

# ANNOTATION OF PERSON MARKING CONSTRUCTIONS IN THE CORPUS OF HUNGARIAN LYRICAL POETRY: PRINCIPLES AND PRACTICES

PÉTER HORVÁTH

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

GÁBOR SIMON

ELTE Eötvös Loránd University  
simon.gabor@btk.elte.hu  
<https://orcid.org/0000-0001-5233-6313>

SZILÁRD TÁTRAI

ELTE Eötvös Loránd University / Jagiellonian University  
tatrai.szilard@btk.elte.hu  
<https://orcid.org/0000-0002-1069-6676>

## Abstract

This paper presents the annotation scheme for the manual annotation of the Corpus of Hungarian Lyrical Poetry. The corpus will consist of 400-600 annotated texts, grouped into four sub-corpora: 20th century lyrical texts from the canon of Hungarian public education, contemporary lyrical texts, slam poetry texts, and song lyrics. The manual annotation is based on automatically generated and manually checked annotations of lemmas, parts of speech and morphosyntactic features. The manual annotation of syntactic properties proposed in the annotation scheme follows a dependency analysis approach and allows us to obtain quantitative data on person marking constructions in Hungarian lyrical texts. Besides the annotation of verb-dependent relations, the paper also presents the annotation of specific phenomena such as auxiliary verbs, vocatives, elliptical structures, and nominal predicates. The annotation scheme was tested using a test corpus of 16 texts. We also provide some examples of the types of quantitative data that can be extracted from the annotated corpus.

**Keywords:** Corpus of Hungarian Lyrical Poetry, person marking, manual annotation, annotation scheme, dependency analysis

## 1. Introduction

The constructions of person marking constitute a subsystem of the grammar of a language (see Cysouw 2009), and may seem to fulfill a mere grammatical function in the paradigm of pronouns and in verb agreement. However, our key theoretical assumption is that these constructions contribute significantly to the poetic character of lyrical texts. The questions of who speaks to whom in a lyrical discourse, or what is the role of apostrophe in lyrical poetry have a long history in literary criticism (see e.g. Culler 2015, Waters 2003, Jackson–Prins

eds. 2014, and Pethő–Tukacs, this volume). Moreover, one can consider the role of personification or anthropomorphisation in the unfoldment of apostrophic addresses. In other words, the identification of figures and characters in poetry can be considered a defining factor in both genre theory and lyric theory. Despite its theoretical significance, the systematic study of person marking in lyrical poetry is yet to be carried out. The main aim of our paper is to narrow the gap between theoretical and empirical research, providing a solid methodological framework for investigating person marking in lyrical poetry in a corpus-based manner.

The contribution of person marking to the emergence of poetic quality can only be partially explored through case studies based on qualitative analyses. To gain more general insight into the functioning of person marking in poetic discourses, quantitative data are needed. In linguistics, there are two ways to obtain quantitative data about a linguistic phenomenon. The first is the use of experimental methods, which aim to collect quantitative data from informants. The second option is to use a corpus and collect quantitative data by observing and measuring linguistic patterns in the corpus. By building the Corpus of Hungarian Lyrical Poetry, we aim to obtain quantitative data in the latter way. Currently, there is no annotated corpus that allows the detailed quantitative analysis of syntactic features related to person marking in Hungarian lyrical discourses.

This paper focuses on the annotation scheme developed for the manual annotation of syntactic features related to person marking in the Corpus of Hungarian Lyrical Poetry. In section 2, we briefly present the sub-corpora and the annotation methods of the Corpus of Hungarian Lyrical Poetry. Section 3 outlines the main principles of the annotation procedure. Sections 4 and 5 present the annotation scheme in detail with examples from a test corpus of 16 texts. Section 6 highlights some quantitative data types that can be extracted from the annotated data. Finally, in section 7, we give a brief summary and suggest some further possibilities for extending the annotation scheme.

## **2. The Corpus of Hungarian Lyrical Poetry**

When designing the corpus, it was a crucial hypothesis that lyricism can be described as a continuum, with more lyrical and less lyrical texts. This means that the defining properties of lyrical texts, such as poetic simultaneity (see Volk 2002), lyrical directness and apostrophic fiction (see Culler 1981: 135–154, Tátrai 2015), are not only present in the canonical lyrical texts of so-called high literature, but also in song lyrics and slam poetry. The total size of the corpus will be 400-600 texts. The corpus will consist of 4 sub-corpora, each containing 100-150 manually annotated texts. The sub-corpora are the following.

- 20th century lyrical texts from the canon of Hungarian public education
- Contemporary lyrical texts
- Song lyrics
- Slam poetry texts

Dividing the corpus into sub-corpora containing different types of lyrical texts has two benefits. On the one hand, it allows for a comparison of trends between discourse types and on the other hand, it offers the possibility to include further sub-corpora in the future.

The manual annotation of syntactic features is based on the automatic annotation methods of ELTE Poetry Corpus. ELTE Poetry Corpus is a database containing all the poems of 50 canonical Hungarian poets (Horváth et al. 2022). Besides the texts of the poems, ELTE Poetry Corpus contains automatic annotations of structural units (titles, stanzas, lines) and sound devices such as rhyme patterns, rhyme pairs, rhythm of lines, alliterations, and phonological features of words (Horváth 2020). In addition to the annotation

of structural units and sound devices, the texts of ELTE Poetry Corpus were tokenized and the lemma, part of speech and morphosyntactic features of the words were also automatically annotated by the e-magyar toolchain (Váradi–Simon–Sass et al. 2018; Indig–Sass–Simon et. al. 2019; Simon–Indig–Kalivoda et al. 2020). In the case of the Corpus of Hungarian Lyrical Poetry, the same workflow and tools are used for the automatic annotation as in the case of ELTE Poetry Corpus.

