The Register-specific Distribution of Personification in Hungarian: A Corpus-driven Analysis

Gábor Simon

Abstract

Although several promising initiations have been proposed recently about how to identify personifications, a comprehensive corpus linguistic analysis of personifying meaning generation still have to be carried out. The aim of the paper is twofold: (i) to present an extended version of the PerSE corpus, the language resource for investigating personification in Hungarian; (ii) to explore the semantic and lexicogrammatical patterns of Hungarian personification in a corpus-driven analysis, based on the current version of the research corpus. PerSE corpus is compiled from online available Hungarian texts in different registers including journalistic (car reviews and reports on interstate relations) and academic discourse (original research papers from different fields). The paper provides the reader with the infrastructure and the protocol of the semi-automatic and manual annotation in the corpus. Then it gives an overview of the register-specific distribution of personifications and focuses on some of its lexicogrammatical patterns.

1 Introduction

Despite its apparent clarity, the category of personification is far from being simple and homogeneous. In the last decades, at least four different conceptual models of personifying meaning-making have been proposed in cognitive linguistics. Beyond the general metaphorical explanation (personification is an ontological conceptual metaphor with a human being as its source domain, see Kövecses, 2010) there is an alternative model within the framework of conceptual metaphor theory (based on the EVENTS ARE ACTIONS generic-level metaphor, see Lakoff, 2006), but a metonymic (Low, 1999) and a conceptual integration model (Long, 2018) are also available in the literature. Moreover, a solid methodological framework for identifying personifications in texts has been proposed by Dorst et al. (2011). However, systematic and extended research on the linguistic variability of personification has not been carried out yet. Although the protocol for identification may serve as a promising vantage point for a comprehensive corpus study, there is not any available language resource in terms of personification annotation. The present paper aims to fill this gap by proposing an extended version of the PerSE[1] corpus, a new language resource for studying personifying language use in Hungarian. Beyond merely demonstrating the corpus, some initial analyses of the register-specific patterns of personification in Hungarian are provided here, too.

The study is based on the extended and improved version of the PerSE corpus introduced previously (Simon, 2022). The former study provided the reader with the annotation protocol, the basic infrastructure of the corpus and some preliminary results of personification identification in a pilot sample of only one register. Compared to it, this paper demonstrates the annotation of personifications on a relatively larger scale (analyzing three different registers) and with advanced infrastructure. This modest expansion of the corpus made it possible to consider the register-specificity of personifying language use. Consequently, the scope of the study also encompasses the quantitative analysis of

[1] The name of the corpus is the abbreviation of the phrase „Personifying Structures Encoded”.

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 99–109

personifying language use within and between registers. For the latter, both a whole corpus design and a linguistic feature design (Brezina, 2018) have been implemented.

The paper is structured as follows. After the introduction, the basic notions and principles of the analysis are discussed (2). Then the material and the methodology of the study are detailed, including corpus building, annotation and quantitative analysis (3). The fourth section deals with the results of the analysis, and the paper ends with some concluding remarks (5).

2 Theoretical Background

According to the glossary of Kövecses's volume on conceptual metaphor theory, personifications "involve understanding nonhuman entities, or things, in terms of human beings. They thus impute human characteristics to things" (Kövecses, 2010). This very basic definition needs to be detailed with the further aspects of personifying meaning-making: it attributes agency to non-human entities (Dorst, 2011), it can rely on the metonymic link between human and non-human entities, and it can be conventionalized in different degrees ("dead personifications", see Dorst, 2011). Therefore, personification as a semantic phenomenon is much more complex than it is implied in its definition.

