Identification and Analysis of Personification in Hungarian: The PerSECorp project

Gábor Simon
Eötvös Loránd University Budapest
H-1088 Budapest, Múzeum krt. 4/A.
simon.gabor@btk.elte.hu

Abstract

Despite the recent findings on the conceptual and linguistic organization of personification, we have relatively little knowledge about its lexical patterns and grammatical templates. It is especially true in the case of Hungarian which has remained an understudied language regarding the constructions of figurative meaning generation. The present paper aims to provide a corpus-driven approach to personification analysis in the framework of cognitive linguistics. This approach is based on the building of a semi-automatically processed research corpus (the PerSE corpus) in which personifying linguistic structures are annotated manually. The present test version of the corpus consists of online car reviews written in Hungarian (10468 words altogether): the texts were tokenized, lemmatized, morphologically analyzed, syntactically parsed, and PoS-tagged with the e-magyar NLP tool. For the identification of personifications, the adaptation of the MIPVU protocol was used and combined with additional analysis of semantic relations within personifying multi-word expressions. The paper demonstrates the structure of the corpus as well as the levels of the annotation. Furthermore, it gives an overview of possible data types emerging from the analysis: lexical pattern, grammatical characteristics, and the construction-like behavior of personifications in Hungarian.

Keywords: personification, annotation, Hungarian

1. Introduction

Personification can be defined as "the treatment of a non-human concept or entity as if it were human" (Thornborrow and Wareing, 1998), e.g., a *car* can be described as *powerful* or having *sensitive qualities*. More than a decade ago, the investigation of this figurative category was deficient in systematic corpus research: as Dorst (2011) put it, "hardly any empirical work" had been done "on the different manifestations of personification in discourse" and it remained "unclear how personifications can be reliably identified and analyzed." Five years later, the state of the art did not seem to be changed: the editors of a representative collection of studies on personification claimed that "the figure's cognitive form and function, its rhetorical and pictorial effects, rarely elicit scholarly attention", which means that "[a]s a communicative device it is either taken for granted or dismissed as mere convention" (Melion and Ramakers, 2016). Although Dorst and her colleagues devoted remarkable attention to exploring the conceptual complexity and linguistic variability of personification in English (see Dorst, 2011; Dorst, Mulder and Steen 2011), the systematic extension of this promising start onto a larger scale of corpus research is yet to be carried out. The aim of the present paper is to take the first steps towards this extension within the framework of cognitive linguistics.

As an example, it has been observed that the PoS category has a significant role in the emergence of personifying meaning; however, we do not know the exact proportions of verbal, nominal, adjectival etc. personifications in our discourse. Or, turning to another factor, Dorst and her colleagues have found in an experiment that the majority of personifications being identified by informants was multi-word expression (Dorst, Mulder and Steen 2011). Nevertheless, neither the construction-like behavior of personification nor its idiomatic character has been systematically explored yet.

Concerning the Hungarian language, the emerging picture of personification is much less grounded in empirical research. According to a comprehensive account of personification as a figure of speech (Sájter, 2008), the main factors of its description are the following:

- the ontology of the personified entity (e.g., abstract things, natural phenomena, physical objects, plants, animals, or groups),
- the way of personification (e.g., performing an action, emotion attribution, having a human figure or having mental capabilities),
- the grammatical structure of personification (e.g., verbal predicate, possessive construction, nominal and adverbial elements or vocatives),
- and the register-specificity of personification (e.g., colloquial, scientific, journalistic, or literal).

There are two problems with this comprehensive approach. On the one hand, no empirical investigation supports the proposed factors, rather they are based on professional intuition and a collection of Hungarian personifications (mainly from literal texts) as illustrative examples. Having a system specified to identify personifications may provide an expansion of data sources used in research. On the other hand, it can be considered an exhaustive enumeration of conceptual and linguistic characteristics of personifying expressions (including semantic, grammatical, and stylistic features as well), but the operationalization of the factors raises different difficulties for the researcher. The first and the second factors require well-elaborated nominal and verbal ontologies; however, in tackling the way of personification, we can also rely on the language-specific vocabulary of emotions. Grammatical analysis can be carried out automatically, but the annotation of register-specificity needs to involve independent human coders. In other words, the list above is a rather heterogeneous taxonomy of the rich variability of personification in Hungarian, which serves as a good vantage point for sophisticated linguistic analysis, but a unified and general schema of annotation cannot be built on it.

To provide a corpus-driven exploration of personifications in Hungarian on a solid empirical base, we need (i) a corpus with a sufficient number of personifying expressions, (ii) a protocol for gathering grammatical information from the corpus and (iii) an annotation schema for identifying personifications in the corpus. The long-term purpose of the **PerSECorp** project is to establish a corpus of Hungarian for investigating personifying language use, i.e., a novel language resource in which personifications are available as a result of a reliable process of annotation. (Hence the name of the corpus: PerSE is the abbreviation of Personifying Structures Encoded). The present study demonstrates the initial stage of the project with a small-scale annotation of a test corpus consisting of online car reviews written in Hungarian. After outlining the theoretical background of the research (2), the paper will discuss the details of corpus building: the linguistic material of the test corpus, its automatic processing, and the protocol for identifying personifications manually in it (3). Then some preliminary results of the annotation will be discussed: the lexical patterns of personification reoccurring in the analyzed texts, as well as the grammatical characteristics and the potential constructions of personifying expressions (4). The paper ends with a brief conclusion (5).

