Named Entities in Modernist Literary Texts: The Annotation and Analysis of the May68 Corpus

Andrejka ŽEJN
ZRC SAZU

Mojca ŠORLI
ZRC SAZU

This paper is a follow-up and elaboration of the paper published in the JTDH 2022 Conference Proceedings on manual semantic annotation of named entities based on a proposed set of annotations for a corpus of modernist literary texts. We first briefly describe the corpus and introduce the annotation scheme, then focus on the results of additional analyses, and conclude with further challenges and issues we identified with respect to established NER systems and practices of related projects. Overall, we identify several categories of proper names, foreign language elements, and bibliographic citations, but focus here on the challenges of annotating names of literary characters and place names, and provide examples of the results of preliminary analyses of these entities in the corpus.

Keywords: modernism, named entities, corpus analysis, Slovenian literature, *Tribuna*, *Problemi*, 1968

Žejn, A., Šorli, M.: Named Entities in Modernist Literary Texts: The Annotation and Analysis of the May68 Corpus. Slovenščina 2.0, 11(1): 118–137.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.118-137

1 Introduction

In literary studies, named entities are most closely associated with research on literary characters and settings. A comprehensive picture of the way characters are named in literature and how place names are used in the text was obtained beyond the renderings of automatic recognition of "Named Entities" (hereafter NEs) by manually annotating these entities in literary texts, first by analyzing the annotation process, and then the data obtained from the annotated corpus itself. In this paper we report on an attempt to identify and annotate three groups of NEs in the "Corpus of 1968 Slovenian literature May68 2.0" (the May68 Corpus, for short)[1] (Juvan et al., 2022), expanding on the analyses first presented as a conference submission (see Šorli and Žejn, 2022). Section 1 provides a brief description of the corpus and the annotation procedure, followed by Section 2 that focuses on the preliminary results of an extended analysis of personal and place names. In Section 3, we discuss the potential for future annotation tasks and improvements to the annotation scheme, as well as the optimal application of the results.

In view of the significance for the Digital Humanities of controlling a large number of texts and their vertical reading, where patterns become visible that cannot be detected with the naked eye or traditional close reading, the corpus size is often seen as a key factor. At the same time, large volumes of text require automation of corpus processing for quantitative analysis, which includes different levels of (linguistic) annotation in the first phase and allows for additional levels of semantic annotation in later phases that enrich the text with metadata. In the presented approach, however, the annotation task is performed on a small, specialized corpus that is easier to control and allows for manual annotation. The identified and manually annotated NEs are distinguished based on semantic criteria, so we consider this an example of semantic annotation.

Together with the theoretical concept, the selection of annotation material, and the definition of guidelines for the annotation process (Pagel et al., 2020), the annotation scheme presented here constitutes a model for the extended annotation of NEs in modernist periodicals,

1 Corpus of 1968 Slovenian literature May68 2.0: http://hdl.handle.net/11356/1491

certain segments of which can be applied to other corpora of literary texts. We focus on both the identified inaccuracies and the advantages of manual annotation of selected groups of NEs in our specialized corpus (for more on the theoretical background and history of automated and manual annotation of NEs, including the different approaches, see Šorli and Žejn 2022: 188-189).

1.1 The May68 Corpus of Slovenian modernist literary texts – corpus description

The May68 Corpus is the result of a project on the literature of the avant-garde and modernism in the period of the worldwide student movement associated with May 1968, whose activities are also reflected in the transformation of literature. The corpus consists of Slovenian modernist literary texts from the late 1960s to the early 1970s and was created according to special criteria defined for the purposes of corpus and stylistic research of modernist texts. The student journals *Tribuna* and *Problemi*, from which the texts for the corpus were selected, played an important role in the theoretical and literary-artistic innovations of the Slovenian student movement. The May68 Corpus 1.0 contains 1,521 texts by 198 known authors published between 1964 and 1972 in the Slovenian periodicals *Tribuna*, *Problemi* and *Problemi.Literatura*. The version May68 Corpus 2.0, which has been further edited and corrected (metadata), contains 647 additional texts from *Tribuna* and *Problemi*. The texts contain complete bibliographic data, are classified by text and language type, degree of presence of non-standard Slovenian, foreign languages, modernism, and visual elements. Author details, i.e., gender and year of birth, are included with the texts. The presence of visual elements is also marked in the corpus (48 texts).[2]

1.2 Annotation procedure

Following the automatic pre-processing (the automatic linguistic annotation included lemmas, morphosyntactic descriptions from MULTEXT-East and morphological features and syntactic annotations from Universal Dependencies) of the May68 Corpus, further manual

2 A detailed description of the corpus was provided in Juvan et al. (2021: 60–64).

annotation was performed to capture more complex linguistic (semantic) phenomena and to provide a more sophisticated annotation model for proper nouns given the recurring representational problems. Manually annotated are (foreign) language variations and registers, but the focus of the present article is on the NEs denoting persons, including cited authors (sources), geographical locations, i.e., various real and fictitious place names, organizations, and miscellaneous entities.