Dorst et al. (2011) operationalize the notion in the following way: if the basic meaning of a lexical unit is human-oriented (i.e., the primary figure of the meaning is typically a human being), and the contextual (or actual) meaning of the unit refers to a non-human entity, it can be labelled as personification. By way of explanation, identifying personification in discourse is a specific process of word sense disambiguation, rendering it possible not only to highlight personifications in a text but also to categorize them in terms of conventionalization. If both the human basic meaning and the non-human contextual meaning are offered by the dictionary, the personification can be considered a conventionalized one. In (1) invasion is a military process in its basic meaning, but it has a more general meaning in the dictionary as well ('Someone or something appears somewhere en masse'), therefore the personifying usage of the noun is conventional. However, if only the human basic meaning can be found in the dictionary, the expression is rather a novel personification. In (2) (referring to an engine of a car) greedy has only a human-related meaning in the dictionary ('[someone] trying to satisfy their desire ardently'), thus the personifying usage of the adjective has not been lexicalized yet. In the case of not referring directly to human beings in the description of the basic meaning by the dictionary, but the prototypical or default figure is human, the personification belongs to the default type. The basic meaning of develop (3) is '<living being> grows, their features evolve gradually', in which the primary figure is not explicitly a human being, but in its default interpretation, it refers typically to people, therefore it has a default personifying usage. Finally, (4) illustrates metonymic personification with the reference of Russia to the leaders or the members of the Russian army.

(1) biológia-i invázió
biology-ADJ invasion
'biological invasion'

(2) nem egy mohó szerkezet
not a.DET greedy gear
'not a greedy gear'

(3) harc-művészet-i hagyomány-ok fejlőd-t-ek
fight-art-ADJ tradition-PL develop-PST-3PL
'[the] traditions of martial arts developed'

(4) Oroszország helikopter-t veszít-ett
Russia helicopter-ACC loose-PST.3SG
'Russia lost [a significant amount of] helicopters'

This lexical semantic approach to personification provides a solid theoretical foundation for a corpus-driven analysis of the linguistic patterns of personifying language use, ensuring such a scope that is broader than in previous research. Low (1999), for instance, considers metonymic and personifying readings as alternatives in meaning-making. In the cognitive linguistic research of the discourse on interstate relations (Twardzisz, 2013) metonymy is completely excluded from the realm of personification. As a consequence of the latter decision, the personifying use of state names proves to be rather infrequent in the journalistic corpus of the previous analysis. However, the present study adopts such an operationalization of the notion of personification that results in a better recall of the corpus analysis without decreasing the level of precision.

The chosen theoretical and methodological orientation is based on the MIPVU protocol for

metaphor identification (Steen et al., 2010), and since a systematic and comprehensive analysis sheds light on the register-specific patterns of metaphorization in English, we can assume that personifying language use will also show different realizations in terms of discourse types in Hungarian. As Steen et al. (2010) observe, the academic register has the highest percentage of metaphor-related words, while fiction is only the third on the list regarding the frequency of metaphors. News texts are almost as metaphorical as academic papers, and conversation is the least metaphorical. Compared to the previous research, the PerSE corpus represents two broad fields of discourse on a higher level of granularity than the Amsterdam Metaphor Corpus: from journalism, I analyzed two specific registers (car reviews and reports on interstate relations), while from academic discourse I sampled original research papers from a wide range of scientific fields, including natural and social sciences.

Considering the linguistic structure of personification, we can claim that it is not limited to only one word in the discourse. In a previous experimental study, Dorst et al. (2011) found that 61.90% of the personifications identified by the informants were word combinations. From a rather theoretical point of view, Long (2018) defines personification as an "extended unit of meaning" (relying on Sinclair's notion), encompassing a node word and its collocations. In my previous analysis (Simon, 2022), personification-related arguments were slightly more frequent than words related directly to personifying meaning, which means that on average every personification had at least one argument in the corpus. Consequently, the present study also focuses on personifications as potentially multi-word expressions and provides data not only on the raw frequencies of personifications in different texts but also on their size and distribution in terms of node and argument structure.

As a result of the overview of the theoretical background of the present study, the central research question is as follows: what are the differences between the patterns of linguistic personification in different registers in Hungarian? This general question can be answered from more than one perspective, regarding the distribution and frequency of personifications on the one hand (whole corpus design, see Brezina, 2018), and on the other hand taking the lexicogrammatical features of personification in different registers into consideration (linguistic feature design, see Brezina, 2018). In this paper, I apply both points of view.

3 Material and Methods

Before turning to the results of the annotation process and the corpus analysis, I introduce the language resource that served as the basis of the research: the PerSE corpus. The process of corpus building, the infrastructure of the research, the annotation protocol and the methods of the quantitative analysis are outlined in this section.

