



HAL
open science

Annotation Guidelines for Corpus Novelties: Part 1 – Named Entity Recognition

Arthur Amalvy, Vincent Labatut

► **To cite this version:**

Arthur Amalvy, Vincent Labatut. Annotation Guidelines for Corpus Novelties: Part 1 – Named Entity Recognition. Laboratoire Informatique d'Avignon. 2024. hal-04715338v2

HAL Id: hal-04715338

<https://hal.science/hal-04715338v2>

Submitted on 2 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Annotation Guidelines for Corpus *Novelties*: Part 1 – Named Entity Recognition

Version 1.0.1

11 June 2024

Arthur Amalvy & Vincent Labatut

Laboratoire Informatique d'Avignon – LIA UPR 4128, Avignon, France

Abstract

The *Novelties* corpus is a collection of novels (and parts of novels) annotated for Named Entity Recognition (NER). This document describes the guidelines applied during its annotation. It contains the instructions used by the annotators, as well as a number of examples retrieved from the annotated novels, and illustrating expressions that should be marked as entities as well as expressions that should *not*.



Contents

Abstract	1	5	Organization Entities (ORG)	19
Contents	2		5.1 Proper Nouns	19
1 Introduction	3		5.2 Definite Descriptions	19
1.1 Notion of <i>Named Entity</i>	3		5.3 Disjointed Entities	19
1.2 Considered Entity Types	4	6	Group Entities (GRP)	21
1.3 Document Conventions	5		6.1 Family Names	21
1.4 Organization	5		6.2 Demonyms & Ethnonyms	21
2 General Principles	6		6.3 Definite Descriptions	22
2.1 Nested Entities	6		6.4 Enumerations	22
2.2 Definite Descriptions	7	7	Miscellaneous Entities (MSC)	23
2.3 Determiners	8		7.1 Temporal Entities	23
2.4 Parts of Names	8		7.2 Cultural Assets	24
2.5 Misspelled Names	8		7.3 Named Artefacts	24
3 Character Entities (CHR)	10		7.4 Other Entities	25
3.1 Proper Nouns	10	8	Type Confusion	26
3.2 Presence vs. Evocation	10		8.1 Characters vs. Locations	26
3.3 Definite Descriptions	11		8.2 Characters vs. Organizations	26
3.4 Titles & Honorifics	12		8.3 Characters vs. Miscellaneous	27
3.5 Nicknames	12		8.4 Locations vs. Organizations	27
3.6 Societal Roles	14		8.5 Locations vs. Miscellaneous	28
3.7 Personification	14		8.6 Organizations vs. Groups	28
3.8 Disjointed Entities	15	9	Concluding Remarks	29
4 Location Entities (LOC)	16	A	Version History	30
4.1 Proper Nouns	16			
4.2 Parts of Names	17	B	Todo List	31
4.3 Definite Descriptions	17			
4.4 Nicknames	18		References	32

1 Introduction

This document aims at providing instructions for the annotation of named entities in the *Novelties* corpus¹. The corpus itself will be the object of a separate description. It was constituted mainly to fulfill two goals: in the short term, train and test NER methods able to handle *long* texts, and in the longer term, be used to develop *Renard* [3], a pipeline aiming at extracting *character networks* from literary fiction. This pipeline includes several processing steps after the NER, including coreference resolution and character unification. Character networks can be used to tackle a number of tasks, including the assessment of literary theories, the level of historicity of a narrative, detecting roles in stories, classifying novels, identify subplots, segment a storyline, summarize a story, design recommendation systems, align narratives, etc. See the detailed survey of Labatut and Bost [11] for more information regarding character networks.

This context drives the elaboration of the corpus, which explains why it exhibits certain differences with many similar NER corpora, such as *CoNLL-2003* [17] or *OntoNotes v5* [20]. We originally based *Novelties* on the literary corpus from Dekker et al. [6] as we describe in Section A of the appendix. Note that there are other literary NER corpora (cf. [10] for a comparison), but they do not contain long texts [4, 6] and/or do not fit our needs [18]. In addition, our end goal and the architecture of our pipeline affects the perimeter of what we consider to be a named entity, a point that we discuss further in Section 1.1. Finally, these aspects also require us to put more emphasis on certain entity types, in particular characters, as explained in Section 1.2. Our guidelines are based on similar instructions previously written for other annotation campaigns and corpora, both in French [1, 14, 15] and English [4, 5, 12]. We adapted them to fit our case and requirements.

1.1 Notion of *Named Entity*

Historically, a named entity is a lexical unit of interest, which traditionally corresponds to a proper noun [7], and refers to an entity from the real world [1]. Certain authors, such as Alrahabi et al. [1] and Bamman et al. [4], use more relaxed definitions, including (proper) definite descriptions [7] in their annotations. A *definite description* is an expression of the form *determiner + noun phrase*, such as *the President of the United States*. A *proper* definite description allows identifying a *unique* entity, e.g. *the 42nd President of the United States*. Bamman et al. [4] call them *common entities*, by opposition to named entities.

