# Annotation of Named Entities in the May68 Corpus: NEs in modernist literary texts

**Mojca Šorli,[*] Andrejka Žejn[†]**

[*] ZRC SAZU, Institute of Slovenian Literature and Literary Studies
Novi trg 2, SI-1000 Ljubljana
mojca.sorli@zrc-sazu.si

[†] ZRC SAZU, Institute of Slovenian Literature and Literary Studies
Novi trg 2, SI-1000 Ljubljana
andrejka.zejn@zrc-sazu.si

**Abstract**

In this paper we present the process of manual semantic annotation of a corpus of modernist literary texts. An extended set of annotations is proposed with respect to the established NER-systems and practices of related projects, i.e. several categories of proper names, foreign language elements and bibliographic citations. We focus on the annotation challenges concerning the names of literary characters seen in transition from common nouns to proper names, as well as giving examples of the results of preliminary analyses of the corpus.

## 1. Introduction

The starting point of the digital humanist literary project presented here is a corpus of literary texts that was created according to special criteria defined for the purposes of this research. In view of the significance for DH 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 text volumes require automation of corpus processing for quantitative analysis, involving different levels of (linguistic) annotation in the first phase, and allowing 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 Named Entities are distinguished based on semantic criteria, so we consider this an example of semantic annotation.

Linguistically annotated corpora have long been a standard tool for linguistic research. Named Entity Recognition (hereafter NER) and analysis has also long been relevant in the social sciences and sociology (Ketschik, 2020), from where the method, like several others, has been transferred via linguistics to literary studies, where named entities are most closely associated with literary character research. A more comprehensive picture of the way characters are named in literature, beyond the automatic recognition of Named Entities (hereafter NEs), can be obtained by manually annotating these entities in literary texts, by analyzing the annotation process, and finally by analyzing the data obtained from the annotated corpus itself.

## 2. The Goal of the paper

In this paper we report on an attempt to identify and annotate three groups of NEs in the "Corpus of 1968 Slovenian literature Maj68 2.0" (short name May68 Corpus) – corpus of Slovenian modernist literary texts from the late 1960s to the early 1970s,[1] discussing these groups from the point of view of three different sources of representation problems that are independent but interrelated: ambiguity, variation, uncertainty. 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 entire annotation system. We list a number of relevant annotated categories, their specific character, and representational problems associated with them. Our choices are discussed when any of the first three listed sources of representation problems apply.

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 is a model of extended annotation of NEs in modernist periodicals that can be applied in certain segments to other corpora of literary texts. We focus both on the identified inaccuracies and on the benefits of manual annotation of selected groups of NEs in our specialized corpus of literary texts. In the concluding part, we present the preliminary results of an analysis performed on the annotated corpus.

Following the automatic preprocessing (i.e., POS tagging and lemmatization) of the May68 Corpus, further manual annotation was performed to capture more complex linguistic (semantic) phenomena and to provide a more sophisticated annotation model for proper names given the recurring representational problems: 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. Here we will focus on some open challenges in the annotation of NEs, in particular problems related to the functional aspects of the annotated elements. We discuss the practical treatment of proper names for the purposes of corpus linguistic and stylistic research, in the hope of

[1] http://hdl.handle.net/11356/1491

improving the reliability of research results and also of NLP models.

## 3. Automated and manual annotation of corpora

In the context of language technologies, universal concepts and tools for automatic corpus annotation have been developed to some extent, especially for individual language groups, while language-specific concepts and tools are also needed. Established levels of automatic tagging for Slovenian, initially based on lexicographic and linguistic projects, include tokenization and related segmentation into sentences, normalization, morphosyntactic tagging, lemmatization, and syntactic parsing (Erjavec et al., 2015). NEs pose a challenge for automatic extraction of information due to their semantic an functional complexities. For Slovenian, the main tool used is StanfordNER, which assigns lexical units to predefined categories (Ljubešić et al., 2012): personal names, geographical names and common proper nouns. The state-of-the-art of the existing NER tools for Slovenian has not been the focus of this research, but a preliminary review of the tools, as well as of the function of NEs in the texts, has shown their limited applicability to a specialized literary corpus that we set out to investigate.

### 3.1. NER-systems for corpora of literary texts

For literary texts, narratology in particular has developed various typologies of protagonists, heroes, or major and minor characters in texts, ways of characterizing them, and strategies for recognizing them. Since the advent of digital tools researchers have had to find a way to translate the definitions formed by literary scholars into computer-readable data (Krautter et al., 2018).