3.1 Sampling and Research Infrastructure

The overall aim of the research is to explore systematically, in a corpus-driven way how personifying meanings are symbolized in Hungarian, i.e. what are the central linguistic patterns of personification in this language. To discover these patterns, it is essential to sample texts from as wide a repertoire as possible. The pilot version of the corpus was compiled from online car reviews written in Hungarian (see Simon, 2022 for a detailed description of this version), representing a variety of personifications in Hungarian without any reference to their register-specific character. The extended version of the corpus includes Hungarian reports on interstate relations published in an online daily news site, which makes a comparison of personifications across journalistic registers possible.[2]

There was only one aspect for sampling in the interstate relations subcorpus: the article needed to describe a prominent geopolitical event or scenario in the time frame of the sampling period (from 2022 October to 2023 June). 6 reports were chosen with the topics of Italian politics (the political agenda of the Meloni government), the French-German relationship, British politics (the political agenda of the Sunak government), the war in Ukraine and the legal investigation against Donald Trump.

Moreover, the PerSE corpus contains online available Hungarian academic texts as well, namely original research papers from online journals in the following fields of research: health

2 It is worth noting that the complete version of the corpus will consist of literary fiction and conversations, too.

sciences, dentistry, hydrology, orientalism and conservation biology. Here, the criteria of sampling were more complex: (i) the journal needs to have an open access declaration; (ii) the paper needs to be published under a CC BY licence; (iii) the size of the paper needs to be short or medium; (iv) the paper needs to be published recently (from 2020 to 2023). The papers fulfilling these criteria were sampled and converted to .txt format. Diagrams, tables, figures, original quotations not written in Hungarian and references were omitted from the samples. Although the abstracts frequently repeat some sentences from the main text, they constitute an essential component of the genre, therefore the papers were sampled to the corpus with abstracts (and keywords, if there were any). Appendix A presents the size of the corpus and its subcorpora.

The plain texts sampled into the corpus were automatically processed with the e-magyar digital language processing system[3] (Indig et al., 2019) with the preset "Raw text to dependency parsing in CoNLL-U format using Stanza Dependency parser". The automatic preprocessing thus included tokenization, lemmatization, PoS tagging and morphosyntactic analysis. The results of the preprocessing were exported in CoNLL-U format. For the manual annotation of personifications, the INCEpTION platform was used (Klie et al., 2018). (The reliability test of the procedure was carried out in WebAnno (Eckart de Castillho et al., 2016)). According to the annotation protocol, I used the Concise Dictionary of Hungarian (Pusztai ed. in chief, 2003) for word sense disambiguation. To estimate the idiomaticity of linguistic personifications, the Hungarian National Corpus[4] (v2.0.5, Oravecz, Váradi and Sass, 2014) was used as a reference corpus.

The integrated result of the automatic preprocessing and the manual annotation was exported in WebAnno TSV v3.3 file format and further analyzed in MS Excel. Statistical analysis was carried out in R (v4.1.0, R Core Team, 2021).

3.2 Annotation Procedure

In this subsection, I give a brief overview of the annotation process, for a detailed description see Simon (2022). The procedure, which is based on the identification protocol proposed by Dorst et al. (2011), and borrows some elements from a MIPVU-inspired, language-specific metaphor annotation protocol (the MetaID protocol, Simon et al., 2023) as well, has the token as its basic unit, and relies on the previously described operationalization of the notion of personification.

The annotation is carried out on three layers (summarized in Appendix A). First, the components of linguistic personification are labelled (ptags). Then the annotator analyses the semantic relations (prel) between the components (if the actual target is a multi-word expression), using the basic semantic categories of cognitive grammar (Langacker, 2013): the trajector for the primary figure of a process or relation (basically, it is the agent) and the landmark for the secondary figure (the patient, experient, recipient of the process, or the element of the setting in the represented scenario). Since possessive relation is also frequent in personification (mainly in body-part constructions), it receives a distinct label. (A technical label for separated elements of a construction (e.g., preverbs) was also used in the procedure, but it is of peripheral importance, thus, I omitted it from the analysis.) Lastly, the semantic quality of personifications is classified by the aforementioned four categories (conventional, novel, default and metonymic personifications, pqual).

Since the semantic categories of personification have been discussed in section 2, here I focus rather on components and their labels. PRW refers to personification-related words, all tokens that initiate personifying meaning-making. For example, in (2) the adjective *mohó* ('greedy') personifies the concept of the engine.