The annotation was implemented using the WebAnno tool (Eckart de Castilho et al., 2016). WebAnno allows annotation of one sentence at a time, which is a disadvantage for longer instances of text marked by the use of foreign language(s). Each annotation round was curated by two curators. However, reiterative annotation was not foreseen, since the primary goal at this stage was not to improve automatic annotation, but to manually annotate the specialized corpus for optimal corpus analysis and stylistic studies.

The following sections and subsections introduce the types and categories of NEs, including the dilemmas encountered in the process of annotation and the rationale behind the decisions made. With a somewhat narrower notion of NER, for the purposes of this paper we are mainly talking about categories of"proper names (personal and place names)" rather than "named entities".

1.2.1 Named entity categories and resolution

At this first stage, a model for identifying and annotating the selected NEs was put in place, with a second stage of the project envisaged in which the texts will be annotated for the use of metaphor. We also discuss the practical treatment of proper names for the purposes of corpus linguistic and stylistic research, in the hope of improving the reliability of results and NLP models. As pointed out in Beck et al. (2020), representational problems in linguistic annotation arise from five different sources (ibid., 61): (i) Ambiguity is an inherent property of the data. (ii) Variation is also part of the data and can occur, for example, in different documents. (iii) Uncertainty is caused by lack of knowledge or information by the annotator. (iv) Errors may be found in the annotations. (v) Bias is a property of the annotation system as

a whole. We list a number of relevant annotated categories, their specific character, and representational problems associated with them. We focused on some open challenges in the annotation of NEs, and in particular problems related to the functional aspects of personal proper names and place names.

There is no universally accepted taxonomy for NEs, except for some coarse-grained categories (people, places, organizations). Since we are interested in a semantically oriented annotation and prefer more informative (fine-grained) categories, we opted for a three-level NE classification as shown in Table 1 (cf. Sevščíková et al., 2007). The first level in our annotation scheme corresponds to the three basic groups: 1. Proper names, 2. Foreign language and register variations, and 3. Cited authors. These groups are labelled as 1. NAME, 2. FOREIGN, 3. BIBLIO, respectively, with the first two further subdivided. The second and third levels provide a more detailed semantic classification. NE resolution is primarily linked to the category PER, which is labelled in terms of whether the character is text/plot-internal or -external. The NAME group includes the following types and subtypes:

- Person (PER), including the adjective derived from a person's name, is subdivided into fictional literary characters (PER- LIT), real characters referring to existing and historical or mythological persons or beings (PER-REAL), literary characters bearing a descriptive name (PER-DES), and members of national and social groups (PER-GROUP).
- Geographical location (GEO) comprises localities in Slovenia (GEO-SI), the former Yugoslavia (GEO- YU), Europe (GEO-EU), and in other countries, including fictitious place names (GEO-ZZ).
- Organizations and institutions (ORG).
- Miscellaneous (XXX).

Once the annotation process was completed, the labels were converted to TEI encoding in WebAnno.5 Following the conversion all proper names (personal names, place names, names of organizations) were labelled with <name>, then divided into types with @person, @ geo, @misc, @personGrp, and @org attributes, three subtypes for literary characters (@literary, @descriptive, @real), and for geographical names (@SI, @EU, @ZZ and @YU).

PERSON (PER) type is divided into PER-LIT, PER-REAL, PER-DES and PER-GRP. While the first three are categorized as subtypes of the same type, PER-GRP is defined as an independent type. As shown in Table 1, the most important NE resolution consists in the subdivision of the PER type (within the NAME group) into real, e.g., historical or real-life, persons appearing in the text, and fictional characters, each of which, however, is further specified according to semantic criteria. We have classified both historical and mythological names as non-fictional, that is, as PER-REAL, unlike, for example the Netscape project where variants of "legend", "mythological" and "fictional" are all subsumed under "fictitious" (cf. de Does et al. 2017, p. 364). The PER type includes names of people (and provisionally for pets), nicknames, pseudonyms, members of national and social groups.