### 3. The main principles of the annotation scheme

The annotation scheme presented in this paper has been elaborated in several phases and the different phases have been tested on a test corpus of 16 texts. The test corpus contains 13 poems written in the 20th century and 3 song lyrics. Besides the authors of this paper, 5 annotators participated in the test annotations.<sup>1</sup> They were all academics or PhD students. For the test annotations we used WebAnno, a corpus tool developed for projects using multiple annotators (Yimam–Gurevych–Eckart de Castilho et al. 2013; Eckart de Castilho–Mújdricza–Maydt–Yimam et al. 2016). WebAnno’s interface made it possible to assess inter-annotator agreement and to detect and modify problematic parts of the annotation scheme. In the course of developing the annotation scheme, we had several meetings, where the annotators reported on the difficulties encountered. This feedback from the annotators was also taken into account when we finalized the annotation scheme.

At the heart of the elaborated scheme is the annotation of verbs and their direct dependents, which play a fundamental role in the expression of person relations. The manual annotations will be based on the automatic annotation of the lemma, part of speech and morphosyntactic features of words. Thanks to the automatic annotations checked manually, only syntactic relations between the elements of verbal structures need to be annotated manually. We mostly annotate relations between the elements of verbal structures in the usual way of dependency analyses. Besides constituency analysis, dependency analysis is the most typical way to annotate syntactic structures. One of the main advantages of dependency analysis is that it is compatible with the syntactic analysis of computational methods and with the cognitive linguistic theoretical framework of this research project (see Geeraerts–Cuyckens eds. 2007, Langacker 2008, Tolcsvai Nagy szerk. 2017). For instance, in the D1 dimension of the multi-dimensional functional cognitive model of the Hungarian sentence elaborated by Imrényi (2013, 2017, 2019), relations between the verb and linguistic elements referring to the participants and circumstances of the event expressed by the verb are described as dependency relations.

The dependency analysis results in a dependency tree for each sentence or verb structure analyzed. A dependency tree is a graph whose nodes are the words of the sentence and the root node, i.e. the node at the top of the graph, is by default the verb of the sentence. The edges between the nodes represent dependency relations between the words. These dependency relations can be described as a relation between a head node and a dependent node. Nodes of the dependency tree that are not at the top or the bottom level are heads and dependents at the same time. For example, the root node of sentence (1) is the verb *elment*. The nouns *lány* and *boltba* functioning in the sentence as subject and adverbial argument are direct dependents of the verb. However, the noun *lány* is not only the direct dependent of the verb, but also the direct head of the preceding article and the adjective *legfiatalabb*. Similarly, the noun *boltba* is the direct head of the article preceding it. The dependency analysis of sentence (1) is shown in Figure 1.

(1) A leg-fiatal-abb lány el-men-t a bolt-ba.

<sup>1</sup> Besides the authors of this paper, the following researchers from the Stylistic Research Group took part in the test annotations: Júlia Ballagó, Ágnes Kuna, Andrea Pap, József Pethő, Réka Sólyom.

The SUP-young-CMPR girl VPFX-go-PST.3SG.NDEF the shop-INE  
 'The youngest girl went to the shop.'

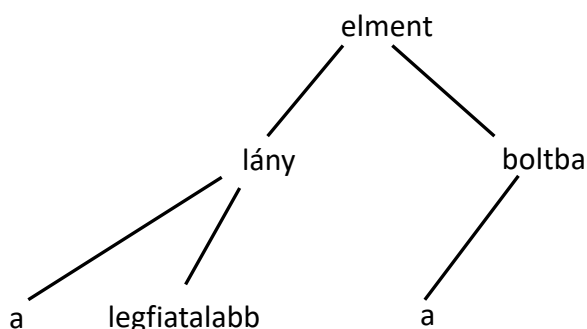


Figure 1.

Although the nodes in dependency analysis are usually words, in the annotation scheme, in some cases, we have allowed the node to be a structure of two or more words placed side by side (noun + postposition structures, verb + preverb structures, vocatives).

When designing the annotation scheme, we followed the principle that the nodes of the dependency tree can only be actual words or structures that are linguistically realized in the sentence. In other words, we do not complete sentences with zero pronouns and zero copulas, nor do we complete elliptical structures with the missing verb (this approach is followed by Vadász 2020). However, we have built the annotation of elliptical structures and predicate structures without copula into the annotation scheme.

In the course of manual annotation, two types of information are annotated. On the one hand, we annotate the verbal structures' elements having different syntactic and semantic roles. On the other hand, we annotate the relations between these elements. This means that the labels of the manual annotation have two main groups: they refer either to a word or a structure consisting of more than one word, or to a relation between words or structures. In some cases, additional information is added to the labels. In most cases, the labels annotating the nodes and the relations between nodes result in redundancy. For instance, preverbs get a Prev label, and the relation between the verb and the preverb also gets a prev label. However, the redundant labeling of relations makes it easier to check and correct annotations and to write scripts converting the corpus to other formats.

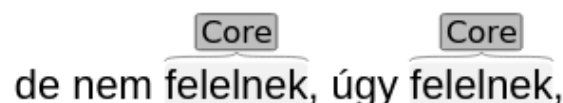
#### 4. The first stage of the annotation of verbal structures

Since the annotators can only focus on a few things at a time, we have defined a first, less detailed stage of the annotation process, which is followed by further stages adding further annotations to the existing ones. This first stage consists of the annotation of verbs constituting the root nodes of dependency trees and the annotation of direct dependents of the verb, auxiliary verbs, preverbs, and vocatives.