While there are no specific NER-systems for annotating literary texts, even though literary texts have a high variation of NEs compared to normal non-fiction texts (Stanković et al., 2019), "universal" systems are often used. However, automatic annotation tends to overlook certain segments of NEs in literary texts (Vala et al., 2015). Attempts are made to overcome these limitations by additional automatic tagging, or to expand the set of annotated entities by manual tagging, often of referential expressions, i.e., linguistic expressions that refer to a specific entity in the text world, where the entities and their references must be interconnected (entity grounding). References and connections themselves can only be inferred from the knowledge of the context (Ketschik, 2020; Papay and Padó, 2020), so in the early stages of research, manual annotation of the corpus is usually required to improve the automatic process.

### 3.2. Background and related work

Compiling lists of NEs, especially for categories of proper names, represents only the basis for the identification of character names and is as yet insufficient for relevant literary analyses, so these lists must be dealt with by multidimensional approaches that shed additional light on proper names in light of the special features of literary text. Empirical analyses of protagonists in the literature can, at the most basic level, for example, study the characteristics of names, their typicality, archaic character, or "unusualness" for a particular society (cf. Calvo Tello, 2021), compare usage and functions of proper names, exploring to what extent they are genre-related (e.g. children's literature, cf. van Dalen-Oskam, 2022).

Empirical analysis of the ratio between female and male characters in a corpus of English literature up to the mid-20th century (cf. Nagaraj and Kejriwal, 2022), for example, showed the quantitative predominance of male characters over female characters. More complex research also deals with characterization analysis, identifying relationships between main and secondary characters, examining the relationship between active and passive character presence, and distinguishing between "actively present" characters and characters from other fictional worlds (Krautter et al., 2018; Brooke et al., 2016; Ketschik, 2020). One of the more established approaches is the application of social network analysis, a method from empirical sociology that builds on the relationship between NEs. The analysis of social networks in the literature (cf. de Does, 2017) is closely related to quantitative approaches to the study of direct and reported speech or narrator speech and character speech in storytelling and drama, where NEs are an essential component of a broader context (cf. Burrows, 2004; Moretti, 2011; Elson et al., 2010; Papay and Padó, 2020). Digitally supported analysis of the broader picture of characters also draws on concepts derived from Bakhtin's concept of chronotope, such as The Text World Theory – a cognitive-linguistic concept of a unity of characters, time and space, or the concept of situation (Krautter et al., 2018; Mikhalkova et al., 2019).

## 4. Model annotation schemes

In designing the model for manual annotation of the May68 Corpus, we relied on familiarity with the texts contained in the corpus and on several other well-known models of manual annotation for similar projects, three of which are presented below.

### 4.1. COST Action ("Distant reading" project)

Distant Reading project for the annotation of the multilingual ELTeC corpus (https://www.distant-reading.net/eltec/)[2] based on European novel provides the following distinct categories: "demonyms (DEMO), professions and titles (ROLE), works of art (WORK) person names (PERS), places (LOC), events (EVENT), organizations (ORG)" (for a brief description of the categories cf. Frontini, 2020). The selection of these categories was partly motivated by the existing possibilities of automated NER, which brings with it certain limitations (Stanković et al., 2019). The project also points out the importance of "cultural references, role models and cosmopolitanism", and these can only be answered "if references to works of art, authors, folklore and periodical publications are detected", which is why in our corpus of

[2] The Distant Reading for European Literary History (COST Action CA16204) started in 2017 with the goal of using computational methods of analysis for large collections of literary texts. It is based on the compilation and analysis of a multilingual open source collection, named European Literary Text Collection (ELTeC).

modernist literary texts we introduced a BIBLIO group to incorporate references to authors, but covered other listed types of references with the "other" group (NAME / XXX). In May68 Corpus, however, we focus for now on proper names.

### 4.2. CLARIN.SI

The annotation scheme adopted largely follows the guidelines provided for Slovenian in the past (e.g. Štajner et al., 2013), perhaps closest in its granularity to the Janes-NER guidelines (CLARIN.SI) as described by Zupan et al. (2017), except for the derived adjectives (DERIV-PER) type, which is given here an independent status unlike in May68 Corpus, where this is subsumed under the PER-LIT and PER-REAL subtypes.[3]

In addition, we decided in the case of May68 Corpus to conceptualize combinations of nouns denoting professions, functions or titles, and personal names as units, therefore labelling the entire strings as literary personal name (PER-LIT) or real personal name (PER-REAL).

### 4.3. Annotation schemes for Czech language

Annotation of NEs in Czech corpora is implemented according to more complex models as described in Sevščíková et al. (2007). Our three-level NE taxonomy is, nonetheless, somewhat less fine-grained. Furthermore, unlike the Czech model, ours does not include numbers, such as in addresses, zip codes, or phone numbers, specific number usages and quantitative expressions – entities typically included in NER.

## 5. May68 Corpus of Slovenian modernist literary texts – corpus description

The Maj68 Corpus is a result of a project on the literature of the avant-garde and modernism in the period of the worldwide student movement, whose activities are also reflected in the transformation of literature. 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 Maj68 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 Maj68 Corpus 2.0 version, which has been further edited and corrected (metadata), contains 647 additional texts from *Tribuna* and *Problemi*.