Table 1: *The main categories of the May68 annotation scheme (WebAnno)*

Group	Type	Subtype	Description
NAME	**PERSON – PER**	**PER-REAL**	Real: Characters referring to real, i.e. existing and historical or mythological persons or beings, e.g. *Greta Garbo*, *Charlie Brown*, *hlapec Jernej*, *Maruška*
		PER-LIT	Literary: Fictional literary characters, e.g. *Ančika*, *Zobec*, *Janko*, *Polona*
		PER-DES	Descriptive: Literary characters that carry a descriptive name (e.g., *dolgolasec*, Eng. the long-haired guy)
		PER-GRP	Group: Members of national and social groups, e.g. *Kranjci*, *Slovenec, Američan*
	GEO	**GEO-SI**	Slovenia, e.g. *Ljubljana*, *slap Savica*, *Crngrob*
		GEO-YU	Former Yugoslavia (except for Slovenia), e.g. *Zagreb*, *Dajla*
		GEO-EU	Europe, e.g. *Frankfurt*, *Minsk*, *Vltava*
		GEO-ZZ	Other, e.g. *Peking*, *Kuba*, *Indija Koromandija*
	ORG	**–**	Names of organizations, institutions (e. g. *Klub nepismenih*, *Slovenska matica*, *Državna varnost*)
	XXX	**–**	Common proper nouns, including titles of books and other art works, artefacts, etc., e.g. *Rdeča kapica*, *Empire State Building*

Group	Type	Subtype	Description
FOREIGN	HBS	–	Serbo-Croatian
	EN	–	English
	DE	–	German
	FR	–	French
	IT	–	Italian
	LA	–	Latin
	XX	–	Other
	DIALECT	–	Dialect
	VERNACULAR	–	Vernacular
	SLANG	–	Slang
BIIBLIO	**–**	**–**	Quoted authors (Sources)

PER-REAL denotes both real, i.e. existing, persons and historical or mythological figures that are basically identifiable in encyclopaedic sources such as online lexicons of proper names, Wikipedia and the like. URL is an additional attribute of the NAME group and is given as a relevant source of information, such as a website, for a group of people appearing in the literary text. The assignment of an URL depends on the context or on extra-linguistic knowledge; we linked names to web resources only when a (personal) name was not assumed to be part of today's common cultural knowledge (e.g. Giorgio Albertazzi, Italian actor and director, or Dave Brubeck, American jazz pianist and composer), if a person can be assumed to be part of common (cultural) knowledge (Descartes, Nietzsche), we chose not to enrich the corpus with encyclopaedic data. All standard personal proper names are labelled as NAME and assigned to one of the closed subtypes.

The label PER-GRP with no subtype is assigned to members of a particular social group, most often nationality (*Slovenec*, *Nemec*), regional (*Kranjci*, *Štajerci*) or family (*Novakovi*) identity, but also smaller social groups defined on the basis of occupational or other criteria (*esesovec*, *vaščan*, *vojak*).

Of the categories introduced specifically for the purposes of the May68 Corpus, NAME / PER-DES proved, as expected, to be the most challenging subcategory. This group of names seems to be used to describe the personality and/or physical appearance of literary characters

(*govorancar*, *starec*, *brkati*), as well as their occupation (*načelnik*, *inšpektor*) or social status (*neznanec*).

Adjectives derived from personal proper nouns are annotated as the corresponding proper nouns, e.g., *Dimitrijev* (Dimitrij's), *Prešernov* (Prešeren's), *dolgolaščev* (pertaining to the long-haired one). Their derived character is revealed by morpho-syntactic tagging.

Given their statistical importance in the context of NER, the same annotation rules apply here as for characters in plays when they do not require special treatment with respect to their function. The labelling of personal names in plays depends on the status and/or function of the name. Names of individual characters that merely announce an individual character's speech, and thus his/her lines of dialogue, have not been annotated, while names in descriptions of their physical actions or behaviour are treated as ordinary proper names on the model of "sb does sth", etc. (*Pandolfo se oglеduje v zrcalu* / Pandolfo looks at himself in the mirror).

Compared to the categories of personal names, significantly fewer dilemmas occurred in the categorization and labelling of place names. Individual unresolved cases (e.g., fictitious places, names referring to localities or objects in space) were assigned to the category "Other".

Geographical names in the broadest sense spanning from names of streets (*84. ulica*), rivers (*Drava*), mountains (*Učka*, *Himalaja*), cities (*Piran*, *Rim*, *Dunaj*) to those of countries (*Slovenija*, *Japonska*) and continents (*Evropa*, *Južna Amerika*), but also abstract (e.g. space-related) or fictitious (text- or plot-internal) (planet *Tuku-Luka*) place names, were taken into account in the manual annotation. Adjectives derived from geographical names were also labelled following the scheme for personal proper names. Both place names and the derived adjectives were classified into four categories according to the wider geographical location: place names in Slovenia, in the former Yugoslavia (with the exception of Slovenia), in Europe and the rest.