PRA is for all the tokens semantically and grammatically linked to an initiator of personification (labelled as PRW), and contribute to the elaboration of a personifying meaning. As an example, the *hagyományok* ('traditions') in (3) constitutes an argument of a multi-word personification.

PRWid stands for idiomatic personification, i.e., for those units that are considered an element in a prefabricated structure in Hungarian. Prefabricatedness is measured by exploring the collocational behavior of the candidate expressions in the reference corpus (see section 3.1). Collocations are identified by the logDice score

[3] The pipeline is available here: http://emtsv.elte-dh.hu:5000/ (last access: 01/03/2024).

[4] The HNC corpus is available here: https://clara.nytud.hu/mnsz2-dev/ (last access: 01/03/2024).

(Rychlý, 2008) with a threshold of 6. PRAid refers to the idiomatic counterpart of the PRA category.

PRWimp stands for implicit personification: in this case, the token refers to some other labelled token via coreference, therefore it implicitly conveys personifying meaning. For instance, in (5) the verb *szankcionálták* ('sanctioned') refers back to the noun *törvények* ('laws') in the context, therefore it has an implicit personification in its semantic structure.

(5) szankcionál-t-ák […] bűn-cselekmény-ek-et
sanction-PST-3PL […] crime-action-PL-ACC
'[they] sanctioned criminal acts'

In the Hungarian construction, *Németország befelé fordul* ('Germany turns inward') the word *befelé* ('inward') collocates with the verb *fordul* ('turn') with a logDice score of 7.716, which means that in the expression the verb can be labelled as PRWid, the adverb *befelé* can be identified as PRAid, and the noun *Németország* ('Germany') is a simple argument of the personification. Additionally, within the idiomatic expression, there is a landmark relationship (*befelé* ('inward') symbolizes the orientation of the action as a container), whereas the nominal component explicates the trajector of the process.

The reliability of the annotation process was tested with two annotators, one of them was the author of the present paper, and the other was a university student who learned the procedure during a workshop and practiced it alone. The test was performed in one text sampled to the corpus, the annotators worked independently. The inter-annotator agreement was automatically calculated in Cohen's Kappa by the WebAnno platform. At two layers out of three, the annotation has good reliability (above or very close to the threshold of 0.8: at the layer of the components, it was 0.79, at the layer of the relations it was 1). Regarding the decision about the semantic quality of personifications, the test demonstrated a more modest agreement (with a Kappa measure of 0.68), but even in this case, the procedure seems to be tentatively reliable (Artstein and Poesio, 2008). Further improvement of the annotation protocol needs to be done in the future to improve the reliability of the process in the latter case, especially in terms of the sematic categories of personification.

3.3 Methods of Quantitative Analysis

Manual annotation is a time-consuming process, and the identification of personifications requires a lot of effort from the analyst. In the PerSE corpus, a relatively small-scale language resource, the sample sizes are low, and the distributions of the data are not normal in every case. As a consequence, only non-parametric statistical tests can be taken into consideration, if we are interested in register-specific tendencies of personifying language use in Hungarian.

As a baseline, two-tailed Wilcoxon tests were performed in pairs of the subcorpora. Then, a non-parametric one-way ANOVA was carried out (with a non-parametric post hoc test) to shed light on the between-group variability of personifications in the whole corpus. Note that only the pqual data have been tested in this way since the preliminary visualizations suggested significant differences only at this layer.

Considering the lexicogrammatical patterns of personification across registers, I focused on the four most frequent personified verbs in all three subcorpora and analyzed their register-specific personified use with Pearson's Chi-squared tests (see Brezina, 2018).

4 Results and Discussion

After introducing an extended version of the PerSE corpus, this section demonstrates why such a language resource is useful in researching figurative language use (especially in cognitive linguistics). First of all, I explore the distributions of the allocated labels in the entire corpus. The statistical testing of register-specific differences is the next focus of the analysis. Finally, I zoom in on some lexicogrammatical patterns of personification in the corpus.

