1826–1909	156	32,525
Vajda, János	1827–1897	199	95,260
Reviczky, Gyula	1855–1889	335	50,485
Komjáthy, Jenő	1858–1895	246	47,908
Ady, Endre	1877–1919	1116	121,526
Kaffka, Margit	1880–1918	102	22,329
Somlyó, Zoltán	1882–1937	379	66,672
Juhász, Gyula	1883–1937	1278	113,222
Babits, Mihály	1883–1941	514	95,758
Kosztolányi, Dezső	1885–1936	630	85,026
Tóth, Árpád	1886–1928	451	60,147
Reményik, Sándor	1890–1941	667	82,309
József, Attila	1905–1937	599	64,717
Dsida, Jenő	1907–1938	1060	119,440
All	1773–1941	11262	2,120,996

3. The conceptual domain of TIMES OF DAY

The first closed conceptual domain investigated is the domain of TIMES OF DAY. We divide days into time spans. The two largest time spans of a day are daytime and night, which are based on the different physical settings related to the position of the sun. In our conceptual system, the default time of day is DAYTIME, since usually this is the time of our active life. For instance, when somebody tells a story that happened to her and does not specify the time of day, we tend to think that it took place in daytime, except in the case of some special activities typically related to night, such as drinking beer in a pub or dancing in a club. This is also true for the reception of poems. When the fictive lyrical speaker does not specify the time of day, we usually do not think that the time of the fictive speech situation or the narrated events is night. We are led to think so only when the fictive speaker makes it explicit by means of linguistic expressions. Signifying that the time of the fictive lyrical speech situation or the narrated scene or events is night is a deviation from the default parameter setting for time. An interesting question is the extent to which different authors have deviated from this default.

The first case study analyzes the frequencies of those poems in the research corpus in which the concept of NIGHT comes up, in other words, where there is a deviation from the default time setting. The analysis used a list of synonymous lemmas as search terms referring to the time of NIGHT. Since the corpus specifies the lemma of each word, it was possible to query the lemmas directly, without spending a considerable amount of time by collecting all of the word forms of a lemma, which is a typical problem of using unlemmatized Hungarian corpora. The search terms are shown in (1) with the English translations. As the noun and the adjective forms of the concept NIGHT are expressed differently in Hungarian, the part of speech labels are also indicated. In addition to the standard forms, archaic forms and spelling variants have also been added to the search terms. I have also used verbs, participles/participial adjectives, and nouns derived by the *-ás/-és* suffix from the verbs with the Hungarian prefix *be* and *rá* as search terms (e.g. *beesteledik* [night is falling], *ráesteledik* [benight]). For the sake of clarity, these prefixed lemmas are not included in the list in (1) below.

(1) *este, estve, est* (evening NOUN), *éjszaka, éjjszaka, éjjel, éjel, éj, éjj* (night NOUN), *éjfél, éjjfél* (midnight NOUN), *alkony, alkonyat, alkonyulat, szürkület* (twilight NOUN), *estefelé, estefele* (around evening ADV), *esti, estvei* (evening ADJ), *éjszakai, éjjszakai, éjjeli, éjeli, éji, éjji* (night ADJ), *éjféli, éjjféli* (midnight ADJ), *alkonyi, alkonyati, alkonyulati, szürkületi* (twilight ADJ), *esteledik, estveledik, alkonyodik, alkonyul, sötétedik* (night is falling VERB), *esteledő, estveledő, alkonyodó, alkonyuló* (darkening ADJ: present participle), *esteledett, estveledett, alkonyodott, alkonyult* (darkened ADJ: past participle), *esteledés, estveledés, alkonyodás, alkonyulás, sötétedés* (nightfall NOUN)

As (1) shows, the method does not take grammatical categories into account. In the list, there are nouns and (participial) adjectives as well as verbs. The goal was to collect (nearly) all of the Hungarian lemmas referring directly to the concept of NIGHT, regardless of grammatical category. I wrote a simple Python script that went through the XML files of the poems and checked if the poems contained any of the lemmas in the list. In the case of each author, the script's output was the number of poems containing at least one lemma from the list. These frequency numbers indicate the measure of deviation of a given author's poetry from the default time setting. The resulting frequency data are presented in Table 2.