2. Theoretical Background

At the first sight, the concept of personification seems to be straightforward. From the perspective of cognitive linguistics, however, which is interested in the conceptual motivation of meaning generation, the picture is much more complicated because of the numerous conceptual operations contributing to the emergence of personifying meaning. The traditional cognitive linguistic approach considers personification as a special type of conceptual metaphor, in which the source domain is the human body and our mind, and the target domain is a non-human or non-living entity (Kövecses, 2010). According to this proposal, personification relies on conceptual mappings between two domains, and it is an appropriate model for nominal expressions (which introduce an analogy between two entities, e.g., DRUG IS AN ENEMY, see Dorst, 2011).

There is another general explanation of personification in cognitive linguistics: according to Lakoff (2006), the conceptual pattern of personification is the general metaphor of EVENTS ARE ACTIONS, and the central participant of the event becomes the actor of the metaphorical action. In this approach, mappings unfold between arguments of two domains and not between the domains themselves (Drost, 2011), which is characteristic in verbal personifications.

Two conclusions can be drawn from this brief overview. On the one hand, these two proposals may not be alternative but rather complementary models of personification: the former can motivate perceptual (having a human figure), emotional or cognitive personification, whereas the latter gives an account of the agency of non-living entities. On the other hand, the linguistic manifestation of personification is not of secondary importance since the grammatical structure of the expression orientates the conceptualizer in meaning generation. Consequently, it is worth starting to explore the variability of personification on the level of the linguistic structure, and a reliably annotated corpus can serve as a vantage point for the investigation of conceptual configuration.

Contemporary cognitive linguistics emphasizes the complexity of the conceptual background of personification. Besides the different types of conceptual metaphors, conceptual metonymy is also considered to have an essential motivating role here. Although Low (1999) argued for a careful distinction between metaphorical personification and metonymy (for instance, the expression *the paper concludes* initiates a metonymic reading without attributing human characteristics to the study), Dorst and her colleagues observed an overlap between personification and metonymy, regarding metonymic personifications as a specific subtype of the category (Dorst, Mulder and Steen, 2011). According to their experiment, metonymic personifications are similar to novel personifications. A possible explanation for this observation is that these conceptualizations are motivated by agency attribution. However, there is an essential difference: whereas metaphorical personifications are established on cross-domain mappings, in metonymies, there is a domain-internal attentional switch (Panther and Thornburg, 2007). Therefore, metonymic personifications need to be identified separately to shed light on the organization of the unfolding personifying meaning.

Furthermore, we can also model personification with conceptual integration, in which two mental spaces are combined into a blended space, but it is the network itself that motivates the figurative meaning (Long, 2018). This approach emphasizes not only the multiplicity of conceptual structures involved in meaning generation but also the multi-word character of linguistic personification: according to Long (2018), personification in discourse “can be regarded as an extended unit of meaning […], whose elements include node word, collocation, colligation, semantic preference, semantic prosody”. Thus, the blend model highlights the complexity of both the conceptual and the linguistic structure of personification: “[m]eaning inconsistency in personification is mainly manifested by incongruity between the node word and its collocation”. The term collocation is used somewhat loosely by Long; nevertheless, he directs our attention to the reoccurring patterns of the linguistic components of personification.

To conclude, we can agree with Dorst (2011) that “the identification and analysis of personifications raise different issues at each level of analysis, and the question whether something should count as a personification may yield a different answer for each level.” My proposal, however, goes beyond the mere distinguishing between the levels of analysis, suggesting that a corpus in which grammatical and semantic features are annotated parallel with labelling personifications may serve as a novel language resource for thorough cognitive semantic analysis, grounding empirically the process of theoretical modelling.

What are the pieces of information being essential to investigate personification relying on a corpus? Based on the literature, there are at least two general features to be annotated: the part of speech category (since it is closely related to the conceptual organization) and the morpho-syntactic structure (since it makes reoccurring grammatical or with another term, colligational relations observable).

A further lexical-semantic dimension is conventionality: the degree of the lexicalization of a personifying usage of a word. According to Dorst and her colleagues’ dictionary-based approach, four categories can be distinguished (Dorst, Mulder and Steen, 2011). In the case of novel personifications, the dictionary entry of the word does not include the personifying meaning. An example of it is the expression *őrködik az elektronika* (‘the electronics watches over’), in which the verb *őrködik* (‘to watch over’) does not have any reference to non-human or non-living entities in the dictionary.[1] The opposite of it is conventional personification, by which the meaning described in the dictionary contains the personifying usage as a sub-meaning. For instance, in the expression *erős autó* (‘a strong car’), the adjective *erős* (‘strong’) has the following sub-meaning in the dictionary: ‘a device or machine that functions in its field with a high level of effectiveness’, thus the adjective can be used conventionally as personification (though its basic meaning refers to the physical, bodily power of a human being). The dictionary entry of a default personification does not refer explicitly to a human being, but the standard interpretation of the word assumes a human agent or figure. For example, the basic meaning of the verb *megbújik* (‘be in hiding’) in the expression *két kipufogóvég bújik meg* (‘two exhaust pipes are in hiding’) is the following: ‘hide oneself in a hiding place, lie flat’, which refers only implicitly (or by default interpretation) to a prototypical human actor, since animals can also lie in a hiding place. Finally, metonymic personifications constitute a fourth category on the scale of conventionality: here, the personifying meaning is neither conventional, nor default, but well-entrenched as metonymy. The example of it is the expression *a Mercedes megcsinálja* […] *ferdehátúját* (‘the Mercedes produces […] its fastback’), where the nominal subject (*Mercedes*) refers metonymically to the engineers at the company. The significance of conventionality is twofold. First, it does not follow from the conventionality of the grammatical structure, thus these categories shed new light on the variability of personification in a particular language. Second, this semantic feature having been introduced in previous research makes the cross-linguistic comparison of annotated data possible.