4.1 The Frequency of Personifications

If we are interested in the overall tendencies of personifying language use in the corpus, we can observe that reports on interstate relations use personifications most frequently (34,759.33 per million words), then come the car reviews (with a relative frequency of 33,985.32 pmw), and the least personifying language use is characteristic of research papers (28,267.87 personifications pmw). Thus, the distribution of personifications across registers appears to be the reverse of the tendencies

of metaphorization observed by Steen et al. (2010) in a previous study: although they didn't find great differences between academic discourse and news texts in terms of metaphor use, the former proved to be the most metaphorical in their research, followed by the latter register. In the PerSE corpus, however, academic texts are relatively deficient in personifying language use compared to the two other journalistic registers. However, these numbers hide important within-group differences: one of the three texts being the richest in personifications comes from the academic subcorpus (after a car review and a report). Since natural and health sciences are overrepresented in the research paper subcorpus, and since the discussion of these topics seems to be weak in personifying language use, one possible explanation behind the overall frequency distribution is that humanities prefer personification more than sciences. Meanwhile, however, the very technical language of car reviews also gives a lot of personifications, thus, any absolute distinction would be hard to make.

The distribution of the components of personifications indicates slight differences. Regarding the relative frequencies of the allocated component labels in the corpus, it is a general tendency that the arguments outnumber the node words of personifications. But this ratio is the greatest in reports on interstate relations (2.60% : 5.24%, which means that almost 2 arguments are linked to a node on average in this subcorpus), and the lowest in car reviews (3.00% : 3.99%, the

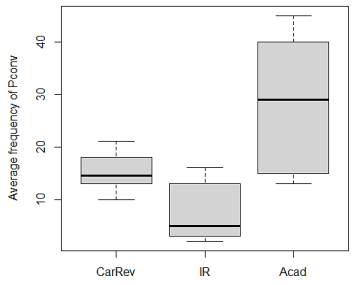

Figure 1: Box plots of the frequencies of pconv label

average argument of a node is close to 1), and the research papers are in between these two extremes (2.30% : 3.85%, the average number of arguments per node is slightly above 1.5). In other words, interstate reports provide the most extended (and elaborated) personifications.

The other issue of the distribution of the components is idiomaticity: again, reports contain the highest number of idiomatic personifications (with a relative frequency of above 0.70%); car

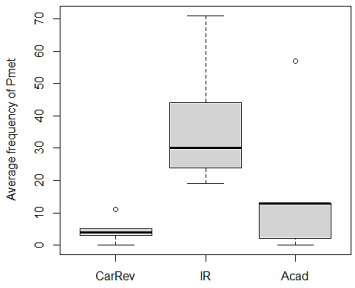

Figure 2: Box plots of the frequencies of pmet label

reviews almost lack idiomatic personifications (the average relative frequency of them is 0.22%); while research papers come close to reports in this respect (with 0.675% average relative frequency). We can claim, thus, that political journalism prefers prefabricated expressions in personifying language use the most.

The difference in allocated semantic relations between the subcorpora is not remarkable. In general, trajector labels are more frequent than landmarks in the entire corpus, which demonstrates that the personified entity (symbolized as the trajector of a process or relationship) receives linguistic elaboration in a higher percentage of the cases than other participants of the scenario. The possessive relationship reaches its maximum in car reviews, due to the preference of body-part personifications in this register.

4.2 The Semantic Quality of Personifications

Based on the observed diversity in idiomaticity in the three subcorpora, I tested first the hypothesis of whether there are statistically significant differences between registers in terms of idiomatic personifications. According to a two-tailed Wilcoxon test, interstate reports use a significantly higher number of idiomatic arguments in personifications ($W=3.5$, $p<0.05$) than car reviews, but no further significant differences could be observed either in other labels or between other registers. This means that there are no other

remarkable structural register-specific patterns of personification.

However, moving on to the semantic quality of the identified expressions, we can assume that the investigated registers significantly differ in more than one aspect.

As can be observed, default personifications are distributed evenly in the corpus, while the other three categories seem to be preferred by different subcorpora. Figures 1-3 show the frequency of the conventional, the metonymic and the novel personifications. It is clear that conventional personifications dominate in the research papers, metonymic expressions are overrepresented in interstate reports, while novel personifications belong to the register of car review.