Table 2. Frequencies of poems referring to NIGHT

Author	**Number of poems**	**Poems with NIGHT**	**NIGHT %**	**Rank**
Csokonai	394	86	21.8%	17
Berzsenyi	136	29	21.3%	19
Kisfaludy	123	47	38.2%	8
Kölcsey	149	58	38.9%	5
Vörösmarty	662	133	20.1%	23
Arany	417	118	28.3%	13
Tompa	491	270	55.0%	1
Petőfi	839	210	25.0%	15
Madách	319	86	27.0%	14
Gyulai	156	50	32.1%	11
Vajda	199	71	35.7%	10
Reviczky	335	69	20.6%	21
Komjáthy	246	50	20.3%	22
Ady	1116	236	21.1%	20
Kaffka	102	45	44.1%	3
Somlyó	379	160	42.2%	4
Juhász	1278	481	37.6%	9
Babits	514	164	31.9%	12
Kosztolányi	630	305	48.4%	2
Tóth	451	174	38.6%	6
Reményik	667	162	24.3%	16
József	599	130	21.7%	18
Dsida	1060	408	38.5%	7
Mean			31.9	
Median			31.9	

The number of poems in which the concept of NIGHT appears is shown in column 3. The proportions of these poems to the number of all poems are given in column 4. The last column shows the ranks of the authors, based on the proportions. Proportionally, Tompa has the most poems containing the concept of NIGHT. In his case, 55% of the poems contains at least one of the lemmas listed in (1). In second place, we find Kosztolányi, with 48.4% of his poems referring to NIGHT. Proportionally the fewest poems containing expressions of NIGHT are found in the case of Vörösmarty, where only 20.1% of poems are categorized as such. The mean and median of the proportions are also indicated in the table, which are 31.9%. The table shows that the number of poems referring to NIGHT is proportionally higher for poets of the early 20th century. The proportions of these poems in the case of Kaffka, Somlyó, Juhász, Kosztolányi, Tóth and Dsida are higher than the mean and median. Interestingly, Ady, who was also a contributor to the modernist Hungarian literary journal *Nyugat*, wrote proportionally far fewer poems referring to NIGHT than the authors mentioned above. In this respect, his poetry is much more similar to Reviczky's and Komjáthy's poetry from the end of the 19th century.

4. The conceptual domain of SEASONS

The second short case study investigates the appearance of the conceptual domain of SEASONS. In this case, there is no default setting based on real life experiences. We cannot say that one season plays a more prominent role in human life than the others. However, it is possible that in a given poetic tradition, referring to one season is more typical than another. It is also possible that there is a typical frequency order of seasons in a poetic tradition. If there is a kind of order, usually there is deviation from that order as well. It is an interesting question which authors deviate from the typical patterns in the use of season concepts, which may shed light on certain idiosyncratic aspects of these authors' poetry. The frequency analysis used the lemmas in (2) referring to the four seasons. The part of speech of each Hungarian lemma is indicated in brackets. As the search terms in list (2) show, the meaning of "spring is coming" can be expressed with one word in Hungarian, but in the case of summer, autumn and winter, there is no one-word equivalent of this kind of meaning, it can only be expressed analytically, as in English.

(2)

I. SPRING: *tavasz* (NOUN), *tavaszi* (ADJ), *tavaszodik* (spring comes VERB), *tavaszodó (*ADJ: present participle)*, tavaszodott* (ADJ: past participle), *tavaszodás* (turning into spring NOUN), *kitavaszodik* (spring comes VERB), *kitavaszodó* (ADJ: present participle)*, kitavaszodott* (ADJ: past participle)*, kitavaszodás* (turning into spring NOUN), *kikelet* (NOUN), *kikeleti* (ADJ)

II. SUMMER: *nyár* (NOUN), *nyári* (ADJ)

III. AUTUMN: *ősz* (NOUN), *őszi* (ADJ)

IV. WINTER: *tél* (NOUN), *téli* (ADJ)

The Python script went through the poems of each author and for each season it produced the number of poems containing at least one expression for that season. Naturally, a poem can belong to several season categories if it contains the expressions of two or more different seasons. The resulting frequency data are presented in Table 3.

Table 3. Frequencies of poems referring to NIGHT

Author	**Seasons %**	**1.**		**2.**		**3.**		**4.**	
		season	**%**	**season**	**%**	**season**	**%**	**season**	**%**
Csokonai	18.0%	spring	9.9%	summer	7.6%	winter	6.6%	autumn	3.3%
Berzsenyi	13.2%	spring	9.6%	autumn	3.7%	winter	2.2%	summer	1.5%
Kisfaludy	20.3%	spring	13.8%	autumn	7.3%	winter	4.9%	summer	2.4%
Kölcsey	16.8%	summer	8.1%	spring	6.7%	winter	4.7%	autumn	2.0%