Even before the advent of corpus linguistic research, which arises from methods that can be applied to larger literary corpora, including corpus stylistics, analyses of geographical names in literary studies took place in two fields: the first field is defined as the geography

of literature, or the study of the spatiality of literary works, which explores space at the level of textuality. Another field is so-called literary geography, which deals with the study of the place-bound nature of writing, publishing and reading and whose results are often presented in literary atlases (Perenič, 2012a, p. 259–260; Gregory et al., 2015, p. 6–8).[3] In recent decades, these two fields have further evolved within DH research, i.e., with distant reading approaches. The potential of entirely new modes and practices for literary scholarship has been suggested, with the aim of complementing existing work with the potential offered by large corpora of literary texts and the development of corpus linguistic and corpus stylistic methods, including the possibilities of data extraction from large machine-readable corpora (Gregory et al., 2015, p. 6–8).

The comparative study of the usage (patterns) and function of place names in literary works that includes both quantitative and qualitative aspects of geographical entities in literary works is called comparative literary onomastics (de Does et al. 2017, pp. 361–362). In the narratological approach of "distant reading", analyses of place names are part of broader research on the relationship between characters, plot, time and space, similar to the analyses within the framework of Text World Theory based on Bakhtin's concept of the chronotope (cf. Šorli and Žejn, 2021, p. 188 and the literature cited there) or the research within the "digital narratology of space" that established the study of space frames based on Lotman's concept of spatial semantics (cf. Viehhauser, 2020, p. 381). Modern quantitative research also includes other aspects in the analysis of space besides the classical narratological categories, such as the analysis of the connections between emotions and space (cf. Grisot and Hermann, 2022). As to the question of literary setting, it follows from the aforementioned types of research that the analysis of geographical entities is only one part of the necessary analysis. At the same time, the data that can be extracted from large corpora allow insights into literature that go beyond the limits of studying a limited corpus of selected "representative" texts.

3 For a survey of relevant research see Hladnik (2012) and, for more recent Slovenian studies, *Prostor slovenske književnosti* (cf. Perenič, 2012b).

2 Statistical analyses

2.1 Literary characters

From the previous analyses it appears that the May68 Corpus is clearly dominated by literary names, PER-LIT (68%), while the other two categories appear in relatively similar proportions: descriptive names, PER-DES (18%), and characters from the non-literary world, PER-REAL (14%). Moreover, a clear preponderance of male characters was found (on average about 80:20 in favour of male characters), and the analysis by the gender of the authors showed that this ratio was somewhat more balanced in female authors, with the exception of real-life characters, where the ratio is independent of gender, most likely due to the real and undisputed presence and roles of men and women in social and cultural history (for more details, see Šorli and Žejn, 2022, p. 193–194).

The annotated May68 Corpus contains texts from three basic literary genres and also enables searching by and comparing among these three basic categories. The following section presents some results of the analysis of the quantitative and proportional distribution of the three types of character names (literary, descriptive, and real names) in drama (see Figure 1), poetry (see Figure 2) and prose (see Figure 3).[4]

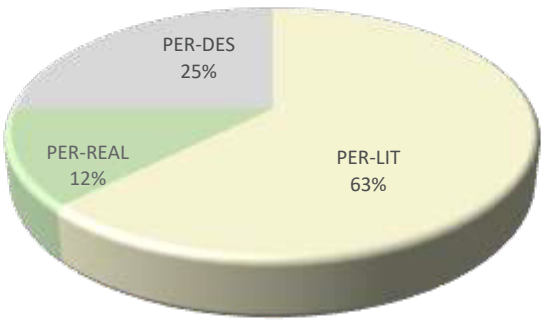

Figure 1: Distribution of character naming types in drama.

4 According to the number of occurrences, poetry accounts for 13.56%, prose for 66.09%, and drama for 18.56%. The remaining 1.77% represent hybrid genres, which are not considered in the analyses both because of their extremely low presence and because of their genre specificity.

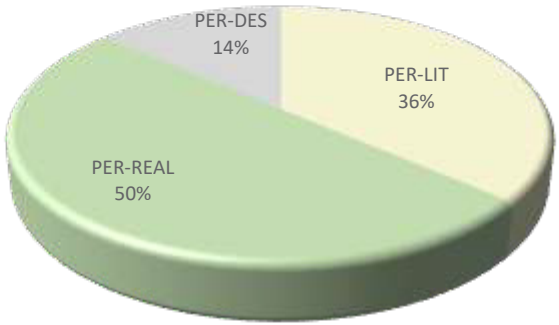

Figure 2: Distribution of character naming types in poetry.