##### 4.1. Annotating verb + dependent relations

Verbs get a label Core and direct dependents get a label Arg. Verbs get the label Core even when they have no dependents. We annotate as dependents the arguments of the verbs

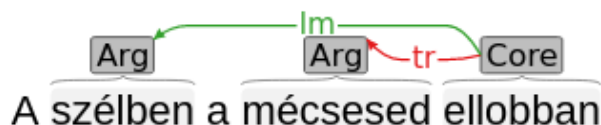
and the nominal adjuncts. Other types of adjuncts are not annotated.<sup>2</sup> For example, adjuncts referring to the time or mode of the action expressed by the verb are usually not labeled, as they are mostly not nouns but adverbs. Adjectives and articles before the dependents are not labeled as part of the dependent. The verb is linked to the dependents by an arrow pointing from label Core to label Arg. The link gets a tr (trajector) or lm (landmark) label. In the terminology of Cognitive Grammar (Langacker 2008), a trajector is in the focus of attention, it is the primary figure of the process expressed by the verb.<sup>3</sup> Usually, its thematic role is agent and its syntactic function is subject. Landmarks are additional, non-agent figures of the process, usually appearing in the sentence as direct or indirect object. Since the morphosyntactic features of words are annotated automatically, it is not necessary to manually annotate the case or number of the verbal dependents. (However, the manual checking of these automatic annotations is necessary.)



(2) 'but they don't reply [Core] , so they reply [Core]'

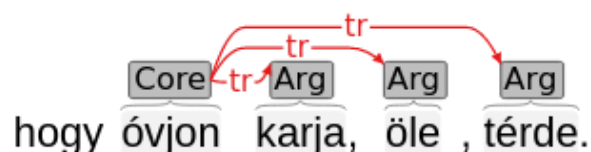


(3) 'an ancient Hungarian song [Arg-tr] still rings [Core] in my ears [Arg-lm]'



(4) 'your candle [Arg-tr] goes out [Core] in the wind [Arg-lm]'

When a verb has more than one direct dependent with the same role, they are labeled and linked to the verb separately.

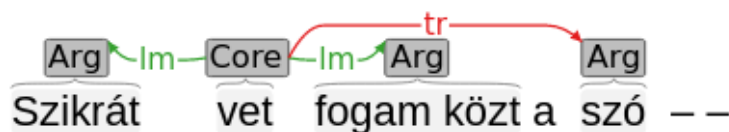


(5) 'in order to protect [Core] you with her/his arm [Arg], lap [Arg], knees [Arg]'

<sup>2</sup> It is not the aim of this research to theoretically clarify the difference between arguments and adjuncts. However, for the annotation to be successful in the future, the annotation scheme should be refined to provide some "practical" aspects that will help the annotator to distinguish between these two categories.

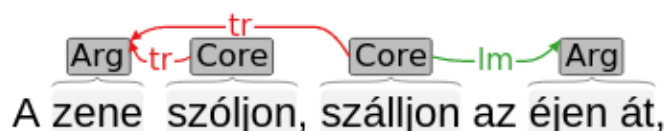
<sup>3</sup> Due to technical reasons, the images demonstrating the examples cannot be edited directly. To ensure the comprehension of both the examples and the marking conventions, we provide the literal translation after each example, with the necessary labels after the English expressions.

In the case of noun + postposition structures, the structure is annotated with one Arg label, and it is linked to the verb as one unit.



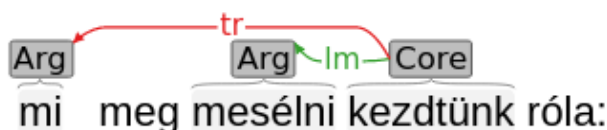
(6) 'The word [Arg-tr] throws [Core] sparks [Arg-lm] in my teeth [Arg-lm]'

When a word is a dependent of more than one verb, the label of the dependent is linked to each verb.



(7) 'Let the music [Arg-tr] play [Core], fly [Core] through the night [Arg-lm]'

In the case of structures consisting of a finite verb and an infinitive, the infinitive is annotated as the dependent of the verb, in the same way as the other dependents. We do not annotate the dependents of the infinitive.



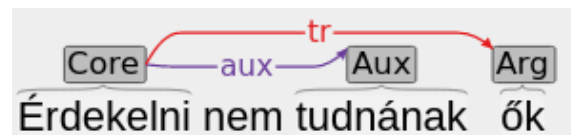
(8) 'We [Arg-tr] started [Core] to tell (stories) [Arg-lm] about him'

## 4.2. Annotating auxiliary verbs and preverbs

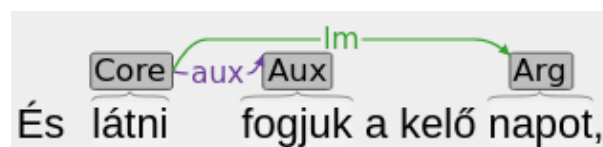
In the case of structures consisting of an auxiliary verb and an infinitive, the infinitive receives the label Core and the auxiliary verb gets the label Aux. The infinitive is linked to the auxiliary verb by an arrow pointing from the label Core to the label Aux. The relation receives the label aux. The dependents of the auxiliary verb + infinitive structure are also linked to the infinitive. It should be noted that the cognitive linguistic literature emphasizes that in Hungarian the auxiliary verb and the infinitive form a semantic unit (Tolcsvai Nagy 2009). This means that it is not necessary to assume a head-dependent relationship between the infinitive and the auxiliary verb. However, due to the dependency approach followed in the annotation system, we had to decide which element of the structure is the head and which is the dependent. The annotation scheme is designed so that in the first phase of the annotation process, all elements are directly linked to the root node (Core element). In other words, we did not want to get chains consisting of more than two elements.<sup>4</sup> To avoid getting