Vörösmarty	13.7%	spring	8.8%	winter	7.1%	autumn	4.1%	summer	2.9%
Arany	21.1%	winter	11.0%	summer	10.1%	spring	8.6%	autumn	8.6%
Tompa	36.3%	spring	20.0%	winter	13.8%	autumn	13.2%	summer	10.8%
Petőfi	17.2%	spring	10.0%	autumn	6.9%	winter	6.3%	summer	3.6%
Madách	28.2%	spring	16.3%	autumn	9.4%	winter	8.5%	summer	7.8%
Gyulai	25.0%	spring	18.6%	autumn	10.3%	summer	6.4%	winter	5.1%
Vajda	26.6%	summer	15.6%	spring	10.1%	winter	8.5%	autumn	4.5%
Reviczky	26.3%	spring	14.6%	summer	9.9%	autumn	7.5%	winter	4.2%
Komjáthy	15.0%	summer	8.9%	spring	5.3%	autumn	2.0%	winter	1.6%
Ady	17.0%	spring	6.3%	autumn	5.4%	summer	5.3%	winter	5.1%
Kaffka	20.6%	summer	9.8%	spring	8.8%	winter	7.8%	autumn	2.9%
Somlyó	30.1%	winter	12.1%	autumn	10.6%	summer	10.0%	spring	6.6%
Juhász	31.1%	spring	15.0%	autumn	11.4%	summer	8.5%	winter	5.9%
Babits	23.9%	winter	10.7%	spring	10.3%	summer	8.4%	autumn	6.6%
Kosztolányi	24.9%	summer	10.0%	autumn	9.7%	winter	9.7%	spring	4.1%
Tóth	22.2%	spring	8.6%	autumn	8.0%	summer	6.2%	winter	4.0%
Reményik	20.4%	autumn	10.9%	spring	6.6%	summer	5.2%	winter	5.2%
József	14.5%	summer	5.7%	winter	5.0%	autumn	4.5%	spring	3.3%
Dsida	27.5%	autumn	12.7%	spring	10.2%	summer	7.7%	winter	5.2%
Mean	22.2%								
Median	21.1%								
I.		spring (12)		autumn (10)		winter (9)		winter (8)	
II.		summer (6)		spring (7)		summer (8)		summer (6) autumn (6)	
III.		winter (3)		winter (3) summer (3)		autumn (5)		spring (3)	
IV.		autumn (2)				spring (1)			

The second column of Table 3 shows the proportions of poems referring to any of the four seasons. The central values are also indicated. The mean is 22.2% and the median is 21.1%. The frequency values show that authors from the end of the 18th century and from the first half of the 19th century (Csokonai, Berzsenyi, Kisfaludy, Kölcsey, Vörösmarty) are below the mean and median. They wrote proportionally fewer poems containing expressions of seasons than the later authors. The lowest frequency values are found in the case of Vörösmarty and Berzsenyi. The highest proportion of poems referring to seasons is found among the works of Tompa. Interestingly, Tompa also wrote proportionally the most poems referring to NIGHT.

Table 3 also shows, in descending order for each poet, the proportions of poems referring to SPRING, SUMMER, AUTUMN and WINTER. For instance, in the case of Csokonai, 9.9% of his poems refer to SPRING, 7.6% of his poems refer to SUMMER, 6.6% of his poems refer to WINTER, and 3.3% of his poems contain reference to AUTUMN. The last four rows indicate the rank order of the four seasons in the four frequency positions. From these fields, it can be seen that 12 authors used SPRING and 6 authors used SUMMER as the most frequent season in their poetry. These numbers show that warm seasons were more popular than cold seasons. Five authors deviate from this pattern: Arany, Somlyó, Babits, Reményik and Dsida. In their poetry, one of the two cold seasons, WINTER or AUTUMN, is the most frequent. It is striking that four of them are from the 20th century. It is also worth noting that until the end of the 19th century, SPRING was the most popular season in the case of the poets analyzed, after which the pattern becomes much more heterogeneous. In the second frequency position, AUTUMN is at the top, which means that AUTUMN is the second most frequent season in the case of most poets. The most typical frequency order of seasons is SPRING, AUTUMN, SUMMER|WINTER.

The pipe indicates that the order of SUMMER and WINTER is reversible, they can be either in the third or in the fourth position. There are 8 authors who follow this pattern.

5. The conceptual domain of COLORS

The last closed conceptual domain analyzed here is the domain of COLORS. The questions are similar to those for NIGHT and SEASONS. Which authors used color concepts more? Are there more typical, more widely used colors in Hungarian canonical poetry? What is the distribution of the most popular colors over time? The investigated lemmas referring to colors are shown in (3). In addition to standard forms, I have also listed archaic forms as well as spelling variants. Contrary to the expressions of NIGHT and SEASONS, in this case there are color terms which can be nouns and adjectives as well. The lemmas searched for included the verbs in (3) with the verbal prefixes *el, meg, be, ki, át* and *össze* (e.g. *megfeketedik, átfeketül*), the present and past participle forms of the prefixed and non-prefixed verbs (e.g. *feketedő, megfeketedő, feketedett, megfeketedett*), as well as nouns derived from the verb stem (with and without the prefixes) by the *-ás/-és* suffix (e.g. *feketülés, megfeketülés, elfeketedés*). These lemmas are not shown in list (3) due to lack of space.