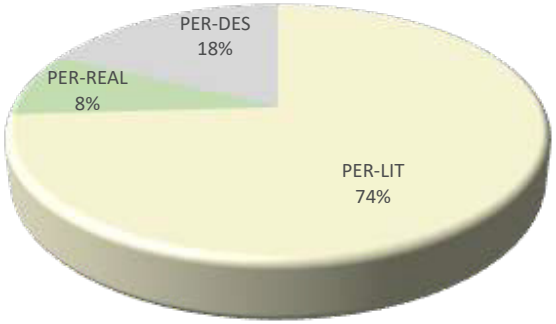

Figure 3: Distribution of character naming types in prose.

A comparison of the three charts shows significant differences in the distribution of the different types of character naming. Literary names (PER-LIT) are more prevalent in drama when compared to descriptive names (PER-DES) and real names (PER-REAL) (63% of all namings), and in prose (nearly 75% of all namings), while in poetry the proportion of literary names comes in third place, at 36%. The results show that in poetry there are fewer direct namings of literary characters and that, in general, these are not prominent. The proportion of descriptive names is largest in dramatic texts (25%), smaller in prose (18%), and even smaller in poetry (14%). It can be concluded that the higher proportion of descriptive names in drama is related to the fact that such texts are primarily intended to be performed on stage over two or three hours, and the descriptive names of characters are used as a means of "economic" characterization. In prose, because of the larger size of the text and the greater likelihood that descriptive nomenclature will take hold

in the text, it is more likely that a particular characteristic of a person, such as a physical trait, an occupation or social status, etc., will serve the function of a proper name.

The relationship between descriptive names in prose and poetry is also characterized by the fact that the first ten descriptive names by frequency (see Table 2) in drama are almost exclusively (with one exception) names that refer to an occupation and/or social status (e.g. chief, principal, mayor); in prose, almost half of the first ten designations indicate a particular physical characteristic of the character (e.g., one-armed, long-haired, old) – such designations are effectively replaced in dramatic texts by descriptions and instructions in the didaskalia or dramatic performance.

Table 2: *The first ten descriptive names in drama and prose by frequency*

DRAMA		**PROSE**	
Lemma	Frequency	Lemma	Frequency
načelnik	38	*senzal*	126
ravnatelj	27	*črni mož*	85
župan	21	*enoroki*	46
gospod šef	19	*inšpektor*	45
novi načelnik	17	*dolgolasec*	42
tovariš župan	16	*Zobčev*[5]	38
taščica	16	*Tomažev*	37
umetnik	14	*stotnik*	37
gospod namestnik	12	*kapitan*	35
bivši načelnik	12	*bela žena*	35
pisar	11	*stari*	31

The list of descriptive names in poetry shows that these descriptive names were annotated in only seven texts, and that a particular type of descriptive name predominates which is not generally a feature of drama or poetry.

In poetry, a large proportion of real-world persons is conspicuous, constituting the majority or even half of the names, while in drama and prose this category of proper names occupies the smallest proportion:

5 *Zobčev* and *Tomažev* are cases where women are named by their (husband's) second name.

12% in drama and only 8% in prose. These results could be an indication of a high degree of referentiality and intertextuality of the poetry in the corpus or of modernist poetry in general.

In the following, we present some analyses of the annotated place names, in accordance with the categories established for manual annotation. The results ensuing from the analysis by these four categories for the entire annotated corpus are shown in Figure 4. This shows that the largest proportion of place names is related to Slovenia (37%), followed by (the rest of) Europe (28%) and countries beyond Europe (25%), with an unexpectedly modest proportion of geographical locations classified in the territory of the former Yugoslavia (only 10%).[6]

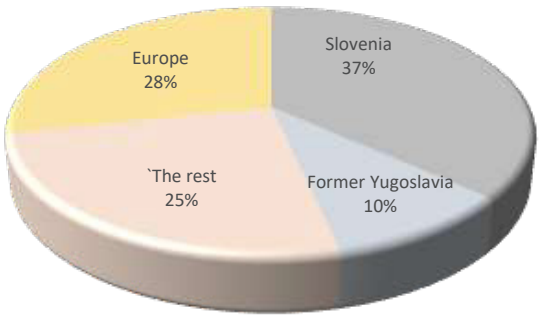

Figure 4: Place names according to the division into four major geographical units.

Similar to character names, we show the results below according to the percentage of geographical locations within each literary genre represented in the corpus.

6 Figure 1 shows data based on the number of occurrences, which show a more relevant picture, since recurring lemmas, in contrast with one or several mentions of a geographical location, are also an indicator of greater importance or presence in the text. The results by the frequency of lemmas show slightly different ratios: locations in Slovenia, Europe and the rest are almost equal each representing a little less than a third (Europe 30%, Slovenia and the rest 29% each); locations in the former Yugoslavia represent 12%. The analysis by the number of occurrences compared to the number of lemmas therefore shows a significantly greater role of geographical locations in Slovenia and the fact that locations outside Europe (the rest) are mostly just brief mentions and not so much actual places of action.