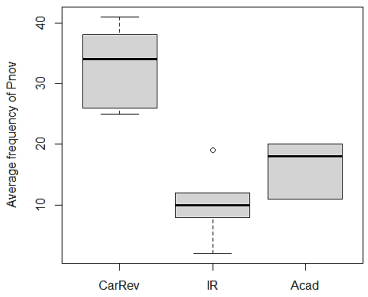

Figure 3: Box plots of the frequencies of pnov label

In the case of conventional personifications, the non-parametric one-way ANOVA did not result in any significant difference ($F(2, 6.7999)=4.7349$, $p>0.05$), thus, the register does not have a significant effect on the conventionality of personifications. However, the register affects the distribution of metonymic personifications ($F(2, 5.5681)=6.0275$, $p<0.05$), and the frequency pattern of novel personifications is affected, too, by it ($F(2, 8.5015)=18.871$, $p<0.001$). Considering the post hoc tests reports on interstate relations contain significantly more metonymic personifications than car reviews ($p<0.001$), while the latter register uses significantly more novel personification than the other two ($p<0.001$) according to a non-parametric version of the Tukey HSD test.

4.3 Verbal Personifications in the Corpus

As a language resource, the PerSE corpus provides data not only about the general distributional patterns of personification in different registers but also about the register-related personifying behavior of specific linguistic structures. While studying the frequencies of the allocated labels can be considered a top-down analysis, personifying language use can be explored from a bottom-up perspective as well, in which the frequency of personification is observed by concrete words. The latter orientation makes it possible to characterize the lexicogrammatical features of personification in Hungarian, e.g., the part-of-speech categories associated with personifying meaning, or the complexity of personifications as constructions. Due to the limitations of the present paper, I provide the reader only with a brief, rather illustrative analysis of the personifying and non-personifying use of some basic Hungarian verbs in the corpus. This analysis can be considered neither exhaustive nor comprehensive, but it may shed light on the perspectives the corpus can open for cognitive corpus linguistics.

First, relying on the verb frequency lists of the corpus, four verbs were selected for further analysis, because they belong to the most frequent verbs in all three subcorpora. The verbs are the following: *tud* (‘know/can’), *ad* (‘give’), *tesz* (‘put/do’) and *vesz* (‘take’). Then I counted all the occurrences of these verbs in the corpus, considering both their personifying and non-personifying use.

The basic tendency of all four verbs is that the non-personifying use is more frequent than personification. There are only two exceptions: *tud* (‘know/can’) is more associated with personification in car reviews, and *tesz* (‘put/do’) is rather personified in research papers. The effect size (Cramer’s V) was moderate in both cases (0.467 and 0.326, respectively). I have found a significant association only between the personifying use of the verb *tud* and register:

$X^2(2)=14.841$, $p<0.001$. Figure 4 shows the register preferences in the usage of the verb *tud*.

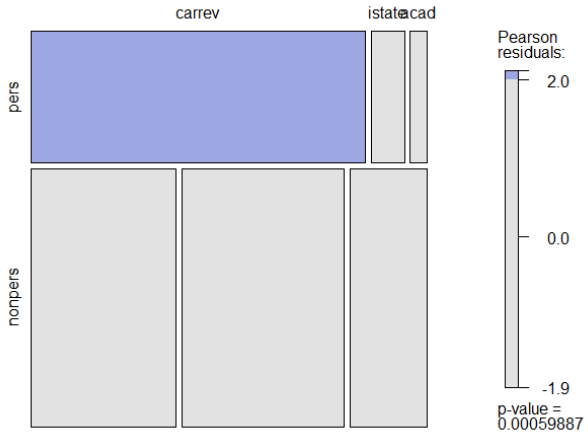

Figure 4: Mosaic plot of personifying and non-personifying use of *tud* ('know/can') in the corpus

This analysis of the linguistic features of personification across registers opens only a small window to the lexicogrammatical patterns of personifying language use in Hungarian. However, it may illustrate the potential of having a manually annotated research corpus of personifications. Moreover, it demonstrates one interesting aspect of register-specificity in personifying language use: cars (and their components) are described as human beings with physical and mental capacities in car reviews, and it is a significant linguistic pattern that cannot be found in other registers.