Finally, beyond the meanings of words, the semantic organization of multi-word structures is also important. Cognitive linguistics (cognitive grammar in particular) offers a useful set of categories in this respect. In cognitive grammar, verbs symbolize temporal processes with one or more participants (Langacker, 2013). These participants are conceptualized as schematic figures of the verb: the primary figure (prototypically the agent of the process) is the trajector, while the secondary figures (instantiating mainly the roles of the patient, the instrument, the recipient, or the circumstance of the process) are the landmarks in the semantic structures of the verb. Since these schematic participants are elaborated by nominal components of the clause in the course of construing the meaning of it, the trajector/landmark alignment characterizes not only the verb but the whole construction. Therefore, labelling construction-internal semantic relationships between the components of a personifying expression constitutes a new aspect of analyzing the linguistic structure of personification. Its significance lies in making it possible to observe the function of the personified entity in a larger conceptual scene. In the case of trajector role, this entity is the central argument of the metaphorical source domain of personification with a high level of agency, whereas, in the case of landmark relationship, the entity contributes to the unfoldment of personification but not as an agent of it. In other words, the cognitive grammatical analysis grasps the construction-like organization of personification with a higher amount of precision. Defining constructions as form-meaning pairings (Goldberg, 2006), the previous research on personification highlighted rather the formal pole of it. Long (2018) describes the linguistic structure of personifications in English with complex grammatical patterns, such as “nonhuman object + predicate verb (used for human beings only) + others” or “others + predicate verb (used for human beings only) + nonhuman object + others”, but these templates seem to be underspecified on the one hand (e.g., what does “others” mean from the perspective of personification?) and too particular on the other hand (e.g., is the order of the components significant?). Hungarian, a language with flexible word order and rich morphology (see Rounds, 2001 for further details) has much more various patterns; thus, the comparison of Hungarian data with these basic templates is difficult. Regarding the semantic pole of personification, Dorst and her colleagues claim that the basic schema includes a verbal, adjectival or adverbial element that set up the frame of personification and a noun that denotes the personified entity (Dorst, Mulder and Steen, 2011). This description is general enough to focus on both the grammatical and the semantic organization of personification, but the exact relationship between the personified entity and the frame of personification remains in the background of the analysis. Consequently, this schematization needs to be extended to the analysis of the semantic relation between the nominal and the verbal/adverbial/adjectival elements, and the cognitive grammatical categories offer exactly this extension.

Concerning the language-specific characteristics of personification in Hungarian, the previous research did not result in rich details. The comprehensive discussion of personification (Sájter, 2008) gives a useful overview of the notion, but it does not share the basic assumptions of cognitive linguistics, therefore only partially harmonizable with it. (For more problematic points on operationalization see the Introduction.) The cognitive linguistic exploration of personifications in Hungarian is not without precedents as well: in Simon (2021), the event structures of personifications in the poetry of Attila József has been investigated. The study mapped the grammatical features, the trajector/landmark alignment and the main conceptual categories of personifications in a small-scale poetic corpus, so it laid the foundations of empirical research. But the results cannot be generalized because of the sampling of the corpus and its genre-specific nature. Moreover, though this previous research adapted the dictionary-based methodology of personification identification, the scope of it was qualitative analysis rather than quantitative exploration. Therefore, the issues of annotation and corpus building have remained unsolved. The PerSE corpus project aims at going one step further towards establishing a general language-specific corpus of personification.

3. Material and Methods

The last section summarized the theoretical challenges of identifying and annotating personifications in a corpus. After proposing solutions to some problems, this section

[1] For assessing the level of conventionality The Concise Dictionary of Hungarian (Pusztai ed. in chief, 2003) was used in the research. See 3.2 for further details of the infrastructure of the project.

deals with the planned structure of the PerSE corpus and the linguistic material of its test version, with the steps of preprocessing the sampled texts, as well as with the method of manual annotation

3.1 The Test Version of the PerSE corpus

To explore the variability of personification in a language, we need a comprehensive collection of texts extending from literal or poetic genres to scientific, journalistic texts and everyday conversations. Consequently, the planned structure of the PerSE corpus will consist of four subcorpora: literal, scientific, journalistic, and conversational. The variety in registers and genres is important not only for obtaining a general picture about the diversity of personification but also for implementing the cognitive linguistic principle that figurative language use is not limited to literature.

For the initial steps of corpus building, however, I did not need the total amount of texts, rather a small-sized but manageable test version of it. After defining the principles of preprocessing and annotation, and implementing them successfully, the test version of the corpus will be able to be extended until reaching its final size and having its planned structure.

Thus, there were only two criteria for sampling texts into the test version of the corpus: first, it had to be an online written text (to avoid a necessary transcription and/or digitalization); and secondly, it has to contain a significant number of personifications. One of the genres that fulfil both conditions is the online car review. This specific discourse type aims to give a detailed evaluation of new car models by describing their advantages and disadvantages, highlighting their capacities, and recommending them to customers. Despite the profit-oriented character of these reviews, they have a semi-professional attitude towards the models, providing the reader with technical data and often being critical of (the products of) the car industry. Moreover, in the spirit of infotainment, these reviews show a continuum of style ranging from a more distanced and objective tone to a more subjective and evaluative use of language. It is typical in this discourse type to refer to cars (or car producing companies) as human beings, either to increase the personal involvement in the topic or to avoid a formal stance on it and to express the informal but professional identity of the author. Although using personification seems to be characteristic of the genre, the more subjective and casual the style of a review, the more personifying expressions it contains.