<sup>4</sup> Chains consisting of more than two elements make it difficult to query the annotations and convert the exported annotations from one format to another. In addition, the annotators' task is probably less difficult if they do not have

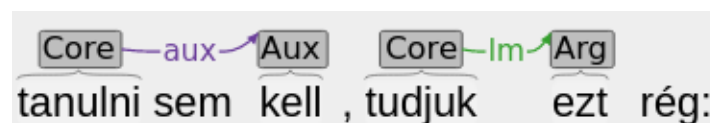
a chain of three elements, we made the infinitive the root node of the structure, since the dependents of the auxiliary + infinitive structure are usually semantically more closely related to the infinitive than to the auxiliary. In a more formal approach, we could say that the selectional restrictions of the infinitive tend to have a stronger effect on dependent choice than the selectional restrictions of the auxiliary verb.



(9) 'I would [Aux] not care [Core] them [Arg]'

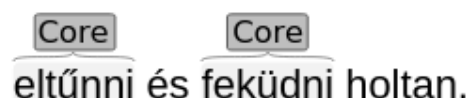


(10) 'And we will [Aux] see [Core] the sun [Arg] coming up'

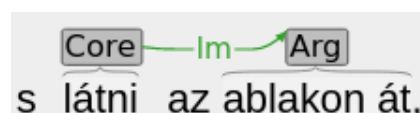


(11) 'you don't even need [Aux] to learn [Core], we've known [Core] this [Arg] for long:'

The infinitives standing alone, without a finite verb or an auxiliary verb, also get the label Core and the dependents of these infinitives are annotated in the same way as the dependents of finite verbs.



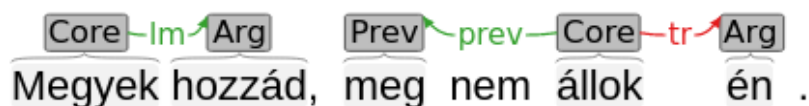
(12) 'to disappear [Core] and lie [Core] dead'



(13) 'and it can be seen [Core] through the window [Arg]'

We also annotate preverbs which are separated from the verb stem, since the verb stem and the preverb form a semantic unit. The preverbs get the Prev label and the relation between the verb and the preverb gets the prev label.

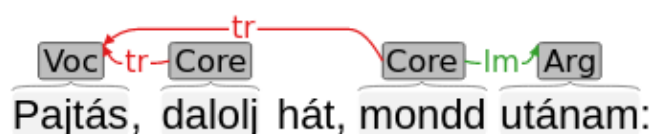
to annotate three-element chains. However, as it can be seen in the next section, we could not avoid the annotation of three-element chains for possessive structures.



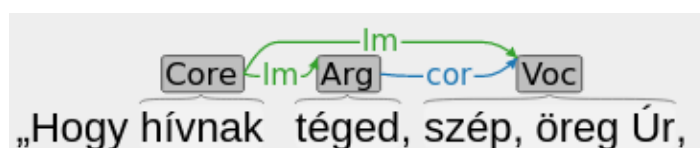
(14) 'I'm coming [Core] to you [Arg-lm], I [Arg-tr]'m not stopping [Core+Prev]'

### 4.3. Annotating vocatives

We usually also annotate vocatives by linking them to the verb, even though syntactically the vocative is not the dependent of the verb. Semantically, however, they elaborate or specify an argument of the verb, so it seemed logical to annotate them as part of the verbal structure. The vocative receives the label *Voc*, which is linked to the verb by an arrow pointing from the label *Core* to the label *Voc*. The relation between the vocative and the verb is labeled either *tr* or *lm*, depending on whether the vocative corresponds to a participant functioning as trajector or landmark. When there is a dependent in the structure that is coreferential with the vocative, then the dependent's *Arg* label is linked to the label of the vocative and the relation is labeled *cor* (coreference). With the label *Voc*, we annotate not only the noun itself but rather the whole vocative structure as one unit, together with adjectives and exclamation words preceding the noun. The advantage of this labeling is that the internal structure of vocatives can also be examined in the future.

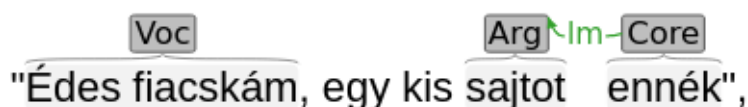


(15) 'Dude [Voc], sing [Core], say [Core] it after me [Arg]:'



(16) 'What do they call [Core] you [Arg], dear old Lord [Voc],'

Sometimes the vocative does not refer to a participant elaborated by an explicit or implicit dependent of the verb. In such cases, we also label the vocative but we do not link it to anything.



(17) 'My dear son [Voc], I'd eat [Core] a piece of cheese [Arg]'

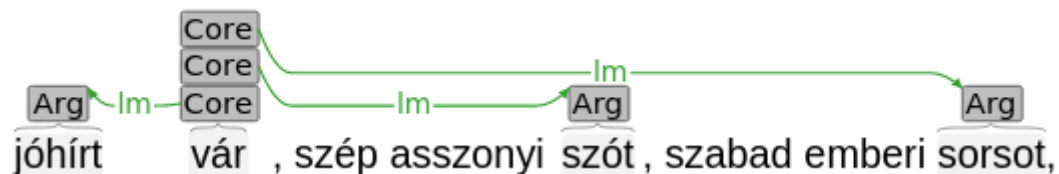


Ó, jaj, barátság, és jaj, szerelem!