On the one hand, we do not want to systematically annotate such expressions, because this quickly leads to *nested* annotations. For instance, in the above example, **United States** is contained in **President of the United States**. Such nested entities, in turn, cause a number of technical complications. First, they make it much harder to define simple and consistent annotation rules: see the many rules and exceptions in the Quaero guidelines [14], for instance. Second, they do not allow the traditional tags-based representation on which many models are based. But on the other hand, in certain novels, some major characters are exclusively referred to through definite expressions. Of course, we do not want to miss any important character, even if it is not properly named in the story. For this reason, there are some exceptions for which we annotate definite expressions in addition to proper nouns.

Even with these clarifications, it is not always obvious to determine what is a named entity and what is not. As stated by McDonald [13], one can distinguish two types of evidence that are helpful in order to come to a decision regarding the annotation of an entity. *Internal* evidence is directly present in the expression of interest. It can consist of criteria such as capitalization, the inclusion of a known name or the presence of titles or abbreviations for corporation types such as “*Ltd.*”. By comparison, *external* evidence is found in the context surrounding the expression of interest: in a novel, the description of a character’s actions is evidence that some of their mentions are named entities. We conducted a first annotation pass over a few chapters to get a beta version of our corpus (cf. Appendix A), and leveraged this experience to identify four types of evidence (two internal and two external) that help to make this decision in the context of novels. While internal evidence is easily interpreted locally, the interpretation of external evidence may necessitate access to knowledge from the entire novel or from its universe, something that can often be done through the help of online

¹<https://github.com/CompNet/Novelties>

wikis dedicated to specific literary universes. Most of the four types of evidence we describe below are neither necessary nor sufficient to really detect an entity, but rather serve as important hints for annotators.

Capitalization (internal evidence) The expression is (possibly partly) capitalized. In English and French, proper nouns are capitalized, so this is a good indication that the expression is a named entity. Some authors are very liberal in their use of capitals, though, so an upper-cased initial does not necessarily mean that the expression is a named entity.

Self-Sufficiency (internal evidence) The expression alone has a meaning and refers to an entity or a group of entities. Contrary to the other three factors, this one is necessary for an entity to be considered as valid. This point helps when dealing with parts of proper nouns, e.g. first names.

Unicity (external evidence) The expression serves to uniquely identify an entity (or possibly a group of entities). This point is related to the notion of *proper* definite expression. The goal here is to exclude generic expressions such as a policeman, a little girl, etc. Unicity is closely linked to the fact that we consider the universe of a literary text as a *closed world*, where two distinct entities would be clearly identified differently, and may not apply to other kinds of texts.

Frequency (external evidence) The expression is frequently used to refer to the entity. It is not possible to define an absolute threshold above which an expression should be considered as frequent enough, so this is left to the appreciation of the annotator. The point is to catch expressions such as nicknames, that are not proper nouns, but still used to refer to certain characters.

1.2 Considered Entity Types

Detecting a named entity in the text is only the first part of the task: the second part consists in determining its category or type. Historically, NER authors are interested in distinguishing between persons, locations, and organizations. Moreover, in many works, the expression *named entity* refers not only to proper nouns, but also to temporal expressions (dates and times) and quantities (amounts of money, percentages) [7].

In our case, because of our end goals, we have a strong focus on one specific category of named entities: those referring to characters. However, we decided to include other types of entities in our annotation, too, as this task does not require much additional work, while increasing the value of the corpus.

Characters This type supersedes the traditional *Persons* category: it includes the usual *anthroponyms* (names of people), but also other entities referring to other types of agents, possibly non-human, such as animals, robots, magical creatures, etc.

Locations This is the standard category used in many corpora, which includes *toponyms* (i.e. names of places). We do not distinguish between natural (river, mountain, island...) and artificial (city, street, building, etc.) locations.

Organizations This type is also standard, in the sense that it appears in many corpora. It gathers all named entities referring to explicitly organized institutions, such as a government, a company, etc. It excludes informal groups such as families, *ethnonyms* (names given to the members of ethnic groups), and *demonyms* (names given to the inhabitants of some places).

Groups This type is more uncommon: it gathers informal groups of persons referred to under the same umbrella name, including family names, ethnonyms, and demonyms. It is used when the text does not identify its members individually. Traditionally, NER corpora annotate these kinds of expressions as *Persons*, but we differentiate them in order to facilitate character detection.