The compilation of the corpus began with an extensive bibliographic inventory of texts in selected publications that have been digitized and are publicly available on dLib. On the basis of these lists, the original texts of Slovenian authors were converted from .pdf format to .docx format and, in a second phase, linked to metadata in Excel spreadsheets. Finally, the corpus was automatically tagged (see Juvan et. al 2021 for more details on the procedure).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 consist only of visual elements, i.e. they do not contain standard text.

Automatic linguistic annotation includes lemmas, morpho-syntactic descriptions from MULTEXT-East, and morphological features and syntactic annotations from Universal Dependencies. As shown here, manually tagged NEs for persons, geographical locations, organizations, and various names, (foreign) linguistic variations and registers, and cited authors (sources) are additionally marked.

The following sections and subsections introduce the types and categories of NEs, including the dilemmas encountered in the process of annotation and the practical reasons for annotation. From here on, and with a somewhat narrower notion of NER, we speak of categories of "proper names (personal and place names)" rather than "named entities" for the purposes of this paper.

### 5.1. Annotation procedure and categories

The annotation was implemented using the WebAnno tool (Eckart de Castilho et al., 2016). To simplify the technical aspect, the whole corpus was divided into 1529 sections of five sentences each, on average 380 chunks per section. WebAnno allows annotation of one sentence at a time, which was a disadvantage for longer instances of text marked by the use of foreign language(s). Each annotation round was curated by two curators.[4] 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.

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 model 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.

The NAME group includes the following types and subtypes:

- Person (PER), including the person-derived adjective, is subdivided into fictional literary characters (PER-LIT), characters referring to real, i.e., 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) is divided into locations in Slovenia (GEO-SI), in former Yugoslavia (GEO-YU), in Europe (GEO-EU), and in others (GEO-ZZ).
- Organizations and institutions (ORG).
- Miscellaneous (XXX).

A group labelled FOREIGN is used to annotate the foreign language: Serbo-Croatian (SBH), English (EN),

[3] Overall and in the same fashion, in May68 Corpus we also favour larger lexical units.

[4] The texts were annotated by A. Jarc, L. Mandić, and K. Žvanut in accordance with the annotation scheme designed by the authors of this paper, who also curated all of the annotations.

French (FR), Italian (IT), Latin (LA), and German (GE), or register variation (DIALECT, INFORMAL, SLANG) in the corpus.

Once the annotation process was completed, the labels in WebAnno were converted to TEI encoding.[5] Following the conversion thus all proper names (personal names, place names, names of organizations, and real names) are 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). Units of text with foreign languages and non-standard Slovenian were labelled as <foreign/> and corresponding attributes according to TEI coding.

### 5.1.1. Person

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. The most important subdivision of the type (within the NAME group) is that between 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. Subcategories include names of people and pets, nicknames, pseudonyms, members of national and social groups.

| Group | Type | Subtype | Description |
|---|---|---|---|
| NAME | PERSON (PER) | PER-REAL | Real: Characters referring to real, i.e. existing and historical or mythological persons or beings (web sources, Wikipedia, etc.), e.g. *Greta Garbo*. |
| | | PER-LIT | Literary: Fictional literary characters, e.g. *Ančika*, *Zobec*. |
| | | 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* |
| | | GEO-YU | Former Yugoslavia (except for Slovenia), e.g. *Zagreb* |
| | | GEO-EU | Europe, e.g. *Frankfurt* |
| | | GEO-ZZ | Other, e.g. *Peking* |
| | ORG | – | Names of organizations, institutions (*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*. |
| FOREIGN | HBS | – | Serbo-Croatian |
| | EN | – | English |
| | DE | – | German |
| | FR | – | French |
| | IT | – | Italian |
| | LA | – | Latin |
| | XX | – | Other |
| | DIALECT | – | Dialect |
| | VERNACULAR | – | Vernacular |
| | SLANG | – | Slang |
| BIBLIO | – | – | Quoted authors (Sources) |

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

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, e.g., a website, for a group of people appearing in the literary text. The assignment of a URL depends on context or extra-linguistic knowledge; if a person can be assumed to be part of common (cultural) knowledge (Descartes, Nietzsche), we do not 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), regional identity (Kranjci, Štajerci; Novakovi), but also smaller social groups defined on the basis of occupational or other criteria.

Of the categories introduced specifically for the purposes of the May68 Corpus, NAME / PER-DES proved, as expected, to be the most challenging subcategory (see 6.1.1.).

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 proper names in plays depends on the status and/or function of the proper name. Names of individual characters that merely announce an individual character's speech, 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 ogleduje v zrcalu* / Pandolfo looks at himself in the mirror). Below is an example of a dialogue showing the distinction between the two and a third subtype (the names in bold are labelled as PER-LIT, PER-DES and PER-REAL respectively):

[5] The annotation task was carried out in collaboration with T. Erjavec (technical aspects and data conversion).