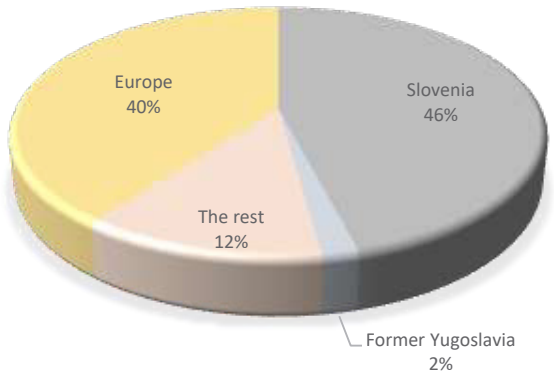

Figure 5: Proportional shares by classification in broader geographical units – Drama.

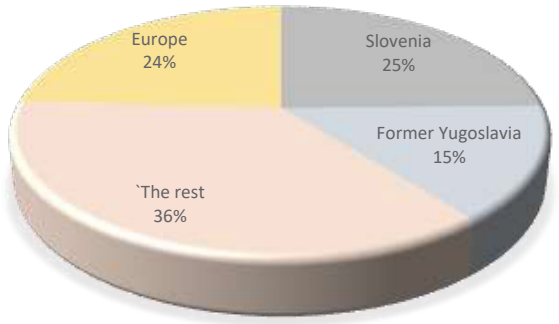

Figure 6: Proportional shares by classification in larger geographical units – Poetry.

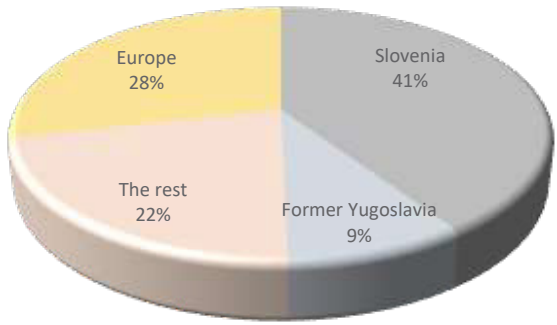

Figure 7: Proportional shares by classification in broader geographical units – Prose.

In the dramatic texts (see Figure 5), a disproportionate share of places in Europe (40%) and Slovenia (46%) is noticeable compared to poetry and prose, as well as a small proportion of places outside Europe (12%) and in the former Yugoslavia (a barely detectable 2%). The proportions within poetry (see Figure 6) are relatively even: places in Slovenia and Europe account for about a quarter, other places for slightly more (36%), and places in Former Yugoslavia for the least (15%). The distribution according to the general geographical classification in the prose (see Figure 7) corresponds most closely to the results for the entire corpus: the largest share is accounted for by geographical places in Slovenia (41%), followed by places in Europe (28%), other places (22%), and only 9% for places in the former Yugoslavia.

Since the number of geographical entities in the corpus is much smaller compared to personal names, the results in this segment are less representative and less likely to be generalizable to modernist literature or literature in general. Nonetheless, they suggest that there are some differences in the selection and listing of geographical locations across genres.

3 Potential benefits of corpus enlargement, additional annotation tasks, and further research

3.1 Conclusions and open issues

The main goal of our annotation task was to provide an adequate representation of a specific set of semantic data, i.e. named entities, and to fully exploit the potential of this type of corpus linguistic data in the context of future literary and linguistic analyses. To this end, we implemented a three-level annotation scheme. The preliminary results and additional analyses presented in this paper provide an argument for annotating the remaining part of the May68 Corpus, possibly with some adjustments to the scheme based on the experience of previous work. Accordingly, in the next phases of annotation we plan to improve the segments that show the lowest degree of consistency and annotator agreement, such as common nouns that serve the referential function of proper nouns and appear to act as a representational

continuum. We have yet to figure out how best to incorporate the various instances of descriptive names (PER-DES) into the annotation scheme, but they are certainly worth considering as a special (sub) category of the NAME group.

Compared to the categories of personal names, significantly fewer dilemmas occurred in the classification and labelling of place names. Individual unresolved cases (e.g., fictitious places, names referring to localities or objects in outer space) were assigned to the category "Other". Due to the high percentage of geographical entities labelled "Other", a further subdivision (in addition to Slovenia, Former Yugoslavia and Europe) of the wider geographical location must be proposed on the basis of the qualitative analysis of the results.