Miscellaneous This heterogeneous category gathers *praxonyms* (names referring to historical, cultural, commercial or sport events), *ergonyms* (names of objects and man-made products) provided they are unique, *phenonyms* (names of meteorological events such as tempests), and titles of intellectual or cultural works (such as books and movies).

1.3 Document Conventions


In the rest of the document, we provide a number of examples to illustrate our guidelines. Offline examples are represented in gray frames, as follows:

This is an example.

Inline examples are inserted in the text using a sans-serif font, e.g. **the Emperor**.

We use boxes to highlight entities in the text. Their color indicates the entity type, and we increase the transparency to distinguish between entities that are the focus of an example, and those that just happen to appear in the example :

<i>Characters</i>	CHR	Red		
<i>Locations</i>	LOC	Blue		
<i>Organizations</i>	ORG	Brown		
<i>Groups</i>	GRP	Orange		
<i>Other entities</i>	MSC	Green		

When we want to specifically highlight that an expression should *not* be annotated, we use a gray box, e.g. .

1.4 Organization

In the following, we first describe guidelines that are applicable for all the types of entities that we annotate (Section 2). Afterward, each remaining section is dedicated to a specific type of entity: characters (Section 3), locations (Section 4), organizations (Section 5), groups (Section 6), and other entities (Section 7). In Section 8, we discuss the possible confusion between certain types of entities. Finally, Section 9 provides our concluding remarks, and Appendix A gives the history of this document.

2 General Principles

This section describes some general rules that apply to all entities, independently of their type. There are some exceptions to these rules, which are described latter, in type-specific sections.

Annotation is conducted manually, and the human aspect of this process must therefore be taken into account, in addition to the more technical points mentioned before. In particular, the annotation rules must be clear enough, simple enough, and not too numerous, in order to avoid human errors. For this reason, we sometimes sacrifice accuracy if it allows providing the annotator with simpler or more consistent instructions. We also noticed that human readers tend to *want* to annotate certain expressions. Not providing any instructions for these cases, or forbidding to annotate them, can be counterproductive. For instance, in an earlier version of these guidelines, we did not annotate languages at all. But annotators did it anyway, probably because the same word is generally used to refer to a language and its speakers (and the latter were already annotated as a group). Therefore, in a subsequent version of the guidelines, we included language annotation as a miscellaneous entity.

2.1 Nested Entities

Nested entities are entities within entities, e.g. **President of the United States**: the **United States** part is an organization, but the whole expression is a character. As explained in Section 1, detecting *nested* entities is quite different from standard NER, almost a different problem. For this reason, as in many guidelines [9], we focus only on *flat* entities in this corpus. This means making a choice between the different levels of entities within the nested structure.

General Rule Certain authors keep the innermost entity, e.g. [1], whereas other focus on the outmost, as explained in [9]. It appears that, in novels, the outmost entity is generally the entity participating in the action, or the entity that the author wants to mention. By comparison, the innermost entity brings some additional information allowing to identify the outmost entity, but is not the main object of the mention. For this reason, our general rule is to annotate the outmost entity.

Examples See this case from Aldous Huxley’s *Brave New World*, for instance:

the **CHR** **Director of Hatcheries and Conditioning** entered the room, in the scarcely breathing silence

This is one of the main characters in the novel. He is also called by his first name (**Thomas**) and nickname (**Tomakin**), but more importantly by the acronym **D.H.C.**, which shows the importance of including the organization **Hatcheries and Conditioning** in the annotation. Annotating the innermost entity, i.e. only this organization, would be much less informative.

Interestingly, his full title is actually **Director of Hatcheries and Conditioning of Central London**, but it is never used under this form. We rather find:

The **CHR** **D.H.C. for Central London** always made a point of personally conducting his new students

Although there is no other **D.H.C.** in the novel, and therefore no possible confusion, we annotate **Central London** as a part of the entity, for the sake of consistency.

Nestedness also concerns other types of entities than characters. Regarding organizations, we can mention the **Knight of the Vale**, from *A Song of Ice and Fire*, the Fantasy series by George R. R. Martin:

“The **ORG** **Knights of the Vale** could make all the difference in this war,” said **CHR** **Robb** [...]

Where the **Vale** is a location. The full form of this name is actually the **Vale of Arryn**, where **Arryn** refers to a person, making this a nested entity, too:

[...] there were no friends of the **GRP** **Lannisters** in the **LOC** **Vale of Arryn**.

Counterexamples There are some exceptions to this general rule, though. When the outmost entity is deemed too generic and/or not frequent, we annotate the outmost *valid* entity. In practice,

there are often only two levels of nestedness, so this amounts to selecting the innermost entity. See this example from Glen Cook’s *The Black Company*:

[...] CHR Bucket’s answer. “He wanted to pick off NOT Black Company guys. That’s obvious.”