5 Limitations of the study

The PerSE corpus can be considered a new language resource in Hungarian that makes it possible to analyze the expressions of personification within and across different registers in a systematic way. It has, however, three major limitations that need to be addressed here. First, it is the manifestation of in-progress research, which means that other texts will be sampled into it on the one hand, and on the other, it needs to be made available for the broader research community. In its present version, it is rather a modest-scale research corpus, thus, the long-term goal of corpus compilation is to provide an open-source database designed to support cognitive corpus linguistic investigations.

Secondly, further refinement of the annotation procedure needs to be carried out to increase the reliability of the identification process and create a benchmark for cross-linguistic exploration of personification. This also means that not only do additional annotators have to be involved in the curation phase of the corpus but also that the reliability of personification identification needs to be tested via alternative empirical (experimental and/or questionnaire-based) psycholinguistic methods as well. The dictionary-based analysis demonstrates currently that the precise identification of potential personifications is feasible, but whether the annotated expressions are true personifications in actual discourse comprehension has remained an open question.

Finally, the corpus paves the way for automatic personification detection providing a precise and comprehensive data set for training large language models to this task. However, it is not clear whether these models would gain enough information from the corpus to produce good results, and if so, how this development could contribute to the current NLP field. Nevertheless, personification is a pervasive phenomenon in discourse, which makes its identification a good start for improving text classification or observing the patterns of how (mental) health issues or other negative factors are construed figuratively in everyday life.

6 Conclusion

Personification is a complex phenomenon in terms of both conceptualization and linguistic organization. Thus, cognitive and corpus linguistics need to cooperate in exploring the functioning of personifying meaning-making. The PerSE corpus is a unique language resource for analyzing personification in Hungarian from a

corpus-driven perspective. With its extended size (exceeding 30,000 tokens), genre and register variability (including technical, political and scientific language use, but also formal and informal styles in three different registers) and hybrid annotation design (extending to automatic preprocessing of grammatical features and manual identification of personifications as well) the PerSE corpus provides the analyst with a vast amount of information on personification, making it possible to approach it from different theoretical perspectives and with a wide range of methods. The present paper introduced the new, extended version of the corpus, outlining its methodological framework, and demonstrated how to exploit this language resource in a corpus-driven analysis.

The corpus annotation relies partly on automatic language processing, and hence the texts in the corpus can be analyzed on different levels of granularity (from lexical density based on tokenization and lemmatization to word class categories, and morphological and syntactic analyses). The identification of personification is based on the operationalization of the notion proposed in the literature. The dictionary-based word sense disambiguation maximizes the transparency of the annotation while minimizing its intuitive nature. Moreover, the protocol extends the task of identification to measuring the idiomatic character of personifying expressions and allocating semantic relation labels in the corpus. Thus, not only the lexicogrammatical patterns of personification can be observed but also their internal semantic organization and their prefabricatedness. This line of analysis can lead us toward the exploration of the construction-like behavior of personification in Hungarian in the future.

Compared to its pilot version, the extended PerSE corpus sheds new light on the language-internal variability of personification as well. The most important findings in this regard are as follows. (i) Journalistic registers use more personifications than academic discourse (although register-specificity is assumable based on the observations). (ii) Personifications in academic texts and interstate reports appear to be more complex in their linguistic structure with a stronger tendency to use idiomatic patterns of Hungarian. (iii) Register has a significant effect on the semantic quality of personification: interstate reports prefer metonymic personifications whereas car reviews exploit the potential of novel personifications. (iv) Some frequent Hungarian verbs are associated more with personifying language use in particular registers (e.g., the verb *tud* ('can/know') in car reviews).

The PerSE corpus also provides a solid methodological grounding for an even more extended analysis of Hungarian personifications in the future. The closest aim of the author of the present paper is to sample literary texts into the corpus and test the well-known assumption that literature would be the richest source of figurative language use. Additionally, the corpus may serve as an input data set for improving large language models in the direction of detecting and automatically identifying personification in language. In other words, the PerSE corpus as a reliable language resource with precise and multi-faceted processing of lexicogrammatical features can be used as a corpus for training existing NLP resources in Hungarian toward automatic personification annotation. Finally, the design of the corpus and the identification protocol may serve as a vantage point for creating other similar language-specific resources in various languages, bringing personification onto the top of the agenda of cross-linguistic cognitive corpus analyses. This would motivate the reinterpretation of personification as figurative language use evaluating it not as a subtype of metaphor but rather as a complex and colorful phenomenon, which is worth investigating in its own right.