BARRÈRE: *Potemtakem moramo danes z njim obračunati.* (***Tallien*** *odide*)
(Davidu): *Si pripravljen s* ***Krepostnim*** *umreti*?
DAVID: *V smrt?*
BARRÈRE: *Se nisi maločas naglas pridušil?*
DAVID: *Čudovit črtež sem zamislil. Kako dviga* ***Sokrates*** *čašo strupa k ustom. Naš dobri prijatelj je tako presunljivo govoril.*

Adjectives derived from personal proper nouns are annotated as the corresponding proper nouns. Their derived character is revealed by morpho-syntactic tagging.

#### 5.1.2. Geographical location (GEO)

Place names are labelled as NAME and the following closed-set subtypes: SI, YU, EU, ZZ, depending on whether the location is in Slovenia, in the former Yugoslav republics, in the rest of Europe, or outside all of these areas.

As with personal names, a distinction is made between real and fictitious geographical names (*Indija* vs. *Eldorado*). Commentators decide whether a place is real or fictitious (such as street names in a fictitious city) based on context and common knowledge. Places typically include continents, countries, regions, cities, towns, and natural geographical objects, as well as streets, squares, and neighbourhoods, and functional infrastructure such as churches, airports, and local cultural and natural sites. Place names used metaphorically, e.g. *Eden*, are categorized as "other" and assigned the label NAME / GEO-ZZ – the same label is used for place names outside the European territory. At this stage, we have not paid special attention to the treatment of proper names (personification) used metaphorically, such as

*Jadra so pogorela,* ***Delfi*** *molčijo …* [The sails have burnt down, and *Delfi* stays silent …]

This type of analysis is planned for the later stages of annotation (which will include the annotation of metaphors).

Adjectives derived from place names, e.g. African, European, were included in the annotation by analogy with geographical names and divided into the same subtypes (SLO, YU, EU, ZZ).

#### 5.1.3. Organizations and common proper nouns

As with geographical names, there are no subgroups for the two groups of so-called common proper names and names for organizations. Capitalization is an obvious but not a necessary condition for this classification. Thus, no distinction is made here between real and fictitious; what matters is that the name be recognized as "common proper" in the literary context of the text.

Organizations and institutions subsume names of museums and other cultural institutions, as well as political and civic organizations. Organizations are labelled as ORG and usually include businesses, institutes, media, cultural, and educational institutions. However, we have treated restaurants, music groups, and other "entertainment" establishments as "miscellaneous" rather than organizations.

Miscellaneous is a category reserved mainly for common proper nouns, as explained above, such as titles of books and other works of art, artefacts, films, documents, brand names, commercial products, events, including place names, such as mythological places, place names used metaphorically, etc. These NEs are labelled as XXX.

For many common nouns, one can observe a transition to the category of proper names, which seems to exist as a continuum. For example, the word *krčma* (Eng. inn, pub) assumes the function of a proper noun referring exclusively to a particular unit/object, in this case "inn". The word is therefore referred to as NAME / XXX.

#### 5.1.4. BIBLIO

BIBLIO is typically used for literary works cited or mentioned in the literary texts. It contains text passages that refer to literary works or other bibliographic units, and is annotated for authors, not titles or citations, e.g.

*The patamus can never reach The mango on the mango tree* ***(T. S. Eliot****: The Hippopotamus)*

#### 5.1.5. Language and register

In the case of language and register variation, we use the FOREIGN group that subsumes (foreign) language and register variation (see Table 1). This group is not directly relevant to this paper.

## 6. Dilemmas of annotation in the framework of representational problems

A number of dilemmas are discussed here in terms of the three categories – ambiguity, variation, and uncertainty – as detailed, for example, in Beck et al. (2020), who outline the main representational problems in linguistic annotation (we disregard the two additional categories addressed in the model: error and bias).

The interpretation of the listed categories is tailored to the nature of our data, and the problems are assigned to the listed categories accordingly. The annotation process is consistently guided by the identified function of the annotated elements. The three dilemmas are described below.

### 6.1. Ambiguity

In principle, ambiguity occurs whenever a unit admits several interpretations. Ambiguities between form and meaning occur in natural language at the phonological, morpho-syntactic, lexical, or pragmatic levels and are a major source of representational problems (Beck et al., 2020).

#### 6.1.1. Transition from personal proper names to "common proper nouns"

The most striking example of ambiguity is the transition from common nouns to those that function as personal names. This is a pervasive and rather complex representational problem. The dilemma concerns the category NAME / PER-DES, i.e., descriptive names of literary characters, especially in relation to the category NAME / PER-LIT, which refers to standard proper names that are recognizable as such because of their form and conventional properties (e.g., capitalization). This group includes examples where common nouns optionally combine with proper names to refer to individual characters like "inšpektor (Kos)" [inspector (Kos)], or "veteran" [the veteran], including capitalized adjectival derivatives, such as "Brezposelni" [The jobless one], functioning as personal names, etc.