Here, one could want to annotate **Black Company guys** as a group. However, if the expression may be frequent, it is also very imprecise. Over the series, it is likely to refer to several distinct subsets of the people constituting the **Black Company**. For this reason, in this case, we would annotate only **Black Company**, as an organization.

In another example, this time from *A Song of Ice and Fire*, the situation is different :

“CHR Daenerys Targaryen has wed some NOT Dothraki horselord. [...] Shall we send her a wedding gift?”

Here, **Dothraki horselord** refers to the character Khal Drogo. It is precise enough, but used only a few times over the whole series, so not considered as frequent. Consequently, we only annotate **Dothraki**, as a group (see Section 6.2).

The same remarks apply for other entity types. See this example from *The Black Company*, that involves a location:

[...] the avenue [...] winds from the LOC Customs House uptown to the NOT Bastion’s main gate.

The expression **Bastion’s main gate** could be annotated as a location, because this name is quite precise. However, it is not frequent at all. We consider it more informative to annotate only **Bastion** as a location.

Similarly, in the following example from J. K. Rowling’s *Harry Potter*:

[...] and a moment later, NOT Dudley’s best friend, CHR Piers Polkiss, walked in with his mother.

Here, **Dudley’s best friend** refers to a specific entity, which is precisely identified. However, this expression is not frequent (used only once), which is why we annotate **Dudley** as a character. We generally do not annotate expressions that refer to an entity through its relation to another entity (here: **Piers** through **Dudley**), except when it is very frequent.

Enumerations Nested entities should be distinguished from *enumerations* of entities that share a part of their name. In this case, we annotate each entity separately, provided the expression is sufficient to recognize it:

[...] with both hands and said, “In the name of CHR Robert of the House Baratheon, the First of his Name, CHR King of the Andals and the Rhoynar and the First Men, CHR Lord of the Seven Kingdoms and CHR Protector of the Realm, by the word of CHR Eddard of the House Stark, CHR Lord of Winterfell and CHR Warden of the North, I do sentence you to die.”

Otherwise, depending on the case, some parts may not be annotated, or the whole expression may be considered as a group (see Section 6).

2.2 Definite Descriptions

General Rule As stated in the introduction, we generally do not annotate definite descriptions. This purposely excludes generic mentions, such as **the boy** or **the city** in the following excerpts of Brandon Sanderson’s *Elantris*:

- NOT The boy, as if realizing his chance would soon pass, stretched his arm forward [...]
- He had hoped NOT the city would grow less gruesome as he left the NOT main courtyard [...]

There are several exceptions to this general rule, though, which we detail in the entity-specific sections. In principle, if a definite expression is frequently used to refer to a specific entity, then it can be annotated. For instance, for a character, this expression could be considered as a nickname (see Section 3.5).

Capitalization As discussed in Section 1.1, capitalization is a good indication that a definite expression should be annotated as an entity. For instance, in the below sentence from the Fantasy series *The Black Company*, the capitals hint at a LOC entity, and not just any undifferentiated hill:

“We’re going to the LOC Necropolitan Hill to eyeball that GRP forvalaka tomb.”

However, this principle does not always apply, as the use of capitals vary widely from one author to the other. For instance, in Aldous Huxley’s *Brave New World*:

[...] until at last they were dancing in the crimson twilight of an NOT Embryo Store [...]

The use of the indefinite article *an* clearly indicates that, despite the capitalization, the author mentions an arbitrary *embryo store*, and not a specific, recurrent place. The frequency rule mentioned above helps deciding whether that expression is recurring or not.

2.3 Determiners

General Rule Except when they are explicitly part of the name they are attached to, we do not annotate determiners in front of entities. This is because some of these entities can be referred to without determiners.

Examples In *The Black Company* series, the eponymous organization is referred too as the **Black Company** in the novels, but also sometimes as only **Black Company**, depending on context. This shows that the determiner is not crucial to the designation of this entity. Therefore, we keep its smallest consistent expression:

The ORG Black Company does not suffer malicious attacks upon its men.

In the same series, the **Lady** is one of the main characters. The same rule applies:

ORG Oar had not yet seen any of the CHR Lady’s champions.

Counterexample On the contrary, sometimes the determiner is part of the name, in which case we include it in the annotation to be consistent the rule of self-sufficiency:

He took the train to LOC The Hague.

2.4 Parts of Names

General Rule Sometimes, an entity is mentioned through a part of its name, instead of its full name. In this case, we annotate this part, but only under the condition that this incomplete name is sufficient to identify the entity, and that it is used frequently.

Examples This particularly apply to characters, when using only a first name, e.g. in the *Harry Potter* series, the eponymous character is often called only by his first name only:

“I’ve come to bring CHR Harry to his aunt and uncle. They’re the only family he has left now.”