(3)

I. BLACK: *fekete* (black NOUN ADJ), *feketeség* (blackness NOUN), *feketés* (blackish ADJ), *feketésség* (blackishness NOUN), *feketül, feketedik* (turn black VERB), *feketél, feketéll, feketéllik* (look black VERB), *feketít, feketit* (make it black VERB)

II. WHITE: *fehér* (white NOUN ADJ), *fehérség* (whiteness NOUN), *fehéres* (whitish ADJ), *fehéresség* (whitishness NOUN), *fehérül, fehéredik* (turn white VERB), *fehérel, fehérell, fehérellik, fehérlik* (look white VERB), *fehérít, fehérit* (make it white VERB)

III. GRAY: *szürke* (gray NOUN ADJ), *szürkeség* (grayness NOUN), *szürkés* (grayish ADJ), *szürkésség* (grayishness NOUN), *szürkül, szürkülödik* (turn gray VERB), *szürkél, szürkéll, szürkéllik* (look gray VERB), *szürkít, szürkit (make it gray VERB)*

IV. RED: *piros* (red NOUN ADJ), *pirosság* (redness NOUN), *pirosas* (reddish ADJ), *pirosasság* (reddishness NOUN), *pirosul, pirosodik* (turn red VERB), *pirosol, pirosoll, pirosollik, piroslik* (look red VERB), *pirosít, pirosit* (make it red VERB), *vörös* (red NOUN ADJ), *vörösség* (redness NOUN), *vöröses* (reddish ADJ), *vörösesség* (reddishness NOUN), *vörösül, vörösödik* (turn red VERB), *vörösöl, vörösöll, vörösöllik* (look red VERB), *vörösít, vörösit* (make it red VERB), *vörösedik* (turn red VERB), *vörösel, vörösell, vörösellik, vöröslik* (look red VERB), *vöres* (red NOUN ADJ), *vöresség* (redness NOUN), *vöreses* (reddish ADJ), *vöresesség* (reddishness NOUN), *vöresül, vöresedik* (look red VERB), *vöresel, vöresell, vöresellik, vöreslik* (look red VERB), vöresít, vöresit (make it red VERB), *veres* (red NOUN ADJ), *veresség* (redness NOUN), *vereses* (reddish ADJ), *veresesség* (reddishness NOUN), *veresül, veresedik* (turn red VERB), *veresel, veresell, veresellik, vereslik* (look red VERB), *veresít, veresit* (make it red VERB)

V. BLUE: *kék* (blue NOUN ADJ), *kékség* (blueness NOUN), *kékes* (bluish ADJ), *kékesség* (bluishness NOUN), *kékül* (turn blue VERB), *kékel, kékell, kékellik, kéklik* (look blue VERB), *kékít, kékit* (make it blue VERB)

VI. GREEN: *zöld* (green NOUN ADJ),[4] *zöldes* (greenish ADJ), *zöldesség* (greennishness NOUN), *zöldül* (turn green VERB), *zöldel, zöldell, zöldellik*, zöldlik (look green VERB), *zöldít, zöldit (*make it green VERB*)*

VII. YELLOW: *sárga* (yellow NOUN ADJ), *sárgaság* (yellowness NOUN), *sárgás* (yellowish ADJ), *sárgásság* (yellowishness NOUN), *sárgul* (turn yellow VERB), *sárgál, sárgáll, sárgállik* (look yellow VERB), *sárgít, sárgit* (make it yellow VERB)

[4] I have not used the noun *zöldség* (zöld[ADJ] + ség), since its main meaning is 'vegetable'.

VIII. BROWN: *barna* (brown NOUN ADJ), *barnaság* (brownness NOUN), *barnás* (brownish ADJ), *barnásság* (brownishness NOUN), *barnul* (turn brown VERB), *barnál*, *barnáll*, *barnállik* (look brown VERB), *barnít*, *barnit* (make it brown VERB)

The method was the same as in the case of the two previous conceptual domains. The Python script assigned a poem into a specific color group when the poem contained any of the expressions in (3) referring to that color. A poem can be categorized as a member of several color groups if it contains the expressions of more than one color. Table 4 shows the resulting frequency data in a similar way as in the case of seasons.