However, capitalization is not a necessary condition for the NAME / PER-DES designation, especially in a corpus of modernist texts that frequently employ modernist and/or

idiosyncratic conventions, with orthographic rules applied to proper names or descriptive linguistic units that typically eschew capitalization (e.g.,"fant" [the boy], "starka" [the old woman]). A key feature of proper names, as it turns out, is "descriptive continuity," which shows that there is no clear boundary between what can be considered a standard proper name (which is traditionally subsumed under onomastics) and what can be understood as an instance of a text that performs the function of a proper name, but does not, strictly speaking, qualify as such.

The assignment of a noun to NAME / PER-DES is decided primarily on the basis of context. Often, a lexical unit (word or phrase) is used to describe a particular property of the character to which the proper noun initially refers, and which is then gradually but clearly transformed into a (descriptive) unit that functions as a proper name (whether capitalized or not), such as "Rdečelasi" [The red-haired one]. The descriptive name is used only when the transition is complete, which must be evident from the broader context. The quantitative criterion (in longer texts) is a minimum of three occurrences of the same designation, such as below:

> Videl je same znane obraze — *inšpektorja Kosa, vratarja Žorža, kurirja Enorokega, Žana*, nekoliko v ozadju pa je stal bledi ***Novinec*** [the (pale) new guy], …

Other examples include *dolgolasec* [the long-haired guy], *mladenič* [young man], *mojster* [the master], *debelušček* [the fatty] and typically correspond to phrases introduced with a definite article in English. In principle, PER-DES is not limited to a maximum number of components, but the likelihood that a lengthy description, such as Zagledal je na tleh sedečega *fanta upadlih lic in kuštravih las* [He saw *a boy with skinny cheeks and messy hair* sitting on the floor], should appear three times at least in the text(s) is minimal. Even if descriptive units tend to recur they normally vary in at least one of their elements.

Capitalization itself does not preclude a lexical unit from being labelled PER-DES, as with *Mož brez imena* [the Nameless Man].

Appellatives, nicknames, and pseudonyms are labelled as ordinary personal proper names (NAME / PER-LIT), except for those expressing description, such as *Dolgi Džon* [John the Longish].

### 6.1.2. Nesting

Another example of ambiguity concerns nesting, which often creates additional annotation problems. Instead of a potential two- (or three-level) nesting model, a single-level nesting is used throughout, taking as the basic annotated unit the largest possible lexical unit, typically a geographical name or the name of an organization composed of one or more proper names: in the case of *Državna založba Slovenije* [National Publishing House of Slovenia], the entire unit is labelled as an organization (ORG) and the proper name *Slovenije* is not nested and labelled on its own as a place name (*Slovenija*); the same goes for for *Društvo novinarjev Slovenije* [Journalists' Association of Slovenia], *Prešernova družba* [Prešeren's Society Publishing], *Direkcija za prehrano Beograd* [Belgrade Food Agency], or, e.g. *Fani* is NOT nested in *gospodična Fani*, but treated as a single-level personal proper name. A general dilemma often arises here as to whether the term should be referred to as a proper name or as a common noun.

## 6.2. Variation

In variation, the same content or value is expressed by multiple, interchangeable variants (Lüdeling, 2017). Variation can be due to extra-linguistic factors, such as the time period, genre, author/speaker of the text, or linguistic conventions.

Like ambiguity, variation is an inherent part of natural language and thus of corpus data. Indirectly related to variation is the case of ambiguity described above in 7.1.1. The descriptive name is not necessarily used exclusively for one and the same literary character; on the contrary, it usually alternates with the character's actual proper name.

Alternation in the mention of literary characters is very common; in fact, it is the rule. Some personal proper names (including their descriptive variants) occur as variants preceded by an attributive noun (always the same), usually referring to their professional or social status (e.g., Inspector Kos). When this type of designation is used consistently, we refer to the entire lexical unit as NAME / PER-LIT, but when the attributive noun (Inspector) becomes an independent descriptive variant, we refer to it as NAME / PER-DES.

Descriptive terms NAME / PER-DES may consist of one or more words, they may be a combination of "object nouns" and standard proper names (*inšpektor Kos*) or of two or more "common nouns" (*kurir Enoroki*), regardless of their capitalization, as long as they function as personal proper names when referring to or naming characters. The same character may be referred to by three, four, or more variants. In our case: inspector Kos, inspector or Kos.

Also treated as single variants are lexical units denoting proper names whose capitalization varies, e.g., *Ministrstvo za kulturo Republike Slovenije* vs. *ministrstvo za kulturo* (Ministry of Culture) and *Zveza borcev* vs. *zveza borcev* (Association of Freedom Fighters).