But the situation also happens for other types of entities, in particular organizations. See this example from *The Black Company*, where the eponymous organization is only called by an abbreviated (but unambiguous) version of its full name:

He and CHR One-Eye have been with the ORG Company a long time.

2.5 Misspelled Names

General Rule There are several situations where the name of an entity is not correctly spelled, in which our general rule is to annotate the mention as if it was correctly written.

Examples In Joe Abercrombie’s *The Blade Itself*, one of the character has a speech impediment, and some character names are written in a way that reflect this trait:

‘Ith CHR Theverar,’ [...] by which CHR Glokta understood that CHR Severard was at the door.

Here, the name of character **Severard** is rendered as **Theverar**. Same thing with **Felix Grandet**’s fake stuttering in Balzac’s *Eugénie Grandet*:

“CHR M-m-monsieur de B-B-Bonfons,” –for the second time in three years CHR Grandet called [...]

In the following example from Herman Melville’s *Moby Dick*, the non-standard spelling is rather a matter of accent:

“Passed one once in LOC Cape-Down,” said the old man sullenly.

The speaker is **Fleece**, the cook of the ship, and by **Cape-Down** he means **Cape Town**.

Sometimes, the misspelling of a name can be due to the speaker’s error. Here is an example from *Moby Dick*, in which **Captain Peleg** makes a mistake when saying the name of a character:

I say, CHR Quohog, or whatever your name is, did you ever stand in the head of a whale-boat?

The character’s actual name is **Queequeg**, but we annotate the wrong name like before, as it is obvious who **Peleg** talks to, given the context.

3 Character Entities (CHR)

The standard approach to annotate characters would be to consider them as persons, and to use the very common PER tag. However, as remarked by Bamman et al. [4] when annotating LitBank, characters are not necessarily persons. For this reason, they use a wider definition and annotate all entities who “engage in dialogue or have reported internal monologue, regardless of their human status”. They still consider them formally as persons, though, and use PER.

Many works of fiction involve non-human agents that have an effect on the story. Therefore, we go further, and annotate any individual entity with some form of sentience and agency in the plot. As a consequence, contrary to other classical NER datasets, we do not annotate *persons*, but rather *characters*. This wider concept encompasses not only human entities, but also other sentient entities such as animals, mythical creatures, magical weapons, robots... To stress this difference, we use a specific tag, CHR, instead of the traditional PER.

3.1 Proper Nouns

General Rule We annotate proper nouns that refer to individual characters, e.g. in Jane Austen’s *Emma*:

CHR Emma Woodhouse, handsome, clever and rich, with a comfortable home [...]

Parts of Names As explained in Section 2.4, it is possible to annotate isolated parts of the name, provided they allow identifying the character without ambiguity. For example, also from Jane Austen’s *Emma*:

[...] and CHR Emma could not but sigh over it, and wish for impossible things [...]

In Dostoevsky’s *The Double*, the main character Yakov Petrovich Golyadkin is mentioned by various combinations of parts of his name:

- A man with a message. “Is CHR Yakov Petrovitch Golyadkin here?” says he.
- “He’s still at the office and asking for you, CHR Yakov Petrovitch.”
- “You’re mischievous CHR brother Yakov, you are mischievous!”
- When he had made this important discovery CHR Mr. Golyadkin nervously closed his eyes [...]

Antonomasia Certain authors mention person names through *antonomasia*, a metonymy consisting in using a proper noun as a common name. It is questionable whether the mentioned person should be considered as a proper character, or just a cultural reference. We decide to annotate such cases when the author uses an initial capital.

Here is an example from *Moby Dick*:

I laugh and hoot at ye, ye cricket-players, ye pugilists, ye deaf CHR Burkes and blinded CHR Bendigoes!

where Burket and Bendigo are 19th century boxers.

Groups If a name refers to several characters at once, we annotate the entity as a *group* instead (see Section 6. Consider, for instance, this excerpt of *Harry Potter*:

They didn’t think they could bear it if anyone found out about the GRP Potters.

Here, the Potters collectively refers to James Potter, Lily Potter and Harry Potter.

3.2 Presence vs. Evocation

Generally speaking, it is possible to explicitly annotate whether a character is present or evoked, as in certain guidelines like [1]. In the former case, the character is physically present and participating in the scene, like in this example from Aldous Huxley’s *Brave New World*:

CHR John began to understand. “Eternity was in our lips and eyes,” he murmured.

In the latter case, the character is just brought up by other entities in their absence, as in this example from the same novel:

“I suppose **CHR** John told you. What I had to suffer –and not a gramme of **MSC** soma to be had.

General Rule In the context of these guidelines, we assume that distinguishing both types of entity mentions (presence vs. evocation) can be done in a later step of our pipeline mentioned in Section 1. Consequently, as a general rule, we annotate indifferently both situations.