Table 4. Frequencies of poems referring to COLORS

Author	**Colors %**	**1.**		**2.**		**3.**		**4.**	
		color	**%**	**color**	**%**	**color**	**%**	**color**	**%**
Csokonai	22.3%	red	9.4%	blue	7.9%	black	5.3%	yellow	5.3%
Berzsenyi	27.9%	green	13.2%	brown	7.4%	blue	4.4%	yellow	3.7%
Kisfaludy	34.1%	brown	12.2%	green	10.6%	red	8.1%	blue	7.3%
Kölcsey	34.9%	green	22.1%	blue	10.7%	brown	8.7%	red	2.7%
Vörösmarty	18.9%	brown	6.2%	green	5.9%	red	4.5%	black	4.4%
Arany	31.9%	green	16.5%	red	11.3%	white	9.8%	blue	9.6%
Tompa	57.2%	green	28.5%	white	15.1%	red	13.8%	blue	13.6%
Petőfi	23.0%	red	7.6%	green	5.7%	blue	5.5%	black	4.9%
Madách	25.1%	brown	8.2%	white	6.3%	blue	5.3%	red	5.0%
Gyulai	27.6%	green	12.8%	blue	10.9%	red	6.4%	white	5.8%
Vajda	25.1%	black	9.5%	green	9.5%	white	9.0%	red	8.0%
Reviczky	17.9%	blue	7.8%	green	3.9%	red	3.6%	white	2.7%
Komjáthy	12.6%	black	4.1%	blue	4.1%	green	3.7%	white	2.0%
Ady	24.5%	red	8.9%	white	7.5%	black	4.1%	blue	3.5%
Kaffka	46.1%	white	26.5%	blue	13.7%	gray	11.8%	black	9.8%
Somlyó	48.3%	white	16.1%	black	14.0%	red	10.3%	blue	7.7%
Juhász	33.8%	gray	9.3%	blue	7.4%	red	7.0%	black	6.4%
Babits	42.2%	blue	11.7%	white	11.3%	green	10.9%	red	10.1%
Kosztolányi	48.4%	white	17.5%	black	12.1%	red	11.9%	blue	11.1%
Tóth	41.2%	yellow	12.2%	blue	9.8%	black	8.9%	red	8.6%
Reményik	29.2%	black	7.5%	white	7.2%	gray	6.6%	blue	6.0%
József	29.2%	red	8.7%	blue	6.8%	white	6.0%	black	5.2%
Dsida	48.2%	white	17.1%	black	12.9%	red	11.6%	blue	10.0%
Mean	32,6%								
Median	29,2%								
I.		green (5), white (4), red (4)							
II.		green (10), blue (10), white (9)							
III.		red (14), blue (13), green (12), white (12)							
IV.		blue (21), red (19), white (15), black (14)							

The second column shows the proportions of poems referring to at least one color from the eight color concepts analyzed. The spread of the proportions is quite large. The highest proportion, 57.2%, is found for Tompa, as in the case of SEASONS and NIGHT. Kaffka, Somlyó, Kosztolányi and

Dsida are at the top of the frequency list as well, with more than 45%. It seems that authors of the early 20th century preferred to use color concepts more than the earlier authors (Juhász, Babits and Tóth are above the mean and median too). These higher frequencies of 20th century poems referring to COLORS, as well as the higher frequencies of 20th century poems referring to NIGHT may reflect the well-known poetic change known as the emergence of classical modernism in Hungarian literature at the beginning of the 20th century. It is worth noting that the proportions for Reményik and József are lower than for the other 20th century poets (except Ady). Similar patterns can be detected in the case of NIGHT and SEASONS. Another interesting result is that the frequency found for Ady is much more similar to the frequencies found for the 19th century authors than to the frequencies found for the early 20th century authors. A similar trend holds for the domain of NIGHT. If we were to separate literary periods solely on the basis of the frequency data of these two conceptual domains, Ady would not be classified with the other authors of the early 20th century. There are also authors who refer to colors much less frequently. For instance, only 12.6% of Komjáthy's poems contain expressions for colors. Vörösmarty and Reviczky are also below 20%.

The bottom four rows show which colors appear most often among the first, the first and second, the first, second and third, and the first, second, third and fourth most commonly used colors. It can be seen that among the first most commonly used colors, GREEN is used by the most authors, and WHITE and RED are used by the second most authors. Looking at the most and second most commonly used colors as one group (row II.), GREEN and BLUE are the most popular color. There is an interesting distribution of the most preferred colors over time. GREEN appears in the poetry of almost all 19th century authors as the most or second most frequent color. However, this color does not appear at all among the most and second most frequently used colors in the poems of the 20th century poets. In the case of WHITE there is a reversed tendency. For most 20th century poets, WHITE is the most or second most frequently used color. On the other hand, in the case of the 19th century authors, WHITE occurs only two times among the most and second most frequent colors. It seems that the disappearance of GREEN and the emergence of WHITE among the most preferable colors also reflect the change between two poetic eras of Hungarian literature at the beginning of the 20th century.