To get a sample large enough for making general observations, six reviews[2] were sampled into the test version of the PerSE corpus, which consists of 10486 words in total. The texts were written by three different authors, thus the proportion and the patterns of personifying language use cannot be interpreted as idiolectal variations of Hungarian, although the genre-specific features of personification need further investigation.

3.2 Preprocessing of the Samples and the Infrastructure of the Project

Before starting manual annotation, the samples were preprocessed with the e-magyar Digital Language Processing System (Váradi et al., 2018).[3] The texts were tokenized, lemmatized, PoS-tagged, morphologically and syntactically parsed by the analyzer. The result of the automatic processing was exported in .conllu format to complete it with manual annotation of personifying expressions.

For manual annotation, I used the Webanno web-based annotation tool (Eckart de Castillho et al., 2016).[4]

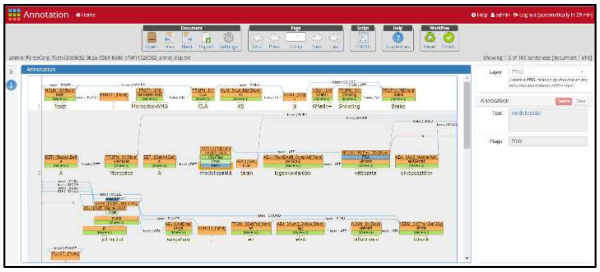

Figure 1: Annotation in preprocessed text with Webanno.

Figure 1 illustrates the process of manual annotation in Webanno. The labels allocated to the words of the text constitute two additional layers of annotation: the ptags set for the components of personification and the pqual set for the degree of its conventionality. Therefore, all tokens of the corpus have a label for the lemma, another designating the PoS category of the lemma and the morphological structure of the word form, as well as two optional tags in case of personifying usage. The syntactic dependency relations are marked with arrows and the corresponding labels on them. It is also the form of designating trajector and landmark relations between the components of the multi-word personification.

For implementing the dictionary-based method of personification identification (for the detailed discussion of the method and its adaptation, see 3.3) I used the second edition of The Concise Dictionary of Hungarian (Pusztai ed. in chief, 2003), which is the only available comprehensive and partially corpus-based dictionary for Hungarian. (The new fully corpus-based dictionary of the language is not finished yet, only eight volumes have been

[2] The reviews are available on the Totalcar website, with the following URLs: https://totalcar.hu/tesztek/2021/07/01/mercedes-amg_cla_45_s_4matic_shooting_brake_teszt/, https://totalcar.hu/tesztek/2021/09/10/mercedes-benz-c-300-limousine-amg-line-w206/, https://totalcar.hu/tesztek/2021/08/02/bemutato_hyundai_kona_n_2021/, https://totalcar.hu/tesztek/2021/07/02/hyundai-ioniq-5-teszt/, https://totalcar.hu/tesztek/2021/07/05/skoda_kodiaq_rs_2.0_tsi_dsg_4x4_facelift_bemutato_menetproba/, https://totalcar.hu/tesztek/2021/07/27/porsche_cayenne_turbo_gt_teszt_bemutato/ (last access: 04/01/2022).

[3] https://e-magyar.hu/en (last access: 04/01/2022) I would like to express my special thanks to Balázs Indig for his technical support and his contribution to the preprocessing of the samples.

[4] https://webanno.github.io/webanno/ (last access: 04/01/2022) Manual annotation has been carried out by the author of the paper. One of the planned improvements in the future is to involve independent annotators and test the reliability of the procedure.

published until now.) Although meaning descriptions in the dictionary used for the project follow professional intuition, the frequency data are based on the first version of the Hungarian National Corpus.[5]

The PerSE corpus contains information about the idiomaticity of the identified personifying expressions (for details, see 3.3). To estimate this measure, I relied on the collocation data of the Hungarian Web 2012 corpus (huTenTen12, for the TenTen corpus family, see Jakubíček et al., 2013),[6] using the logDice association score (Rychlý, 2008) with the threshold value of 6.

The data arising as the result of automatic and manual annotation were exported from the Webanno in .tsv format. The further analysis of these data was accomplished with the use of MS Excel.

3.3 Protocol for Manual Annotation of Personification

The identification process of personifying expressions in Hungarian follows the methodological proposal by Dorst and her colleagues (Dorst, Mulder and Steen, 2011). It is based on the MIPVU protocol for metaphor identification (Steen et al., 2010) and its adaptation to Hungarian (Simon et al., 2019). The identification of personifying usage of a word is a word sense disambiguating process: the analyzer defines the lexical item's basic and contextual meaning using a dictionary. The former (being given normally as the first meaning of the word in the dictionary) is more concrete and human-oriented, whereas the latter is typically more abstract, and in the case of personification it does not refer to human beings. If the contextual meaning coincides with the basic meaning, there is no need to allocate any tags to the lexical item. But if the non-human contextual meaning can be compared to human basic meaning, the lexical item can be tagged as personification.

3.3.1 The Tagsets of the Annotation

The original protocol makes it possible to identify personification at the level of lexical items without shedding light on the internal organization of multi-word expressions. Therefore, in the course of adapting the method, I made a distinction between two different layers of the annotation, establishing two sets of labels.

The ptags set refers to the linguistic components of personification with the following labels.