(18) 'Oh, alas, friendship and alas, love'

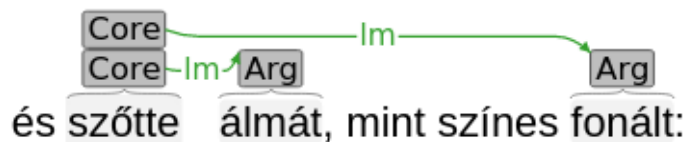
#### 4.4. Annotating elliptical structures

In lyrical texts, we find elliptical clauses quite often. These clauses have an argument structure in which the verb supplying the head of the dependency tree is omitted: it occurs only in a previous (or sometimes in a subsequent) clause. As we have noted, we do not complete such elliptical clauses with the missing verb. When annotating such structures, another Core label is added to the verb having two or more argument structures and the dependents in the elliptical structure are linked to this label. For the Core label of an elliptical structure, ticking a checkbox indicates that the verb is only implicitly present in the annotated argument structure.



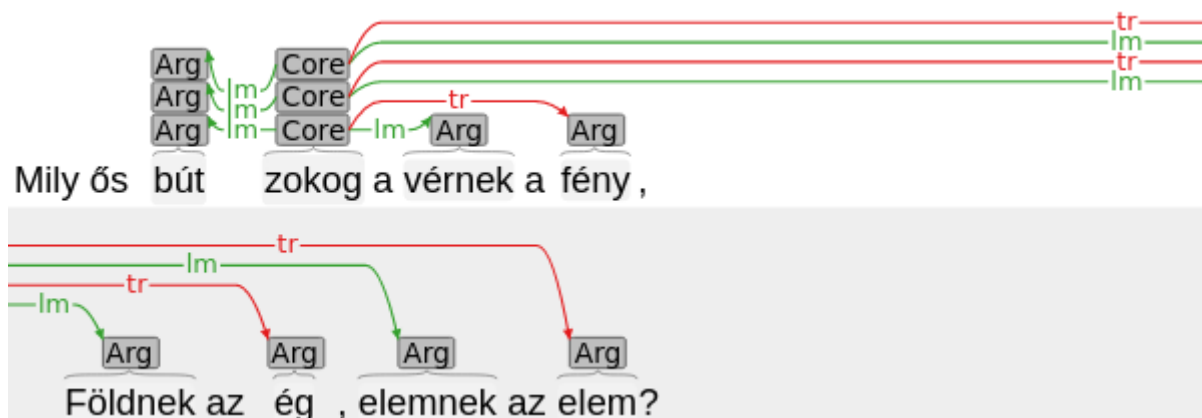
(19) 'he is waiting [Core] for good news [Arg], a nice word [Arg] of a woman, a free human destiny [Arg]'

Conventional similes containing the conjunction word *mint* 'like, as' are treated as similar elliptical structures.



(20) 'and (s)he wove [Core] her/his dream [Arg] like a colorful yarn [Arg]'

In some cases, one or more dependents are also omitted in the elliptical structure in addition to the verb. In such situations, not only the explicit verb but also the explicit dependent gets a second label (Arg) in the preceding non-elliptical structure and this label is linked to the second label of the verb in the same way as for explicit dependents. The only difference is that the checkbox indicating that the argument is implicit has to be ticked.



(21) 'What ancient sorrow [Arg-lm] does the light [Arg-tr] cry [Core] to the blood [Arg-lm], the sky [Arg-tr] to the ground [Arg-lm], element [Arg-tr] to element [Arg-lm]?'

If an auxiliary verb + infinitive structure has an additional elliptical argument structure then the auxiliary verb and the infinitive both receive a further label of Aux and Core, the checkbox is ticked and the dependents of the elliptical structure are linked to the Core label of the infinitive. We have not found examples of this case in the test corpus.

## 5. Further stages of the annotation of verbal structures

In the further stages of the manual annotation of verbal structures, we plan to annotate additional phenomena which enable us to investigate more complex constructions. The annotation of these phenomena is built on the annotation of the first stage. These further stages include the annotation of nominal predicates, possessive nouns, negation words, and implicit arguments.

### 5.1. Annotating nominal predicates

Nominal predicates receive the label CoreN. When the subject is linguistically elaborated in the structure, it receives the label Arg and it is linked to the nominal predicates with the label tr. If the structure contains adverbial nouns (as well), then an Arg tag is added to it, and it is linked to the CoreN tag, with the relation receiving the lm tag.

Mégis győztes, mégis új és magyar.

(22) 'yet victorious [CoreN], yet new [CoreN] and Hungarian [CoreN]'

tölgykerítés, barakk oly lebegő,

(23) 'oak fence [Arg], barack [Arg] so floating [CoreN]'

hogy milliók közt az egyetlenegy.

(24) 'that among millions [Arg] (s)he is the one [CoreN]'

As we have already noted, similes containing the conjunction word *mint* are treated as a special elliptical structure that implies a preceding explicit predicate. When annotating them, the nominal predicate in the first part of the structure is given a second CoreN tag and the checkbox indicating the omission of the predicate is ticked. The standard of comparison in the second part of the structure is linked to this CoreN tag.

mely messze, mint az ég .