We conclude, on the basis of the high variation in referential expressions, that in potential future projects an additional step should be linking the different names of the same character (so-called "nesting"), the same applies to geographical entities, where several spelling variants occur (e.g., *Švica*, *švajc* = Switzerland).

Some NER projects report on the automatic linking of proper names to entries in Wikipedia (cf. de Does et al. 2017), which assists in the named entity resolution to distinguish between plot-internal and plot-external names. As shown in the introduction, we linked names to web resources only when a (personal) name was not assumed to be part of today's common cultural knowledge. This is another issue that needs to be resolved for future undertakings.

3.2 Future projects

The decision in favour of manual annotation of the May68 Corpus was based on the fact that this is a specialized corpus for which established automatic labelling of named entities would not predictably yield adequate and satisfactory results, as well as on experience from related research on named entities in various national literary corpora (cf. Stanković et al., 2019; Vala et al., 2015; Ketschik, 2020; Papay and Padó, 2020). As Won et al. (2018) have noted, using historical texts as an example, a single automatic tagging tool is not optimal for automatic tagging of place name and instead a clever combination of

multiple approaches is required. The fully annotated corpus will allow empirical testing of the differences between the results of manual and automatic annotation. Despite some adjustments, the three-level scheme for manual annotation is perhaps closest in granularity to the Janes-NER guidelines (CLARIN.SI) (cf. Zupan et al., 2017), whose categories are considered the standard for automatic annotation of Slovenian corpora. The results of this comparison could contribute to the optimization of tools for automatic labelling of named entities for corpora of literary texts.

Last but not least, the results of labelling, especially of character names and geographical entities, are crucial for the construction of an NE database. The database of the May68 Corpus could be the cornerstone for the compilation of a database of proper names in Slovenian literature of different literary genres, directions and periods, and the data on geographical entities could contribute to the research of spatiality of literary works on the level of textuality.

In a second stage of the project that is envisaged, the texts will also be annotated for the use of metaphor, which – financial means permitting – will result in a database of literary metaphor and metonymy. The goal of this additional annotation task will be to optimize the annotation procedure and apply the knowledge acquired about the use of metaphor in modernist literary texts for the purposes of future literary and linguistic analyses.

Acknowledgments

ARRS (Slovenian Research Agency) P6-0024 “Literarnozgodovinske, literarnoteoretične in metodološke raziskave [Research into literary history, literary theory and methodology].”

References

Beck, C. Booth, H., El-Assady, M., & Butt, M. (2020). Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In *The 14th Linguistic Annotation Workshop* (pp. 60–73). Barcelona, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.law-1.6.pdf

de Does, J., Depuydt, K., van Dalen-Oskam, K., & Marx, M. (2017). Namescape: Named Entity Recognition from a Literary Perspective. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 361–370). Ubiquity Press. Retrieved from http://www.jstor.org/stable/j.ctv3t5qjk.37

Eckart de Castilho, R., Mújdricza-Maydt, E., Muhie Yimam, S., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) (pp. 76–84). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclanthology.org/W16-4011.pdf

Gregory, I., Donaldson, C., Murrieta-Flores, P., & Rayson, P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, *9*(1), 1–14. doi:10.3366/ijhac.2015.0135

Grisot, G., Herrmann, B. (2022). Emotions and space: an investigation of "urban" vs. "rural" emotional language in Swiss-German fiction around 1900. *Distant reading closing conference*. Accessed at https://www.distant-reading.net/events/conference-programme/

Hladnik, M. (2012). Prostor v slovenskih literarnovednih študijah: kritične izdaje klasikov. In U. Perenič (Ed.), *Prostor v literaturi in literatura v prostoru = Space in literature and literature in space* (pp. 271–282). Ljubljana: Slavistično društvo Slovenije. Retrieved from http://www.dlib.si/details/URN:NBN:SI:DOC-EFDJCFIF

Juvan, M., Šorli, M., & Žejn, A. (2021). Interpretiranje literature v zmanjšanem merilu: »Oddaljeno branje« korpusa »dolgega leta 1968«. *Jezik in slovstvo*, *66*(4), 55–76.