- **PRW**, personification-related word: the word has a personifying contextual meaning. For instance, the group of cars belonging together regarding their production is represented in a review as *modellcsalád* ('family of models'), which can be considered a personification without the contextual support of any other lexical unit.
- **PRA**, personification-related argument: the word contributes to an unfolding personifying meaning, but it does not have a personifying usage in its own right. An example of it is the nominal component of the expression *Így tol ki* […] *387 lóerőt* ('That is how it pushes out […] 387 horsepower'): the verb *kitol* ('to push out') has a human basic meaning ('to get something to outside through pushing'), but here it refers to the performance of the car's engine. Thus, the nominal form *lóerőt* ('horsepower) can be tagged as the argument of verbal personification.
- **PRWid**, idiomatic personification-related word: the lexical unit counts as personification in itself; however, it has a collocational relationship with another word according to its co-occurrence pattern in the huTenTen12 corpus. For instance, in the Hungarian expression *ki lehet hozni a sodrából* (lit. 'it can be taken out of its current', figuratively 'it can be made lose its temper') the verb *kihoz* ('to take out something from somewhere') has the contextual meaning of 'provoke somebody', and it refers to trying out a car. Moreover, the verb has a strong association (logDice = 10.8) with the nominal component of the expression *sodrából* (lit. 'out of its current', figuratively 'out of its temper'); therefore, it can be marked as an idiomatic node of a personification.
- **PRAid**, idiomatic personification-related argument: the lexical unit contributes to a personifying meaning generation through idiomatic relationship to another word. In the example above, it is the nominal form sodrából (lit. 'out of its current", figuratively 'out of its temper') which counts as an idiomatic argument of personification.
- **PRWimp**, implicit personification-related word: the lexical unit (usually a pronoun in Hungarian) has a coreferential relationship with a personification related word of the text. As an example, consider this sentence: *Érezhetően tudna az okos C-osztály magától közlekedni a gondosan felfestett és kitáblázott utakon, ha megengedné neki a jogi környezet* ('Perceivably, the smart C-Class could travel on its own on carefully painted and signposted roads if the legal environment allowed it to it'). The entity denoted by the expression *C-osztály* ('C-Class') is personified; thus, the nominal *neki* ('to it') referring back to the entity can be tagged as an implicit personification related word.

Allocating labels to the structural components of a personifying expression render it possible to mark the semantic relationships between them. The prel tagset contains the following four relational labels.

- **tr**, trajector relationship: the argument (marked as PRA or PRAid) elaborates the primary schematic figure (i.e., the agent) of the verb. The expression *a Mercedes megcsinálja* […] *ferdehátúját* ('the Mercedes produces […] its fastback') also gives a good example of the trajector relationship, since the *Mercedes* nominal (tagged as PRA) specifies the primary figure of the verb *megcsinál* ('to produce').
- **lm**, landmark relationship: the argument (marked as PRA or PRAid) elaborates the secondary schematic figure (i.e., the patient, the instrument, the recipient or other circumstance) of the verb. In the aforementioned example, the nominal argument *ferdehátúját* ('its fastback') specifies the secondary figure of the verb *megcsinál* ('to produce'), thus a landmark relation can be marked between these components.

[5] http://corpus.nytud.hu/mnsz/ (last access: 05/01/2022)

[6] https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhutenten12_hp2 (last access: 05/01/2022)

- **poss**, possessive relationship: this semantic relation is typical in the case of body-part personifications (e.g., *a repülő hátán* 'on the back of the plane'), where the figure of the human body (or one part of it) is mapped onto a physical object. Since there are no arguments in this construction (which can be modelled as a reference point configuration in cognitive grammar, see Langacker, 2013), both components are marked as PRW.
- **r**, unspecified semantic relationship: this label is used when the components of a multi-word expression occur separately from each other in the sentence because of word order patterns, such as inversion or the infiltration of auxiliaries (for the grammatical details, see Rounds, 2001). It is used only for a technical reason: it marks the connection of the elements of a discontinuous personifying expression without any specification of their relationship.

Besides the ptags and prel labels, I adopted the conventionality categories proposed by Dorst and her colleagues in the pqual tagset (Dost, Mulder and Steen, 2011). The pnov tag designates novel personifications, de pconv tag is for conventional personifications, the pdef is used to mark default personifications, and the pmet tag is for metonymical personification. (For the detailed discussion of these categories see section 2.)

3.3.2 The Protocol of the Annotation

At the end of this section, I summarize the process of manual annotation in a step-by-step manner.

1. Find personification-related words (PRWs) and arguments (PRAs) by examining the text on a word-by-word basis.
   a. When the basic meaning of a word refers to a human being, but it has a non-human contextual meaning, mark the word as PRW.
   b. When there is a strong associative relationship (logDice ≥ 6 in the huTenTen12 corpus) with another word, mark the word as PRWid.
   c. When the word contributes to a personification as the argument of another word, mark it as PRA.
   d. When there is a strong associative relationship between the word and another word marked as PRWid, allocate the PRAid tag to it.
   e. When the word is used for creating a coreferential relationship with another word in the text, and this other word is marked as PRW, mark the word as PRWimp.
2. Mark the semantic relationships between words labelled as PRWs and PRAs.
   a. When the argument of another word specifies the primary figure of it, create a tr relationship from PRW to PRA.
   b. When the argument of another word specifies the secondary figure of it, create an lm relationship from PRW to PRA.
   c. When there is a possessive relationship between two words, and the personifying meaning relies on this relationship, create a connection between the two words (being marked as PRW) with the label of poss.
   d. When two components of the personifying expression occur discontinuously, create an r relationship between them.
3. Evaluate the conventionality of the personifying meaning of the words marked as PRW based on the dictionary and allocate the corresponding tag to the word (pnov, pconv, pdef and pmet, respectively).