We are aware that when variants are expressed as a single interpretation, the property of variation as a whole is lost. However, a semantic annotation based on the function of linguistic elements is less prone to structural diversity than, for example, spelling variations in historical texts that reflect, for example, dialectal and/or temporal differences (cf. Beck et al., 2020), which is why, apart from our own specific research goals, we did not choose to preserve (proper name) variations.

## 6.3. Uncertainty

Uncertainty arises whenever there are multiple possible interpretations of data, but the relevant or reliable knowledge to make an informed decision about interpretation is not available (see Bonnea et al., 2014, in Beck et al. 2020). Most examples involve the inability to distinguish between the subtypes PER-REAL and PER-LIT in texts that do not provide sufficient clues to the "origin" of the character, although this seems to be rather rare.

In such cases, manual annotation provides the opportunity for discussion and collective decision, which we see as an advantage, since cases where the uncertainty (or ambiguity) cannot be resolved are reduced to the absolute minimum, for example:

> ***Maruška*** – [PER-REAL, author's wife] *peče domači kruh …*
> ***Milenko, Andraž, Marko, David*** – [PER-REAL, members of OHO Slovenian art group: Milenko Matanović, Andraž

Šalamun, Marko Pogačnik, David Nez – established on the basis of extra-textual knowledge].

## 7. Preliminary results

Apart from the problems encountered in the annotation itself, the preliminary research results of the annotated corpus can also contribute to the study of characters in a selected corpus of literary texts. Based on the query and the results in NoSketchEngine, Figure 1 shows the quantitative relationship between three subtypes of the type PERSON (literary names, descriptive names, and names of characters from the non-literary world). It can be seen that the majority are literary names (PER-LIT, 68 per cent) whose predominance was to be expected - followed quantitatively by descriptive names (PER-DES, 18 per cent, and then by names of characters from the non-literary world (PER-REAL, 14 per cent).

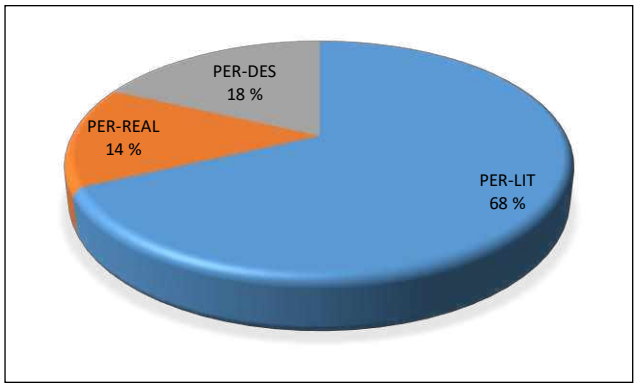


Figure 1: The ratio between the subtypes literary, descriptive and real of the PERSON type.

### 7.1. Categories of descriptive names and real names

Using the lists of the three types of personal names, we can create an approximate typology of character names according to the given typologies and evaluate the consistency of labelling. Because of their special characteristics, we limit ourselves to the subtypes descriptive and real, leaving aside the subtype literary, which includes mostly "ordinary" personal names.

Descriptive names are most often occupational (e.g., chief, inspector, captain, mayor; foreman, waitress, secretary, lab assistant); second are names expressing physical characteristics (e.g., one-armed, long-haired, "the one with the moustache" the handicapped), followed by names describing character (e.g., bully, beast, monster), beast, bloodthirsty),family relations (e.g., aunt, uncle, godmother), generational affiliation (e.g., old man, young man), while longer descriptive lexical strings are rarer (man with no name, brother in Christ, the long-haired one). Among the names for women, forms that formally express possession but function as gendered common proper names are frequent in Slovenian (e.g. Tomaž's (one), the manager's wife). This is statistically almost as significant as feminine names for occupations.

As can be seen from the annotated corpus, we identify five subcategories and include them in the subtype for real persons: 1. Real persons from social (Brutus, Lenin, Kidrič) and cultural history (Prešeren, Heidegger, Descartes, Shakespeare, Mozart); 2. Mythological figures (Cain, Poseidon, Ishtar); 3. Characters from other works of Slovenian and world literature (Pegam, Lambergar, Servant Jernej, Charlie Brown, Odysseus, Pinocchio); the last two groups are represented, on the one hand, by characters from the contemporary world of the authors, such as real-life celebrities (Tomaž Terček, Andraž Šalamun, Milenko Matanovič, Brigitte Bardot, Gérard Philipe, Giorgio Albertazzi, Sylvie Vartan) and, on the other hand, by characters from the authors' immediate (family) environment (Ana, Maruška).