Interjections However, it happens that the name of a person is used as an interjection. This is particularly the case of divinities, e.g. in Dostoevsky’s *The Double*:

“I’m very well, thank **CHR** God, **CHR** Anton Antonovitch,” said **CHR** Mr. Golyadkin, stammering.

Ideally, it would make sense to ignore such mentions, as **God** is not a character actually participating in the story, in this case. However, this decision could be considered too subjective. Therefore, to simplify the annotation task, we decide to annotate all these invocations as characters too. A specific step of our pipeline could determine later whether one entity should be kept, depending on it being a proper character.

The novel *Brave New World* exhibits an interesting case of divine invocation, as **Henri Ford**’s name is almost always used in place of **God**’s, as an interjection:

“Oh, **CHR** Ford!” he said in another tone, “I’ve gone and woken the children.”

As explained before, we annotate **Ford** as a character even if he does not intervene directly in the story.

Special Case The distinction is sometimes more fuzzy, as certain novels involve divinities as characters while also using their names in interjections. This is the case of **God** in Douglas Adams’s *Hitchhiker’s Guide to the Galaxy*:

- **CHR** God, what a terrible hangover it had earned him though.
- “Oh dear,” says **CHR** God, “I hadn’t thought of that,” and promptly vanishes in a puff of logic.

Another example is **Hood**, the god of death in Steven Erikson’s *Malazan Book of the Fallen* Fantasy series:

- Clear the streets? How in **CHR** Hood’s name do we manage that?
- **CHR** Hood glanced down at the spatter on its frayed robes.

It is difficult for the reader to guess whether these divinities are supernaturally permanently listening to the people, and hear them pronouncing their names. For the sake of simplicity, we annotate not only situations where the divinity appears explicitly as a character, but also interjections, as in the above examples.

3.3 Definite Descriptions

General Rule As explained more generally in Section 2.2, contrary to Bamman et al. [4], we do not annotate definite descriptions, in general:

He was still determined to not mention anything to **NOT** his wife.

Exceptions Some characters are only mentioned using a definite description. For instance, this is particularly true for Carlo Collodi’s *The Adventures of Pinocchio*, in which many characters are never properly named: the Judge, the Innkeeper, the Falcon, the Owl, the Farmer, etc. In these cases, we annotate such expressions:

The CHR Judge was a Monkey, a large Gorilla [...] The CHR Judge listened to him with great patience.

Another example is the already previously discussed Director of Hatcheries and Conditioning in *Brave New World*, which is very often called just Director. There are just two other directors in the whole novel, and each one is mentioned only once. For this reason, we annotate Director as a character, as there is no ambiguity, and the use is frequent:

Tall and rather thin but upright, the CHR Director advanced into the room.

This case is also related to the situation where we annotate societal roles (cf. Section 3.6).

3.4 Titles & Honorifics

General Rule We annotate honorific titles as part of CHR entities, even when they are lowercase. This choice is driven by our end application (character network extraction), where titles carry important information: they can be used to disambiguate between several characters, or to detect their gender.

Examples Consider the following sentences, for which the general rule directly applies:

- We talked it all over with CHR Mr. Weston last night.
- CHR Lord Eddard Stark dismounted and his ward CHR Theon Greyjoy brought forth the sword.

The first sentence comes from *Emma*, and the second from *A Song of Ice and Fire*.

Titles are sometimes necessary to distinguish between certain characters. For instance, in Balzac’s *Eugénie Grandet*:

Monsieur and CHR Madame Guillaume Grandet, by gratifying every fancy of their son [...]

Here, Monsieur is not annotated because it cannot stand by itself (see the *Isolated Titles* paragraph, below). Madame Guillaume Grandet refers to Guillaume Grandet’s wife, and without the title Madame, this mention would be mistakenly understood as referring to her husband.

During the elaboration of these guidelines, we considered annotating titles separately from characters, as a distinct entity type. However, this would be very close to handling nested entities, which we want to avoid (see Section 2.1).

Titles include family-related relations. For instance, in *A Song of Ice and Fire*, Jon Snow is the nephew of Benjen Stark:

CHR Uncle Benjen studied his face carefully. “The LOC Wall is a hard place for a boy, CHR Jon.”

Honorific titles can be completely fictional, like for instance High Fist in the *Malazan Book of the Fallen* Fantasy series:

CHR High Fist Dujek Onearm entered, the soap of his morning shave still clotting the hair in his ears.

Isolated Titles Since entities must be self-sufficient, we do not annotate isolated titles as CHR, in general:

Thank you, NOT sir! Please, come again.

We make an exception: it is possible to consider such an isolated title as unique and frequent, similarly to what we do with nicknames in Section 3.5. This case is very close to the annotation of societal roles that we describe in Section 3.6.