Juvan, M., Žejn, A., Šorli, M., Mandić, L., Tomažin, A., Jež, A., Balžalorsky Antić, v., & Erjavec, T. (2022). *Corpus of 1968 Slovenian literature Maj68 2.0, ZRC SAZU*, http://hdl.handle.net/11356/1430

Ketschik, N., Blessing, A., Murr, S., Overbeck, M., & Pichler, A. (2020). Interdisziplinäre Annotation von Entitätenreferenzen. Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung. In N. Reiter, A. Pichler & J. Kuhn (Eds.), *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt* (pp. 203–236). Berlin, Boston: De Gruyter. Retrived from https://doi.org/10.1515/9783110693973-010

Pagel, J., Reiter, N., Rösiger, I., & Schulz, S. (2020). Annotation als flexibel einsetzbare Methode. In N. Reiter, A. Pichler & J. Kuhn (Eds.), *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in*

der CRETA-Werkstatt (pp. 125–142). Berlin – Boston: De Gruyter. doi: 10.1515/9783110693973-010

Papay, S., & Padó, S. (2020). RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 835–841). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.104.pdf (1. 12. 2022)

Perenič, U. (2012a). Space in literature and literature in space. In U. Perenič (Ed.), *Space in literature and literature in space* (pp. 265–270). Ljubljana: Slavistično društvo Slovenije. Retrieved from: http://www.dlib.si/details/URN:NBN:SI:DOC-6P13WHOU

Perenič, U. (Ed.) (2012b). *Space in literature and literature in space*. Ljubljana: Slavistično društvo Slovenije. Retrieved from http://www.dlib.si/details/URN:NBN:SI:DOC-6P13WHOU

Stanković, R., Santos, D., Frontini, F., Erjavec, T., & Brando, C. (2019). Named Entity Recognition for Distant Reading in Several Languages. In G. Pálko (Ed.), *DH_Budapest_2019*. Budapest: ELTE. Retrieved from http://elte-dh.hu/dh_budapest_2019-abstract-booklet/

Ševščíková, M., Žabokrtský, Z., & Krůza, O. (2007). Named Entities in Czech: Annotating Data and Developing NE Tagger. In V. Matoušek & P. Mautner (Eds.), *Text, Speech and Dialogue: Proceedings of the 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3–7, 2007.* Berlin – Heidelberg: Springer-Verlag. Retrieved from https://ufal.mff.cuni.cz/~zabokrtsky/publications/papers/tsd07-namedent.pdf

Šorli, M., & Žejn, A. (2022). Annotation of Named Entities in the May68 Corpus: NEs in modernist literary texts. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities 2022* (pp. 187–195) Ljubljana: Institute of Contemporary History. Retrieved from: https://www.sdjt.si/wp/dogodki/konference/jtdh-2022/zbornik/

Vala, H., Jurgens, D., Piper, A., & Ruths, D. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 769–774). Lisbon, Portugal: Association for Computational Linguistics.

Viehhauser, G. (2020). Zur Erkennung von Raum in narrativen Texten: Spatial frames und Raumsemantik als Modelle für eine digitale Narratologie des Raums. In N. Reiter, A. Pichler & J. Kuhn (Eds.), *Reflektierte*

algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt (pp. 373–388). Berlin – Boston: De Gruyter. Retrieved from https://doi.org/10.1515/9783110693973-015

Won, M., Murrieta-Flores, P., & Martins B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities* 5. Retrieved from https://www.frontiersin.org/articles/10.3389/fdigh.2018.00002

Zupan, K., Ljubešić, N., & Erjavec, T. (2017). *Annotation guidelines for Slovenian named entities: Janes-NER*. Technical report, Jožef Stefan Institute, September. Retrieved from https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1.1.pdf

Imenske entitete v modernističnih besedilih: ročno označevanje in analiza korpusa Maj68

V članku najprej predstavimo korpus Maj68, tj. korpus modernističnih literarnih besedil slovenskih avtorjev iz revij *Tribuna* in *Problemi* iz obdobja študentskega gibanja 1968. Korpus je bil avtomatsko oblikoskladenjsko označen, nato je sledila ročna semantična anotacija z namenom naprednejše analize korpusa. Cilj raziskave je bil, da v označeno gradivo zajamemo kompleksnejše semantične pojave in tem prilagodimo označevalni model, ki bi uspešno naslovil dileme označevanja literarnih besedil, in sicer dvoumnost, nejasnost in variantnost. Trinivojska označevalna shema ima tri osnovne kategorije, od katerih se prvi dve delita še nadalje: 1. lastna imena, 2. tuji jeziki in slovenske jezikovne varietete ter 3. bibliografske navedbe. Predstavljene so izbrane vsebinske analize imenskih entitet (imena likov in geografska imena) glede na tri temeljne literarne zvrsti. Rezultati analiz pokažejo določene razlike med zvrstmi, ki jih je mogoče interpretativno postaviti v širši literarni kontekst. V sklepih razmišljamo o možnostih izboljšave sheme, njene dodatne nadgradnje ter o potencialni nadgradnji rezultatov.

Ključne besede: modernizem, imenske entitete, korpusna stilistika, slovenska literatura, *Tribuna*, *Problemi*, 1968