4. Results and Discussion

4.1 Overview of the Data

Altogether 958 ptag labels have been allocated in the test version of the PerSE corpus. It means that the relative frequency of personifying tokens in the test corpus is 9.15%. However, one token may receive more than one label, since an argument can belong to more than one verb; moreover, a lexical unit may be identified as personification in its own right (hence being marked as PRW) and also an argument of another multi-word personification. As an example, consider the following expression: *a hátsó futómű* […] *követi a kocsi orrát* ('the rear running gear […] follows the nose of the car'). The nominal component *orrát* ('[the car's] nose') can be marked as PRW because according to its basic meaning it refers to the human olfactory organ. Furthermore, the nominal can be tagged as PRA as well, since it is the secondary figure of the schematic semantic structure of the verb *követ* ('to follow'), i.e. that participant which is followed. Therefore, the annotator can simultaneously allocate two tags to the *orrát* nominal.

Taking only one allocated tag into consideration by every annotated token the total number of ptag labels is 818, which means a somewhat lower relative frequency of 7.81%. In other words, almost 8% of the words in the test corpus instantiates a personification or at least contributes to it. Although there are differences between the reviews in the corpus, the overall picture does not change significantly if we zoom in on the individual texts. The result of the manual annotation is presented in Table 1.

Number of the review	Size of the review (in tokens)	Number of tokens tagged as ptag	Relative frequency (%)
R1	2190	152	6.94
R2	1577	145	9.19
R3	1536	152	9.90
R4	2148	144	6.70
R5	1535	111	7.23
R6	1482	114	7.69

Table 1: The frequency of ptag labels in the corpus.

Focusing on the ptag tagset, PRA labels have the highest proportion in the sample. The second most frequent tag is the PRW, on average every word marked as PRW has at least one annotated argument. This observation supports the claim that the typical structure of a linguistic personification consists of more than one word. Idiomatic personifications have much fewer occurrences in the corpus (with only 5% altogether), and the proportion of implicit personifications does not take 1% of all allocated tags. Figure 2 illustrates the distribution of ptags in the corpus.

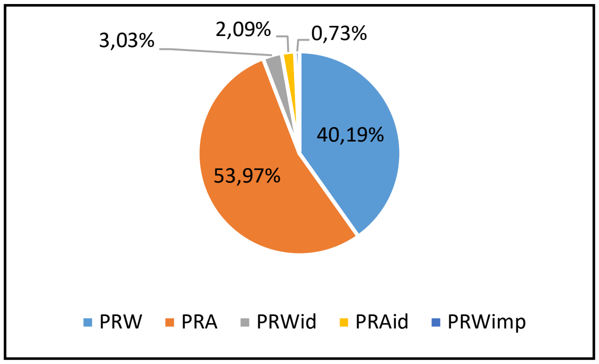

Figure 2: The proportions of ptag labels in the corpus.

Turning to the pqual layer, it can be claimed that novel personifications dominate the sample: more than half of the identified expressions belong to this category. The number of conventional personifications takes only a quarter of all labels. Default personification is much less frequent than the first two types, and metonymic personification has the lowest proportion in the corpus. The results show that the style of online car reviews does not only abound with personifications, but these figurative expressions are creative (non-conventional) in the majority of the cases. Figure 3 presents the exact percentages of the categories.

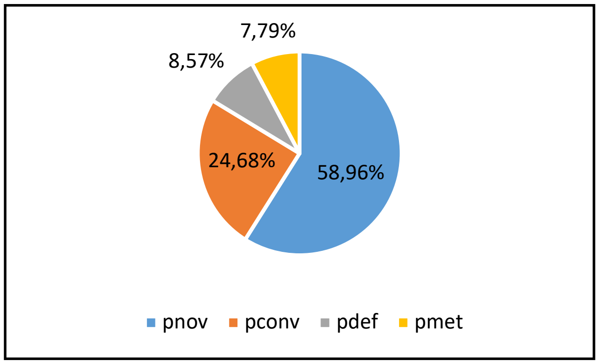

Figure 3: The proportions of pqual labels in the corpus.

4.2 The Lexical Pattern of Personifications

One of the data types emerging from the annotation is the lexical pattern of personifications in the genre of online car reviews, i.e., the lexical units frequently reoccurring in the texts as personifications. The table below presents the twenty most frequent lemmata, with their raw frequency in the second column, the number of the texts they occur in (FreqT) and their pqual category. (In the case of PRA annotation and verbal prefixes the item does not receive a pqual tag at all.)

Lemma	**Freq**	**FreqT**	**pqual tag**
tud ('know, can')	20	6	pmet, pnov
ki ('out')	10	4	–
motor ('engine')	9	5	–
erős ('strong')	8	5	pconv
meg (perfectivizing verbal prefix)	8	5	–
tart ('keep, hold')	8	5	pmet, pnov
segít ('help')	7	3	pnov
dolgozik ('work')	6	5	pconv, pnov
autó ('car')	6	3	–
maga ('him/her/itself')	5	4	pconv
csinos ('pretty')	5	3	pconv
minden ('all')	5	3	–
orr ('nose')	5	3	pconv
ő ('he/she/it')	5	3	pdef, pmet
tesz ('do')	5	3	pconv, pmet, pnov
okos ('smart, intelligent')	4	4	pnov
el ('away')	4	3	–
fenék ('bottom')	4	3	pnov
lóerő ('horsepower')	4	3	–
rendszer ('system')	4	3	–

Table 2: The most frequent lemmata in the test corpus.