Another question is which authors used the most colors to a large extent. To answer the question, the mean and the standard deviation of the frequencies of the eight colors have been calculated for each poet. These data are shown in Table 5. A high mean with a low standard deviation indicates that the author used many colors with a similarly higher frequency. For instance, Table 5 shows that in the case of Babits, the mean is high, 9.03, and the standard deviation is only 2.33, which is a fairly low value compared to the other standard deviations associated with high means. This leads to the conclusion that in Babits's poetry there are not just one or two salient, more frequent colors but rather he used several colors in similarly high frequency, which implies a more impressionistic poetic attitude, more attentive to sensory impressions.

Table 5. Means and standard deviations of the frequencies of poems referring to different colors

Author	Mean	Standard deviation
Csokonai	4.29	3.34
Berzsenyi	4.31	4.26
Kisfaludy	5.69	4.46
Kölcsey	5.7	7.81
Vörösmarty	4.1	1.48
Arany	9.01	3.94
Tompa	12.88	7.64
Petőfi	4.36	2.14
Madách	4.66	2.38
Gyulai	4.96	4.87

Vajda	6.19	3.52
Reviczky	2.85	2.35
Komjáthy	1.99	1.72
Ady	3.88	2.87
Kaffka	10.16	7.63
Somlyó	9.11	3.99
Juhász	5.95	2.22
Babits	9.03	2.33
Kosztolányi	9.97	4.25
Tóth	8.05	2.48
Reményik	4.7	2.53
József	5.36	1.85
Dsida	9.16	4.82

6. Correlation of conceptual domains

The frequency data of the three conceptual domains display similar tendencies. For instance, we have seen that Tompa used the concepts of NIGHT, SEASONS and COLORS with the highest frequency. It has also been detected that the frequencies of the conceptual domains (especially NIGHT and COLORS) are higher in the case of many authors from the 20th century than with a number of earlier authors. Based on these results, it seemed to be a plausible assumption that there is a positive correlation between the frequencies of the three conceptual domains in the case of the canonical Hungarian authors analyzed. To test this hypothesis, Pearson correlation coefficient has been calculated for the three possible pairs of the three domains. The coefficient is always between -1 and 1. Zero means that there is no correlation at all between the two datasets, 1 and -1 mean perfect positive and perfect negative correlation. The resulting correlation coefficients are shown in Table 6.

Table 6. Correlation of the frequencies of poems referring to NIGHT, SEASONS and COLORS

	Pearson's r
NIGHT – SEASONS	0.67
NIGHT – COLORS	0.88
SEASONS – COLORS	0.58

All three correlation coefficients are above 0.5, which means that there is a strong positive correlation between the frequencies of the conceptual domains. The value of 0.88 indicates a particularly strong correlation between the two variables. The strong positive correlation between the frequencies of the conceptual domains means that an increase in the number of occurrences of one conceptual domain usually goes together with an increase in the number of occurrences of the other two conceptual domains, and the decrease of the occurrences of one domain usually goes together with the decrease of occurrences of the other two domains. In other words, the majority of authors referring to NIGHT to a higher extent refer to SEASONS and COLORS to a higher extent as well and the majority of authors referring to SEASONS to a higher extent also refer to COLORS to a higher extent. The positive correlation between the frequencies of the three conceptual domains may mean that referring to these domains is part of the same poetic toolkit, which aims to highlight the sensory impressions of the physical setting of the lyrical situation.

7. Co-occurrences of the concepts from different conceptual domains

The concepts from different conceptual domains can co-occur in the same poem in different combinations. It is an interesting question whether certain combinations are more typical than others. It can be revealing as well if some combinations are less typical than others. For the assessment of typicality and atypicality of combinations, the mutual information scores have been calculated. Although in linguistics, mutual information was introduced as an association measure for surface proximity and syntactic co-occurrence (Church–Hanks 1990), it is also suitable for measuring the co-occurrence of concepts in the same poems (this is the case of textual co-occurrence, see Evert 2009: 1220–1224). An MI-score greater than zero means that the two concepts occur together in the same poem more times than would be expected by chance. In this case, the subcorpora of the different authors were taken as one single corpus and the calculations were carried out on this basis. The formula of MI-score is shown in (4). A is the number of poems containing concept 1, B is the number of poems containing concept 2, C is the number of all poems in the corpus, and O is the number of poems containing concept 1 and concept 2 as well.