(25) 'which [Arg-tr] is as far [CoreN] as the sky [Arg-tr]'

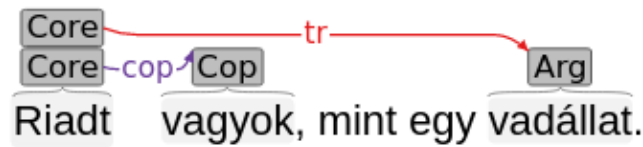
In the case of nominal predicates with a copula, the copula receives a Cop label and it is linked to the nominal's CoreN tag. The relation receives a cop tag.

Fáradt vagyok.

(26) 'I am [Cop] tired [Core].'

Milyen volt szőkesége,

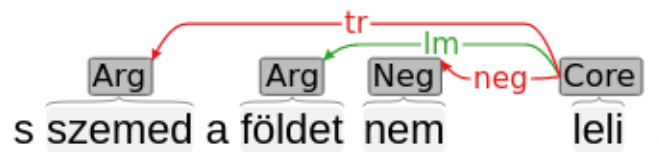
(27) 'how [CoreN] was [Cop] her blondness [Arg]'



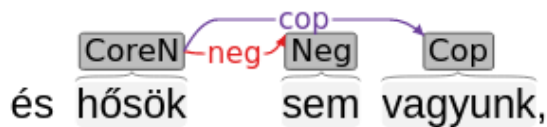
(28) 'I am [Cop] scared [Core] like a wild animal [Arg].'

## 5.2. Annotating negation words and possessive nouns

We also plan to annotate negation words of verbal and nominal predicates. Negation words receive the label Neg, and are linked to the verb or the nominal part of the nominal predicates by an arrow pointing from the Core or CoreN tag to the Neg tag. The relation receives the label neg.

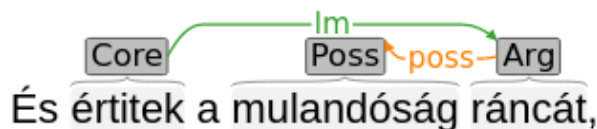


(29) 'and your eyes [Arg-tr] do not [Neg] find [Core] the ground [Arg-lm]'

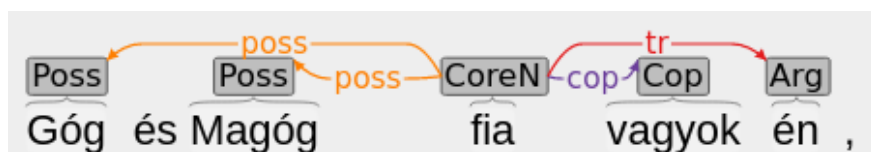


(30) 'and we are [Cop] not [Neg] heroes [CoreN] either'

After the first stage of the annotation, we also intend to annotate possessive nouns modifying verbal dependents and nominal predicates. Since possessive nouns often anchor the possessions to persons, their annotation has particular importance in the context of this research. The possessive noun is given a Poss tag, which is linked to the Arg tag of the verbal dependent or to the CoreN tag of the nominal predicate. The relation receives the label poss.



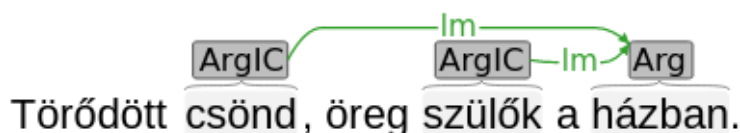
(31) 'And do you understand [Core] the wrinkles [Arg] of transience [Poss]'



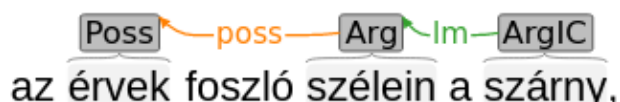
(32) 'I [Arg] am [Cop] the son [CoreN] of Gog [Poss] and Magiog [Poss]'

### 5.3. Annotating argument structures without verb

In some cases, the elliptical structure does not have a verb in a preceding or following argument structure. This means that the dependents have no head at all. In this case the dependent functioning as the trajector (subject) receives an ArgIC label (Argument + Implicit Core) and when there are additional dependents, they are connected to it by a link with an Im label. Negation words and vocatives are also linked to the ArgIC tag.



(33) 'Tired silence [ArgIC], old parents [ArgIC] are in the house [Arg].'



(34) 'the wing [ArgIC] on the parting edges [Arg] of arguments [Poss].'



(35) 'Climates [ArgIC]. Conditions [ArgIC].'

### 5.4. Extending grammatical annotations

Following the first stage of annotation, we plan not only to label additional elements, but also to elaborate on the annotations of elements labeled in the first phase. On the one hand, this allows us to search for more specific patterns. On the other hand, by extending existing syntactic annotations, we can make better use of the theoretical insights of cognitive and construction grammars.

When a verbal dependent labeled is part of the clausal core, it is indicated by ticking a checkbox. The clausal core is the minimal unit within a Hungarian clause which expresses the grounded process profiled by the clause, thus it is constituted by elements necessary for evoking a process type and for grounding an instance of that type (through markers of tense, mood, person and number) (Imrényi 2017: 703, see Imrényi 2019: 83). Usually the clausal core is the verb of the clause. However, there are cases where the process type of the sentence is expressed by the verb and a dependent of the verb together, forming a closer semantic unit (e.g. *intézkedést hoz*, *feleségül vesz*). By annotating dependents that are part of such clausal cores, we obtain a more accurate picture of typical events in lyrical texts.

Annotations of pronouns are also extended if they refer anaphorically or cataphorically to a noun in a preceding or following structure. The antecedent or postcedent of the pronoun can be entered in an empty field. When the pronoun refers to a whole subordinate clause, this is indicated by entering the abbreviation 'sub'.