It is not surprising that entities belonging to cars are on the list: they are the personified objects in the discourse (*autó* 'car', *motor* 'engine', *rendszer* 'system') or the arguments of a personification (*lóerő* 'horsepower'). What is more interesting is that the most frequent personification (*tud* 'know, can') represents the technological potentialities of cars as mental capacities or skills. Another data supporting this subpattern is the adjective *okos* ('smart, intelligent') that refers to the cars as mental agents. There are verbs among the recurring lemmata denoting relatively general processes (e.g., *tart* 'keep, hold', *segít* 'help', *dolgozik* 'work'): they attribute agency to the non-human objects of the discourse. The last group of words represent cars as having a human body or figure: the adjectives *erős* ('strong') and *csinos* ('pretty') make the physical power and appearance of the cars salient, while the nouns *orr* ('nose') and *fenék* ('bottom') describe the form of the cars being similar to a human body. Regarding the conventionality of these lexical units as personifying expressions, we can claim that mental agency attribution counts as novel personification, whereas body part terms and the description of the figure of the car rather belong to conventional personifications. General agency attribution is instantiated with both novel and conventional personifications. Metonymic and default personifications are not so frequent in the lexical pattern.

4.3 Grammatical Characteristics of Personifications

The automatized preprocessing of the texts in the corpus makes sophisticated grammatical analyses possible. Due to the limitation of the length of the paper, in the present subsection, I focus only on the relationship between the PoS categories and the ptag and pqual labels, as well as on the pattern of tr and lm semantic relations, since the latter can inform us about the construction-like behavior of personifying expressions.

There is an interesting difference between the PoS patterns of ptag and pqual labels. Whereas 39.04% of all of the ptag labels was allocated to nominal tokens (with only 24.22% of verbal personifying components), verbs received the majority (51.95%) of pqual labels (with only 14.29% of nominal forms and 23% of adjectives being tagged as one category of pqual). It is worth remembering that only words marked as PRW, PRWid or PRWimp can receive a pqual evaluation according to the annotation protocol (see 3.3.2). Thus, the results show that although the most frequently tagged personifying components are nominal tokens in the corpus, verbs and adjectives constitute the node of personifications in the vast majority of the cases (74.29% altogether). Moreover, while 15.86% of all nouns in the test corpus was tagged as a component of personification, this proportion is 19.27% by the verbs. At the layer of pqual, only 2.33% of all nominal tokens received a label, however, 16.61% of verbal tokens was allocated as one type of personification. (The proportions of adjectives were relatively low in both respects: 6.81% and 5.3%, respectively.) These findings support again the thesis of personification as a multi-word expression on the one hand, and on the other hand, it may contribute to the development of a semi-automatic annotation of personifications based on PoS tagging.

On the grounds of these results, it is not unexpected that in the group of PRWs the verb is the most frequent category (48.83%), with the adjective in the second and the noun in the third place of the list (23.12% and 17.66%). In contrast, among PRAs nouns are dominant (with 56.09%), followed by pronouns (17.79%) and proper names (12.38%). The pattern is similar in the realm of idiomatic personifications: whereas verbs (65.52%), adjectives (17.24%) and adverbs (6.90%) are the three most frequent PRWid categories, in the case of PRAid nouns (80%), pronouns (15%) and adverbs (5%) make the list. Implicit personifications are instantiated mainly with pronouns and their adverbial derivations (85,72% altogether). Put it simply, verbs, and adjectives constitute the most salient personifications in the corpus, while nouns, pronouns and names are typically arguments of a personifying expression.

There is further evidence for the salience of verbal and adjectival personifications. In all four groups of the conventionality dimension, these are the most frequent PoS categories. Table 3. summarizes the results.

PoS	**pnov (%)**	**pconv (%)**	**pdef (%)**	**pmet (%)**
verb	58.15	42.11	39.39	50
adj	17.18	32.63	30.30	26.67
noun	13.22	18.95	15.15	6.67

Table 3: The distribution of PoS categories at the layer of pqual.

The last aspect of the grammatical analysis concerns the semantic relations marked between the node and the arguments of a multi-word personification. Since possessive relations occur infrequently in the corpus (there are 17 of such labels in total), I concentrate here on trajector and landmark relations. The first was allocated 236 times in the corpus, while there are 198 lm tags in it. A plausible explanation of the higher amount of trajector relations lies in the number of adjectives and adverbs among personifications. These structures (amounting to 15.56% of all ptag labels) function as modifiers or adverbials in the clause having a modified or specified nominal argument, which refers typically to the personified entity elaborating the primary figure (i.e., the trajector) in their schematic semantic structure. Therefore, the high number of adjectival and adverbial personifications also increases the number of trajector labels.

From the distribution of semantic relations, three typical constructions of personification can be abstracted. The first is centered around a personifying verb, with a nominal argument elaborating its primary figure (the personified entity) and one or more other arguments specifying the event structure of the verb (e.g., [*a biztonsági rendszer*] *mindenre halálosan figyel* '[the security system] pays attention to everything'. The second consists of two components: an adjective or an adverb (providing the conceptual frame of personification) and a nominal argument specifying the primary figure of the adjective/adverb as the personified entity (e.g., *cinikus reménytelenség* 'cynical hopelesssness'). The third construction is the nominal personification (typically body-part expressions), in which two nouns are connected via a possessive relationship (e.g., *egy repülő hátán* 'on the back of an airplane'). The least frequent case is when there is an isolated, individual word initiating a personifying meaning without involving further components in the clause (e.g., describing a car as being *powerful*/*strong*).

5. Conclusion and Future Perspectives

The present paper provided a detailed description of the actual phase of the PerSECorp project. It has the aim of performing a systematic analysis of personifications in Hungarian applying corpus linguistic methodology. This analysis is built on the PerSE corpus, which contains both general linguistic information (e.g., PoS labels, morpho-syntactic analysis) and the annotation of personifications. The corpus is planned to have four subcorpora (literary, journalistic, scientific, and everyday discourses).