3.5 Nicknames

General Rule We annotate nicknames if they are frequent and allow identifying the entity in a reasonably unique way (see Section 1.1).

Examples For instance, in *A Song of Ice and Fire*, character Mance Rayder is often referred to as follows:

[...] he was a wildling, his sword sworn to CHR Mance Rayder, the CHR King-Beyond-the-Wall.

Another example is White Whale (although this could also be considered a definite description), an expression frequently used to refer to the eponymous whale in *Moby Dick* (note the capitalization):

[...] many brave hunters, to whom the story of the CHR White Whale had eventually come.

It is possible that the nickname concerns only a part of the original name, e.g. the first name for Eddard Stark:

“You’re CHR Ned Stark’s bastard, aren’t you?” CHR Jon felt a coldness pass right through him.

Incidentally, observe that we do not annotate the whole expression Ned Stark’s bastard as a character, because it is not frequent enough (see Section 2.1).

Some characters have several nicknames. For example, in the *Harry Potter* series, the main antagonist, Tom Marvolo Riddle, is known under various nicknames: Lord Voldemort, He-Who-Must-Not-Be-Named, the Dark Lord, and others. We annotate all significant nicknames:

Rejoice, for CHR You-Know-Who has gone at last!

It is worth stressing that some characters are referred to using *only* their nicknames, so discarding these would mean missing these characters entirely. In *The Black Company*, one of the main characters is called the Captain, and his true name is *never* revealed:

LOC Beryl had ground our spirits down, but had left none so disillusioned as the CHR Captain.

Attributes When an attribute follows the name of the character, we treat the whole expression as a nickname, as in this example from J. R. R. Tolkien’s *The Lord of the Rings*:

It has seldom been heard of that CHR Gandalf the Grey sought for aid [...]

The attribute is sometimes itself the name of a distinct entity, so this is consistent with our decision to annotate the outmost entity in case of nested entities (cf. Section 2.1). Here are some examples from *A Song of Ice and Fire*:

- The one I want is with a highborn girl, the daughter of CHR Lord Stark of Winterfell.
- This is the will and word of CHR Robert of House Baratheon, the First of his Name [...]

Note that in the above examples, we do not annotate Winterfell as a location or House Baratheon as an organization.

Frequency We do not annotate very punctual nicknames or insults (even personalized ones). In the following example from *A Song of Ice and Fire*, we would annotate only the word Arya:

CHR Jeyne used to call her NOT Arya Horseface, and neigh whenever she came near.

And in this example from *Harry Potter*, only Potter:

“NOT Saint Potter, the GRP Mudbloods’ friend,” said CHR Malfoy slowly.

In this sentence from *The Black Company*, Goblin uses a creative nickname to provoke his friend One-Eyed:

CHR Goblin chortled, “You ain’t winning even when you deal, NOT Maggot Lips. [...]”

Here, we would not annotate any part of the expression Maggot Lips, which is used only once in the whole book.

3.6 Societal Roles

General Rule We annotate societal roles according to the same general principles as before, i.e. when they refer to a specific character without ambiguity, and they are mentioned frequently enough. Put differently, we consider them a bit as if they were nicknames.

Examples For instance, in Robin Hobb’s *Farseer Trilogy*, **King Shrewd** is the grandfather of the protagonist, and an important character. Moreover, he is the only king for most of the first book, therefore we annotate as follows:

But our father the **CHR King** is not a hasty man, as well we know.

On the contrary, some roles are too generic or too common to be annotated, as they do not ensure the unicity of the entity:

He pointed, and **CHR Arya** saw it. The body of the **NOT soldier**, shapeless and swollen.

Capitalization Capitals are a good indication to detect important societal roles, however many words are capitalized without having such meaning. Moreover, the use of capitals varies significantly from one author to the other. For instance, in *Brave New World*, Aldous Huxley capitalizes a lot of expressions:

The **NOT Chief Bottler**, the **NOT Director of Predestination**, 3 **NOT Deputy Assistant Fertilizer-Generals**

Each one of these three expressions appears only once or twice in the whole novel, so we do not consider them as entities.

3.7 Personification

General Rule As per our wide definition of what a character is, we annotate personified animals or items as **CHR** when relevant.

Artificial Beings This includes robots and other manufactured beings such as Marvin the Paranoid Android from *The Hitchhiker’s Guide to the Galaxy* series:

[...] **CHR Marvin** managed to convey his utter contempt and horror of all things human.

On the same note, the sentient sword **Stormbringer** in Michael Moorcock’s *Cycle of Elric* has its own will, so we annotate it as a **CHR** entity:

CHR Stormbringer whined almost petulantly, like a dog stopped from biting an intruder.