(4) $$MI = log_2 \frac{O}{\left(\frac{A \times B}{C}\right)}$$

The higher the mutual information score, the stronger the association between the two concepts, in other words, they are more likely to co-occur in the same poem. Negative MI-score means that there is a tendency that the two concepts do not occur together. For the calculation, a script was used, which counted the co-occurrences of all possible two concepts from different conceptual domains in the same poem and calculated the mutual information scores. I excluded poems longer than 300 words from the analysis as these longer texts distort the results of the co-occurrence analysis. Table 7 shows the top 20 highest scoring pairs of concepts belonging to different conceptual domains. These concepts are more likely to appear in the same poems than the other concepts under study. It can be said that they attract each other in Hungarian canonical poetry. The fourth column shows the MI-score and the fifth column shows the number of poems in which both concepts appear.

Table 7. MI-score of concept pairs occurring in the same poems

	Concept 1	Concept 2	MI-score	Occurrence
1	autumn	yellow	1.86	87
2	spring	green	1.28	115
3	summer	yellow	1.03	40
4	autumn	gray	0.98	54
5	winter	green	0.96	59
6	summer	green	0.84	58
7	spring	blue	0.82	100
8	autumn	green	0.81	69
9	winter	white	0.76	67
10	spring	yellow	0.74	48
11	autumn	red	0.72	81
12	autumn	blue	0.7	76
13	night	white	0.67	318
14	night	gray	0.65	166
15	summer	red	0.62	62

16	summer	blue	0.61	59
17	night	black	0.59	247
18	winter	yellow	0.56	27
19	night	brown	0.54	141
20	night	yellow	0.53	133

It is worth noting that the concept of NIGHT is not associated with any SEASONS with an MI-score greater than 0.5. It seems that in Hungarian canonical poetry, NIGHT and SEASONS do not attract each other as much as NIGHT and certain COLORS or SEASONS and certain COLORS. It is also striking that the concept of NIGHT does not appear among the top 12 highest scoring pairs of concepts. This may be due to the simple fact that sensory information, especially color, fades at night. This explanation is supported by the fact that the first four of the five concepts occurring with NIGHT in the top 20 highest scoring pairs are “colorless” colors: WHITE, GRAY, BLACK and BROWN. There is one pair with a negative MI-score: SPRING – BROWN (MI: -0.22, occurrence: 26). The negative MI-score means that these concepts repel each other, that is, they tend not to occur in the same poems.

8. Summary and some final remarks

The paper has presented a quantitative approach of poetry based on the distribution of the concepts of closed conceptual domains. Closed conceptual domains are well-delimited conceptual fields consisting of a finite number of concepts. In the present study, three closed conceptual domains are analyzed in Hungarian canonical poetry: the domains of TIMES OF DAY, SEASONS and COLORS. The frequencies of the concepts of these domains highlight certain aspects of the physical setting of the authors’ poetry. We have seen a general trend that the frequencies of the conceptual domains NIGHT and COLORS are usually higher for early 20th century authors than for earlier authors. It has also been shown that the most frequent seasons for 20th century poets are much more varied than for earlier poets, and that while GREEN was the most popular color until the end of the 19th century, WHITE was the most popular color afterwards. The change in these frequencies seems to reflect a poetic change known as the emergence of classical modernism in Hungarian literature at the beginning of the 20th century. Another interesting result is that the frequency of NIGHT and COLORS found for Ady, who is considered by literary historians to be the first great poet of Hungarian classical modernism, is much more similar to the frequencies found for the 19th century authors than to the frequencies found for the early 20th century authors.

It has also been demonstrated that in the poetry of the Hungarian canonical authors under study, there is a strong positive correlation between the proportions of poems referring to the three conceptual domains. This means that in the case of authors where the frequency of the author’s poems referring to one of the three conceptual domains is higher, the frequency of poems referring to the other two conceptual domains is usually higher as well. Such positive correlation between the frequencies of the three conceptual domains implies that referring to the three conceptual domains is part of the same poetic toolkit aiming to highlight the sensory aspects of the physical setting of poems. Finally, the mutual information scores of all pairs of concepts from different conceptual domains were calculated. This method was applied to identify pairs of concepts which occur more often in the same poem than would be expected by chance.

It is worth mentioning that the three conceptual domains analyzed have a strong metaphorical potential. The concept NIGHT is usually a metaphorical source domain for SADNESS, LONELINESS, DEATH, INACTIVENESS, NON-EXISTENCE, etc. Similarly, WINTER can be a metaphorical source domain for such concepts as well. On the other hand, SPRING and SUMMER are typical source domains for LOVE, LIFE, HAPPINESS, ACTIVENESS, etc. (on conceptual metaphors, see Lakoff–Johnson 1980; Lakoff 1992). The analysis presented here did not take these concepts as metaphorical source domains into account. It is a future task to elaborate more sophisticated methods, which can combine the quantitative analysis of closed conceptual domains with the description of semantic functions.

Acknowledgements

This paper was supported by the National Laboratory for Digital Heritage and by the project No. K-137659 (Corpus-based cognitive poetic research on person marking constructions) of the National Research, Development and Innovation Office of Hungary.