The results show the least consistency for the descriptive name subtype with the lowest degree of intersubjectivity, especially with respect to the relationship between the transition from common noun to proper name and the aptronyms or nominative determinism, which Barthes considers a kind of "economic" characterization (Lahn and Maister, 2016). The relatively high presence of this subtype suggests a modernist blurring of the boundary between fiction and reality, which is reinforced by postmodernism.

### 7.2. Relationship between male and female characters

The second graph (cf. Figure 2) shows the quantitative ratio between male and female characters as they occur in the May68 Corpus (based on the number of tokens).

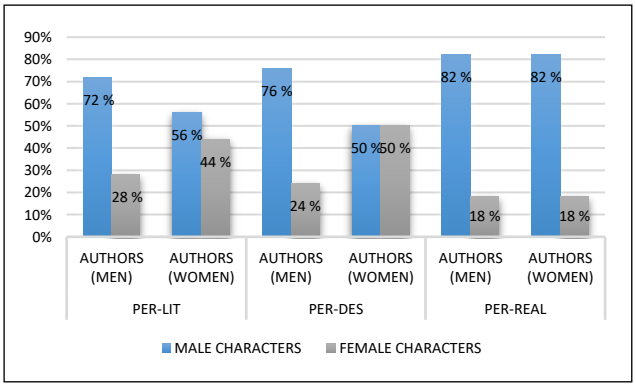


Figure 2: The quantitative relationship between male and female characters in the May68 Corpus.

The results confirm findings from other research (cf. Nagaraj and Kejriwal, 2022) that the proportion of male characters is significantly higher than that of women.

We supplement this account by comparing male and female characters by author gender, which gives a very disproportionate picture: Metadata analysis has shown the predominance of male authorship in the corpus (81 per cent) - only 7 per cent of authors are women, and there are no data for the remaining 12 per cent (Juvan, et al., 2021).

If we start from the gender of the authors when analyzing the occurrence of male and female characters, we find (see Figure 3) that in the works by men, male characters outnumber female characters by 44 per cent in the subcategory literary names, while this difference is much smaller in the works by women (12 per cent). In the category descriptive names, this ratio is difficult to assess due to the low occurrence among women authors, but a large difference between female and male characters in men authors goes in favour of the latter.

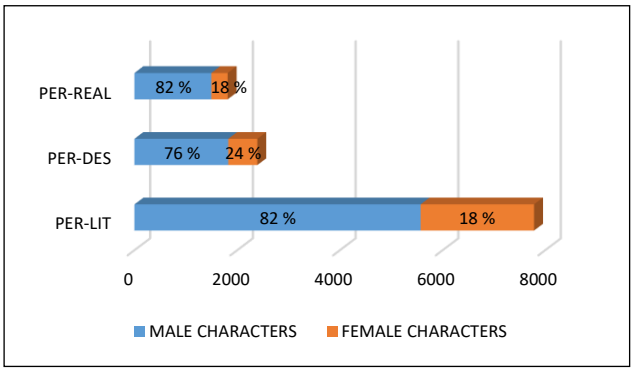


Figure 3: Male and female characters according to the gender of authors.

In the subcategory real, there is no significant difference in terms of author gender, which is probably due to the actual and undisputed presence of men and women in social and cultural history.

## 8. Conclusions and open challenges

The main goal of our annotation task was to provide an adequate representation of a specific set of semantic data (=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 process. We conclude on the basis of high variation in referential expressions that in potential future projects an additional step should be linking the different names of the same character.

In the present work, we sought to identify and interpret different types of representational problems based on the model proposed by Beck et al. (2020) in order to improve our understanding of the linguistic and extra-linguistic properties of the texts in a (literary) corpus. It is hoped that this will lead to a more nuanced understanding of the challenges of NER, and that this in turn may inform future resources in ways that are more appropriate to the data they represent.

In the next phases of annotation, we plan to improve the segments that have the lowest level of consistency and agreement among annotators, such as common nouns that perform the referential function of proper names, seemingly operating as a representational continuum.

We have yet to work out the best approach to fully incorporate the various instances of PER-DES in the annotation scheme, but these are certainly worth considering as a special (sub)category of the NAME group.

## 9. Acknowledgements

ARRS (Slovenian Research Agency) J6-9384 “Maj 68 v literaturi in teoriji (May '68 in Literature and Theory)”

## 10. References

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In: The 14th Linguistic *Annotation Workshop*, pages 60–73, Barcelona, Spain, December 12, 2020.

Julian Brooke, Timothy Baldwin, and Adam Hammond. 2016. Bootstrapped Text-level Named Entity Recognition for Literature. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 344–350, Berlin, Germany, August 7–12.

John Burrows. 2004. Textual analysis. In: S. Schreibman, Ray Siemens, and John Unsworth, eds., *A Companion to Digital Humanities*. Blackwell, Oxford.

José Calvo Tello. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press, Bielefeld.