Acknowledgments

The research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00382/21).

References

Ron Artstein, and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4):555–596.

Vaclav Brezina. 2018. *Statistics in Corpus Linguistics. A Practical Guide.* Cambridge University Press, Cambridge, UK.

Aletta G. Dorst. 2011. Personification in discourse: Linguistic forms, conceptual structures and communicative functions. *Language and Literature* 20(2):113–135. https://doi.org/10.1177/0963947010395522

Aletta G. Dorts, Gerben Mulder, and Gerard J. Steen. 2011. Recognition of personifications in fiction by

non-expert readers. *Metaphor and the Social World* 1(2):174–201. https://doi.org/10.1075/msw.1.2.04dor

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chirs Biemann. 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.

Gábor Simon. 2022. Identification and Analysis of Personification in Hungarian: The PerSECorp project. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022).* Marseille, France: European Language Resources Association, pp. 2730–2738.

Gábor Simon, Tímea Bajzát, Júlia Ballagó, Zsuzsanna Havasi, Emese K. Molnár, and Eszter Szlávic. 2023. When MIPVU goes to no man's land: a new language resource for hybrid, morpheme-based metaphor identification in Hungarian. *Lang Resources and Evaluation* (2023). https://doi.org/10.1007/s10579-023-09705-9

Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai. 2019. One format to rule them all. The emtsv pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop.* Florence, Italy: Association for Computational Linguistics, pp. 155–165.

Jan-Christoph Klie, Michael Burgert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018).* Santa Fe, New Mexico, USA, pp 5–9.

Zoltán Kövecses. 2010. *Metaphor. A Practical Introduction.* Oxford University Press, Oxford, UK.

George Lakoff. 2006. The contemporary theory of metaphor. In Dirk Geeraerts (Ed.), *Cognitive Linguistics: Basic Readings.* Mouton de Gruyter, Berlin, GE and New York, NY, pp. 185–238. https://doi.org/10.1515/9783110199901

Ronald W. Langacker 2013. *Essentials of Cognitive Grammar.* Oxford University Press, New York, NY.

Deyin Long. 2018. Meaning construction of personification in discourse based on conceptual integration theory. *Studies in Literature and Language* 17(1):21–28. http://dx.doi.org/10.3968/10361

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

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).* Reykjavik, Iceland: European Language Resources Association, pp. 1719–1723.

Ferenc Pusztai (Ed. in chief). 2003. *The Concise Dictionary of Hungarian*. Akadémiai Kiadó, Budapest, HU.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN.* Masaryk University, Brno, CZ. pp. 6–9.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, AU. URL https://www.R-project.org/.

Gerard J. Steen, Aletta G. Dorst, Berenike J. Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijnjte Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU.* John Benjamins, Amsterdam, NL and Philadelphia, PA. https://doi.org/10.1075/celcr.14

Piotr Twardzisz. 2013. *The Language of Interstate Relations. In Search of Personifications.* Palgrave Macmillan, Houndmills, UK and New York, NY. https://doi.org/10.1057/9781137332707

A The PerSE Corpus: numbers and labels

Register	No. of texts	Size in tokens
Car reviews	6	10,468
Reports on interstate relations	5	7,938
Research papers	5	11,500

Table 1: The structure of the PerSE corpus

Ptags (components)	PR W	PR A	PR Wid	PR Aid	PRW imp
Prel (relations)	tr	lm	poss	r	
Pqual (qualities)	pco nv	pnov	pdef	pme t	

Table 2: The layers of the manual annotation

B Personifying Usage of the Most Frequent Verbs in the Corpus

Verb	Car reviews	Interstate reports	Research papers
tud ('know/can')			
personifying	20	2	1
non-personifying	17	19	9
ad ('give')			
personifying	3	2	5
non-personifying	12	3	8
tesz ('put/do')			
personifying	5	4	8
non-personifying	9	10	4
vesz ('take')			
personifying	2	2	5
non-personifying	10	3	6

Table 3: Contingency table
of (non-)personifying use of verbs in the corpus