The present test version of it consists of 6 online car reviews (10468 words in total). The tokenization, lemmatization, PoS tagging, morphological and syntactic parsing of the linguistic material has been carried out with the e-magyar NLP tool. For manual annotation, I elaborated the protocol of identifying personifications with two tagsets (one is for marking the linguistic components and the other is for evaluating the conventionality of personifying meaning) and implemented it in Webanno. As a result of the annotation process, both the lexical pattern and the grammatical characteristics of personifications in Hungarian have been explored.

There are multiple possibilities for the further development of the PerSECorp project. First of all, the fine-grained morpho-syntactic annotation of the texts can be exploited to obtain mode detailed information on the specific constructions of personification in Hungarian. An additional opportunity is to extend the scope of annotation to the conceptual domains of personification improving a process that relies on other existing language resources (e.g., lexical-semantic databases for Hungarian). Of course, the main direction of the development is to enlarge the present version of the corpus establishing its final structure and scale. The linguistic knowledge accumulating in the course of corpus building will certainly serve as a vantage point for finding solutions to the challenge of automatic personification identification.

6. Acknowledgements

The project and this paper were supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00382/21).

7. Bibliographical References

Dorst, A. G. (2011). Personification in discourse: Linguistic forms, conceptual structures and communicative functions. *Language and Literature* 20 (2):113–135.

Dorst, A. G., Mulder, G. and Steen, G. J. (2011). Recognition of personification in fiction by non-expert readers. *Metaphor and the Social World* 1 (2):174–201.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the LT4DH workshop at COLING 2016.* Osaka, Japan, pp. 76–84.

Goldberg, A. (2006). Constructions at Work. The nature of generalization in language. Oxford, New York: Oxford University Press.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. and Suchomel, V. (2013). The TenTen corpus family. In A. Hardie and R. Love (Eds.), *Proceedings of the 7th International Corpus Linguistic Conference CL*. Lancaster: UCREL, pp. 125–127.

Kövecses, Z. (2010). Metaphor. A Practical Introduction. New York: Oxford University Press, 2nd edition.

Lakoff, G. (2006). The contemporary theory of metaphor. In D. Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings*. Berlin, New York: Mouton de Gruyter, pp. 185–238.

Langacker, R. W. (2013). Essentials of Cognitive Grammar. New York: Oxford University Press.

Long, D. (2018). Meaning construction of personification in discourse based on conceptual integration theory. *Studies in Literature and Language* 17 (1):21–28.

Low, G. (1999). "This paper thinks…": Investigating the acceptability of the metaphor AN ESSAY IS A PERSON. In L. Cameron and G. Low (Eds.), *Researching and Applying Metaphor*. Cambridge: Cambridge University Press, pp. 221–248.

Melion, W. S. and Ramakers, B. (2016). Personification: An Introduction. In W. S. Melion & B. Ramakers (Eds.), *Personification. Embodying Meaning and Emotion*. Leiden, Boston: Brill, pp. 1–41.

Panther, K-U. and Thornburg, L. L. (2007). Metonymy. In D. Geeraerts and H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics*. New York: Oxford University Press, pp. 236–263.

Pusztai, F. (Ed. in chief), The Concise Dictionary of Hungarian. Budapest: Akadémiai Kiadó.

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka, A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN*. Brno: Masaryk University, pp. 6–9.

Rounds, C. (2001). Hungarian. An essential grammar. London, New York: Routledge.

Sájter, L. (2008). Megszemélyesítés [Personification]. In I. Szathmári (Ed. in Chief), *Alakzatlexikon* [Lexicon of figures of Speech]. Budapest: Tinta Könyvkiadó, pp. 383–388.

Simon, G. (2021). The event structure of personification in the poetry of Attila József. In J. Tóth and L. V. Szabó (Eds.), Ereignis *in Sprache, Literatur und Kultur* [*Event* in Language, Literature and Culture]. Berlin: Peter Lang, pp. 67–79.

Simon, G., Bajzát, T., Ballagó, J., Havasi, Zs., Roskó, M. and Szlávich, E. (2019). Metaforaazonosítás Magyar nyelvű szövegekben: egy módszer adaptálásáról. [Metaphor identification in Hungarian texts: a methodological adaptation.] *Magyar Nyelvőr* 143 (2):223–247.

Steen, G. J., Dorst, A. G., Herrmann, B. J., Kaal, A. A., Krennmayr, T. and Pasma, T. (2010). A Method for Linfuistic Metaphor Identification. From MIP to MIPVU. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Thornborrow, L. and Wareing, S. (1998). Patterns in Language. An introduction to language and literary style. London and New York: Routledge.

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R. and Vincze, V. (2018). E-magyar – A Digital Language Processing System. In N. Calzolari, K. Choukri, Ch. Cieri, Th. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris: ELRA (European Language Resources Association), pp. 1307–1312.

Appendix

Abbreviation	Term
ptag	tags for the structural components of personifications
PRW	Personification-related Word
PRA	Personification-related Argument
PRWid	Idiomatic Personification-related Word
PRAid	Idiomatic Personification-related Argument
PRWimp	Implicit Personification-related Word
prel	tags for the semantic relations within multi-word personifications
tr	trajector (primary focal figure)
lm	landmark (secondary focal figure)
poss	possessive relationship
r	unspecified semantic relationship
pqual	tags for the conventionality (i.e., quality) of personifications
pnov	novel personification
pmet	metonymical personification
pdef	default personification
pconv	conventional personification

Table 4: Glossary of the abbreviated terms