References

Church, Kenneth W. – Hanks, Patrick. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.

Culler, Jonathan 1981. *The pursuit of signs: Semiotics, literature, deconstruction*. London – New York: Routledge.

Evert, Stefan 2009. Corpora and collocations. In: Lüdeling, Anke – Kytö, Merja (eds.): *Corpus linguistics: An international handbook.* Vol. 2. Berlin – New York: Walter de Gruyter. 1212–1248.

Fónagy, Iván 1959. *A költői nyelv hangtanából* [Studies in the phonology of poetic language]. Budapest: Akadémiai Kiadó.

Indig, Balázs – Sass, Bálint – Simon, Eszter – Mittelholcz, Iván – Vadász, Noémi – Makrai, Márton 2019. One format to rule them all – The emtsv pipeline for Hungarian. In: Friedrich, Annemarie – Zeyrek, Deniz – Hoek, Jet (eds.): *Proceedings of the 13th Linguistic Annotation Workshop.* Stroudsburg: Association for Computational Linguistics. 155–165. https://doi.org/10.18653/v1/W19-4018

Jékel, Pál – Papp, Ferenc 1974. *Ady Endre összes költői műveinek fonémastatisztikája* [Phoneme statistics of the complete poetic works by Endre Ady]. Budapest: Akadémiai Kiadó.

Jékel, Pál – Szuromi, Lajos 1980. *Petőfi metrumai* [Petőfi's meters]. Debrecen: Kossuth Lajos Tudományegyetem.

Labádi, Gergely 2018. Az olvasó gép: Berzsenyi Dániel versei távolról [The reading machine: distant reading of poems by Dániel Berzsenyi]. *Digitális Bölcsészet* 1: 17–34. https://doi.org/10.31400/dh-hun.2018.1.126

Lakoff, George 1992. The contemporary theory of metaphor. In: Ortony, Andrew (ed.): *Metaphor and thought.* Cambridge: Cambridge University Press. 202–251. https://doi.org/10.1017/CBO9781139173865.013

Lakoff, George – Johnson, Mark 1980. *Metaphors we live by.* Chicago: The University of Chicago Press.

Lesi, Zoltán 2008. Automatikus formai verselemzés [Automatic formal poem analysis]. *Alkalmazott Nyelvtudomány* 8(1–2): 197–208.

McIntyre, Dan 2013. Language and style in David Peace's 1974: a corpus informed analysis. *Études de stylistique anglaise* 4: 133–146. https://doi.org/10.4000/esa.1498

Moretti, Franco 2000. Conjectures on world literature. *New Left Review* 1: 54–68.

Moretti, Franco 2005. *Graphs, maps, trees: Abstract models for literary history.* Lodon – New York: Verso.

Rayson, Paul 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 3(4): 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Simon, Gábor 2020. Géppel mért műfajiság: Esettanulmány a modern magyar elégia korpuszalapú vizsgálatához [Machine-measured genre: A case study for a corpus-based analysis of modern Hungarian elegies]. In: Kulcsár-Szabó, Zoltán (ed.): *Hagyomány és innováció a magyar és világirodalomban* [Tradition and innovation in Hungarian and world literature]. Budapest: Eötvös Kiadó. 45–64.

Tátrai, Szilárd 2011. *Bevezetés a pragmatikába* [Introduction to pragmatics]. Budapest: Tinta Könyvkiadó.

Tátrai, Szilárd 2015a. Context-dependent vantage points in literary narratives: A functional cognitive approach. *Semiotica: Revue publiée par l'association internationale des semiotique = Journal of the international association for semiotic studies.* 9–37. https://doi.org/10.1515/sem-2014-0076

Tátrai, Szilárd 2015b. Apostrophic fiction and joint attention in lyrics: A social cognitive approach. *Studia Linguistica Hungarica* 30: 105–117.

TEI Consortium 2019. *TEI P5: Guidelines for electronic text encoding and interchange: Version 3.5.0.* https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Váradi, Tamás – Simon, Eszter – Sass, Bálint – Mittelholcz, Iván – Novák, Attila – Indig, Balázs – Farkas, Richárd – Vincze, Veronika 2018. e-magyar – A digital language processing system. In: Calzolari, Nicoletta – Choukri, Khalid – Cieri, Christopher – Declerck, Thierry – Hasida, Koiti – Isahara, Hitoshi – Maegaard, Bente – Mariani, Joseph – Moreno, Asuncion – Odijk, Jan – Piperidis, Stelios – Tokunaga, Takenobu (eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Paris: European Language Resources Association. 1307–1312.

Zsilka, Tibor 1974. *Stilisztika és statisztika* [Stylistics and statistics]. Budapest: Akadémiai Kiadó.