Karine van Dalen-Oskam. 2022. *Distant Dreaming About European Literary History*. Evening keynote at the Distant Reading Closing Conference. https://www.distant-reading.net/events/conference-programme/

Jesse de Does, Katrien Depuydt, Karina van Dalen-Oskam, and Maarten Marx. 2017. Namescape: Named Entity Recognition from a Literary Perspective. In: J. Odijk, and A. van Hessen, eds., *CLARIN in the Low Countries*, pages 361–70. Ubiquity Press. https://www.ubiquitypress.com/site/chapters/10.5334/bbi.30/download/1046/.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 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)*, pages 76–84. Osaka, Japan. The COLING 2016 Organizing Committee.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting Social Networks from Literary Fiction. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Tomaž Erjavec, Peter Holozan, and Nikola Ljubešić. 2015. Jezikovne tehnologije in zapis korpusa. In: V. Gorjanc, P. Gantar, I. Kosem and S. Krek, eds., *Slovar sodobne slovenščine: problemi in rešitve*, pages 262–76. Znanstvena založba Filozofske fakultete, Ljubljana.

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named Entity Recognition for Distant Reading in ELTeC. In: *CLARIN Annual Conference 2020, Oct 2020*, str. 37–41, Virtual Event, France.

Marko Juvan, Andrejka Žejn, Mojca Šorli, Lucija Mandić, Andrej Tomažin, Andraž Jež, Varja Balžalorsky Antić, and Tomaž Erjavec. 2022. *Corpus of 1968 Slovenian literature Maj68 2.0*, ZRC SAZU. http://hdl.handle.net/11356/1430

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

Nora Ketschik, André Blessing, Sandra Murr, Maximilian Overbeck, and Axel Pichler. 2020. Interdisziplinäre Annotation von Entitätenreferenzen. Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung. In: N. Reiter, A. Pichler, and J. Kuhn, eds., *Reflektierte Algorithmische Textanalyse.*

*Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*, pages 203–36, Berlin.

Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand. 2018. In: T. Weitin, ed., *Eponymous Heroes and Protagonists – Character Classification in German-Language Dramas*. LitLab. Pamphlet # 7.

Silke Lahn, and Jan Christoph Meister. 2016. *Einführung in die Erzähltextanalyse*. Stuttgart, Metzler.

Nikola Ljubešić, Marija Stupar, and Tereza Jurič. 2012. Building Named Entity Recognition Models For Croatian And Slovene. In: T. Erjavec, and J. Žganec Gros, eds., *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012, Ljubljana, Slovenia: proceedings of the 15th International Multiconference Information Society - IS 2012, volume C*, pages 129–34. Ljubljana, Institut Jožef Stefan.

Anke Lüdeling. 2017. Variationistische Korpusstudien. In: M. Konopka, and A. Wöllstein, eds., *Grammatische Variation. Empirische Zugänge und theoretische Modellierung. IDS Jahrbuch 2016*, pages 129– 144. de Gruyter, Berlin.

Elena V. Mikhalkova, Timofei Protasov, Anastasiia Drozdova, Anastasiia Bashmakova, and Polina Gavin. 2019. Towards annotation of text worlds in a literary work. In: *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, pages 101–10. Issue 18, Supplementary Volume 18.

Franco Moretti. 2011. Network Theory, Plot Analysis. *New Left Review*, 68:80–102.

Akarsh Nagaraj, and Mayank Kejriwal. 2022. Robust Quantification of Gender Disparity in Pre-Modern English Literature using Natural Language Processing. arXiv:2204.05872v1 [cs.CY] 12 Apr 2022.

Sean Papay, and Sebastian Padó. 2020. RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Janis Pagel, Nils Reiter, Ina Rösiger, and Sarah Schulz. 2020. Annotation als flexibel einsetzbare Methode. In: N. Reiter, A. Pichler, and J. Kuhn, eds., *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*, pages 125–142. Berlin.

Ranka Stanković, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. Named Entity Recognition for Distant Reading in Several Languages. In: G. Pálko, ed., *DH_Budapest_2019*. Budapest, ELTE. http://elte-dh.hu/dh_budapest_2019-abstract-booklet/

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named Entities in Czech: Annotating Data and Developing NE Tagger. In: V. Matoušek, P. Mautner eds., *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3–7, 2007. Proceedings*. Berlin – Heidelberg, Springer-Verlag. https://ufal.mff.cuni.cz/~zabokrtsky/publications/papers/tsd07-namedent.pdf

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. In: *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012, Ljubljana, Slovenia: proceedings of the 15th International Multiconference Information Society - IS 2012, volume C*, pages 191–96, Ljubljana, Institut Jožef Stefan.

Katja Zupan, Nikola Ljubešić, and Tomaž Erjavec. 2017. Annotation guidelines for Slovenian named entities: Janes-NER. *Technical report, Jožef Stefan Institute, September*. https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1.1.pdf.