Although we have not yet taken a definitive position on how to annotate them, we also plan to annotate implicit arguments.

## 6. Quantitative data extracted from annotations

In the course of the test annotation, 16 lyrical texts were annotated. The annotation was carried out on the basis of the first stage of the annotation procedure. Among the texts annotated, there are 13 poems from the 20th century and 3 song lyrics.<sup>5</sup> Naturally, 16 texts are not enough to draw general conclusions about lyrical discourses. However, this test corpus is suitable for presenting some of the quantitative data types that can be extracted from the annotations. Table 1 shows the number of tokens in the test corpus and the number of occurrences of some annotated phenomena.

**Table 1.** Quantitative data from the test corpus

Token	3824
Core label	545
Explicit Core (verb or infinitive)	505
Elliptical structure with implicit Core element (verb or infinitive)	40
Vocative	22
Vocative related to verbal structure	15
Vocative standing alone	7
Trajector dependents	275
Landmark dependents	408
Verbal structures with linguistically elaborated trajector	278
Verbal structures without linguistically elaborated trajector	267

From the data, it can be seen that 7% of the Core tags refer to implicit verbs of elliptical structures. The use of vocatives, which is an integral element of apostrophe, is probably one of the central features of lyrical discourses. This is not contradicted by the data of the test corpus, since there are 22 vocatives in the 16 poems. Although it should be noted that 10 of the 22 vocatives occur in the poem *Nagyon fáj* written by Attila József. There are seven vocatives that stand alone, i.e. which do not elaborate any verbal dependents (trajector or landmark). The significantly lower number of trajectors than landmarks is also in line with the prototype for lyrical discourses in which the agent or experiencer of the events expressed by verbs is typically the fictive speaker or the fictive addressee, who are usually only referred to by verb inflections. In the last two rows, the number of argument structures with and without an explicit trajector is also listed. The number of argument structures with an explicit trajector also includes those cases where the trajector role is elaborated only by a vocative.

<sup>5</sup> The test corpus consists of the following poems and song lyrics: *A Sion-hegy alatt* and *Góg és Magóg fia vagyok én...* by Endre Ady; *Az örök folyosó* by Mihály Babits; *Nagyon fáj* and *Reménytelenül* by Attila József; *Milyen volt...* by Gyula Juhász; *Mesteremberek* by Lajos Kassák; *Ének a semmiről* and *Halotti beszéd* by Dezső Kosztolányi; *Között* by Ágnes Nemes Nagy; *Apokrif* by János Pilinszky; *Hetedik ecloga* by Miklós Radnóti; *Lélektől lélekig* by Árpád Tóth; *Csavad fel a szőnyeget* by István S. Nagy; *Az utcán* by János Bródy; *Zsákmányállat* by András Lovasi.

The frequency of verbal structures with different dependent numbers can be interesting as well. Table 2 shows the number of occurrences of verbal structures with different dependent numbers. Verbal structures with a single dependent are the most frequent in the test corpus.

**Table 2.** Numbers of verbal structures with different dependent numbers

Number of dependents	Number of verbal structures
0	122
1	241
2	128
3	42
4	7
5	2
6	1
7	1
9	1

Although the test corpus does not contain the morphosyntactic properties of words, it should be stressed that the manual annotation presented in this paper is based on automatically created and manually checked morphosyntactic annotations. As the result of automatic annotation, the corpus will specify the lemma, the part of speech and the morphosyntactic features of words. By integrating automatic and manual annotations, we can obtain numerous additional quantitative data. It will be possible to investigate typical lexical realizations of different types of verbal constructions or possessive constructions. For example, it could be investigated which are the typical lexemes that appear as the subject, direct object or indirect object of certain types of verbs. We could also find out which are the typical possessive nouns of third-person characters. One could also look at the types of events that have been conceptualized as non-factual, that is, by conditional or imperative verb forms, with negation words or with auxiliary verbs. It will also be possible to analyze the typical structural and lexical properties of vocatives.

## 7. Conclusion

In this paper, we have presented the annotation scheme of the Corpus of Hungarian Lyrical Poetry. The corpus under construction will consist of four sub-corpora: 20th century lyrical texts from the canon of Hungarian public education, contemporary lyrical texts, slam poetry, and song lyrics. In total, it will contain 400-600 manually annotated texts. The manual annotation allows us to obtain quantitative data on the syntactic structures of Hungarian lyrical texts, in particular on person marking constructions. It extends the automatic and manually checked annotations of lemma, part of speech and morphosyntactic features of words. We have outlined the main principles behind the annotation scheme and presented in detail the proposed way of annotating specific linguistic phenomena. The annotation

scheme was tested using a test corpus of 16 texts. We have also provided some examples of the types of quantitative data that can be extracted from the annotated corpus.

Naturally, the annotation scheme presented here can be extended in the future to annotate additional phenomena related to person marking. For instance, the annotation of place and time deixis can be a further extension of the annotation scheme, since deictic reference to place and time implies a reference to the speaker and/or hearer of the fictive lyrical speech situation as well. Besides deictic reference, the personification of inanimate entities is also a fairly common phenomenon related to person marking. It is part of our future plans to elaborate the categorization system of various types of personification (see Simon 2022) and to integrate it into the annotation scheme.

### Acknowledgements

This paper was supported 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

- Culler, Jonathan 1981. *The pursuit of signs. Semiotics, literature, deconstruction*. London, New York: Routledge.
- Culler, Jonathan 2015. *Theory of the lyric*. Cambridge, MA: Harvard University Press.
- Cysouw, Michael 2009. *The paradigmatic structure of person marking*. Oxford: Oxford University Press.
- Eckart de Castilho, Richard – Mújdricza-Maydt, Éva – Yimam, Seid Muhie – Hartmann, Sylvana – Gurevych, Iryna – Frank, Anette – Biemann, Chris 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In: *Proceedings of the LT4DH workshop at COLING 2016*. Japan: Osaka.
- Geeraerts, Dirk – Cuyckens, Hubert (eds.) 2007. *The Oxford handbook of cognitive linguistics*. Oxford: Oxford UP.
- Horváth, Péter 2020. A vershangzás jellemzőinek automatikus feltárása József Attila verseiben. [Automatic analysis of sound devices in Attila József's Poems] *Digitális Bölcsészettudományi Közlemények* 3: M:3-M:27. <https://doi.org/10.31400/dh-hun.2020.3.422>.
- Horváth, Péter – Kundráth, Péter – Indig, Balázs – Fellegi, Zsófia – Szilávi, Eszter – Bajzát, Tímea Borbála – Sárközi-Lindner, Zsófia – Vida, Bence – Karabulut, Aslihan – Timári, Mária – Palkó, Gábor 2022. ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry. In: Calzolari, Nicoletta – Béchet, Frédéric – Blache, Philippe – Choukri, Khalid – Cieri, Christopher – Declerck, Thierry – Goggi, Sara – Isahara, Hitoshi – Maegaard, Bente – Mariani, Joseph – Mazo, Hélène – Odijk, Jan – Piperidis, Stelios (eds.): *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Paris: European Language Resources Association (ELRA). 3471–3478.
- Imrényi, András 2013. *A magyar mondat viszonyhálózati modellje*. [A relational network model of Hungarian clauses.] Budapest: Akadémiai Kiadó.
- Imrényi, András 2017. Az elemi mondat viszonyhálózata. [The network structure of clauses.] In: Tolcsvai Nagy Gábor (ed.): *Nyelvtan*. [Grammar] Budapest: Osiris, 664–760.
- Imrényi, András 2019. Toward a cognitive dependency grammar of Hungarian. In: Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019). Proceedings. Paris: Association for Computational Linguistics (ACL). 81–88.



- Indig, Balázs – Sass, Bálint – Simon, Eszter – Mittelholcz, Iván – Kundráth, Péter – Vadász, Noémi – Márton, Makrai 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*. Florence: Association for Computational Linguistics (ACL). 155–165.
- Jackson, Virginia – Prins, Yopie (eds.) 2014. *The lyric theory reader. A critical anthology*. Baltimore: Johns Hopkins University Press.
- Langacker Ronald W. 2008. *Cognitive grammar. A basic introduction*. Oxford: Oxford UP.
- Simon, Eszter – Indig, Balázs – Kalivoda, Ágnes – Mittelholcz, Iván – Sass, Bálint – Vadász, Noémi 2020. Újabb fejlemények az e-magyar háza táján. [New developments in the tool e-magyar.] In: Berend, Gábor – Gosztolya, Gábor – Vincze, Veronika (eds.): *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. [XVI. Conference in Hungarian Computational Linguistics.] Szeged: SZTE Informatikai Intézet, 29–42.
- Simon, Gábor 2022. Identification and Analysis of Personification in Hungarian: The PerSECorp project. In: Calzolari, Nicoletta – Béchet, Frédéric – Blache, Philippe – Choukri, Khalid – Cieri, Christopher – Declerck, Thierry – Goggi, Sara – Isahara, Hitoshi – Maegaard, Bente – Mariani, Joseph – Mazo, Hélène – Odijk, Jan – Piperidis, Stelios (eds.): *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Paris: European Language Resources Association (ELRA). 2730–2738.
- Tátrai, Szilárd 2015. Apostrophic fiction and joint attention in lyrics: A social cognitive approach. *Studia Linguistica Hungarica* 30: 105–117.
- Tolcsvai Nagy, Gábor 2009. A magyar segédige + igenév szerkezet szemantikája. [The Hungarian auxiliary + infinitive construction.] *Magyar Nyelvőr* 133: 373–393.
- Tolcsvai Nagy, Gábor (ed.) 2017. *Nyelvtan*. [Grammar.] Budapest: Osiris Kiadó.
- Vadász Noémi 2020. KorKorpusz: kézzel annotált, többretegű pilotkorpusz építése. [KorKorpusz: building a manually annotated, multi-layered pilot corpus.] In: Berend Gábor – Gosztolya Gábor – Vincze Veronika (eds.): *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. [XVI. Conference in Hungarian Computational Linguistics.] Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 141–154.
- Váradi, Tamás – Simon, Eszter – Sass, Bálint – Mittelholtz, 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 – Goggi, Sara – Hasida, Koiti – Isahara, Hitoshi – Maegaard, Bente – Mariani, Joseph – Mazo, Hélène – Moreno, Asuncion – Odijk, Jan – Piperidis, Stelios – Tokunaga, Takenobu (eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 1307–1312.
- Volk, Katharina 2002. *The poetics of Latin didactic. Lucretius, Vergil, Ovid, Manilius*. Oxford: Oxford UP.
- Water, William 2003. *Poetry's touch. On lyric address*. Ithaca, London: Cornell University Press.
- Yimam, Seid Muhie – Gurevych, Iryna – Eckart de Castilho, Richard – Biemann Chris 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In: *Proceedings of ACL-2013, demo session*. Bulgaria: Sofia.