Cheeses are usually inanimate objects, which makes **Horace the Cheese** a more extreme example of personification. This character appears in the *Discworld* series by Terry Pratchett:

CHR Horace was the only cheese that would eat mice and, if you didn’t nail him down, other cheeses.

Animals Lewis Carroll’s *Alice in Wonderland* involves many personified animals, such as **Mouse**:

CHR Mouse, do you know the way out of this pool?

However, we do not annotate common entities without any significant role in the story:

[...] and she soon made out that it was only a **NOT mouse** [...]

Abstract Concepts Very often, abstract concepts such as fate or death are personified in novels. We annotate them only if they are actual characters. Consider for instance this example from Terry Pratchett’s *The Color of Magic*:

CHR Death, insofar as it was possible in a face with no movable features, looked surprised.

Death is a well-known character of the Discworld series, so we annotate him. On the contrary, the following excerpt from *Moby Dick* is a counterexample:

Of such a letter, NOT Death himself might well have been the post-boy.

This strong personification might suggest that **Death** is a proper character of the novel, but this is not the case, so we do not annotate it.

3.8 Disjointed Entities

General Rule Disjointed names are annotated as characters if each individual entity is self-sufficient, i.e. if the expression referring to this entity is enough to recognize them.

Examples Here is an example from Jane Austen’s *Pride & Prejudice*:

CHR Elizabeth, CHR Kitty and CHR Lydia Bennet are sisters.

In the above sentence, we assume that both **Elizabeth Bennet** and **Kitty Bennet** can be identified by their first names. Even though the family name **Bennet** is implicitly shared by all three mentions, our annotation only associates it to the last character.

This stays true if the shared part of the name is plural, as in this example from Alexandre Dumas’ *The Three Musketeers*:

[...] [t]o prevent CHR MM. Bassompierre and CHR Schomberg from deserting the army, a separate command had to be given to each.

Here, **MM.** stands for **Misters** (plural), but we associate it only with **Schomberg** in our annotation.

Counterexample Otherwise, the entire span is annotated as a group entity (cf. Section 6). See this sentence, also from *Pride & Prejudice*:

[...] CHR Mr. Collins’s scruples of leaving GRP Mr and Mrs Bennet for a single evening during his visit

In the below example, **Mr** is not self-sufficient, so we annotate the whole expression as **GRP**.

4 Location Entities (LOC)

We consider that the term *location* denotes physical or metaphysical entities that embody a specific place or region. Locations are devoid of any agency: if an entity is described as performing an active action, it cannot be a LOC entity. In particular, names typical of a location but that refer to geopolitical entities in this context should be annotated as **ORG** (cf. Section 8.4).

4.1 Proper Nouns

General Rule A number of locations are referred to using a proper noun, in which case we annotate them.

Physical Locations Physical entities include neighborhoods such as **Flea Bottom** in *A Song of Ice and Fire*, cities such as **London**, regions such as **Derbyshire**, countries such as **England**, continents such as **Westeros** (also from *A Song of Ice and Fire*):

- She had been sleeping in **LOC Flea Bottom**, on rooftops and in stables [...]
- [...] I did not feel quite certain that the air of **LOC London** would agree with **CHR Lady Lucas**.
- [...] not all his large estate in **LOC Derbyshire** could then save him [...]
- [...] with a gesture whose significance nobody in **LOC England** but the **CHR Savage** now understood
- Remember, child, this is not the iron dance of **LOC Westeros** we are learning [...]

These examples come from *A Song of Ice and Fire* (#1, #5), *Pride and Prejudice* (#2, #3), and *Brave New World* (#4).

Physical locations also include man-made structures such as buildings like **Harrenhal**, a fortress from *A Song of Ice and Fire*:

[...] now he's marching north toward **LOC Harrenhal**, burning as he goes.

There are also commercial buildings like the **Cattery**, a brothel in *A Song of Ice and Fire*:

[...] and the brothel called the **LOC Cattery**, where he got strange looks but no help.

Physical locations can refer to natural structures or areas, e.g. **Blackwater Bay** in *A Song of Ice and Fire*:

LOC Blackwater Bay was rough and choppy, whitecaps everywhere.

Similarly to what we do for characters' titles (cf. Section 3.4), we include qualifiers in the annotation, such as **Bay** in the previous example, or **Southern** in the following excerpt of *The Black Company*:

LOC Southern Forsberg remained deceptively peaceful.

Finally, stars and planets can also be considered as locations. See this example from Douglas Adams' *Hitchhiker's Guide To The Galaxy*:

[...] **CHR Ford Prefect** was in fact from a small planet somewhere in the vicinity of **LOC Betelgeuse**.

In certain cases, the celestial object is not a place, though, so context must be considered. For instance, in *Moby Dick*:

What a fine frosty night; how **NOT Orion** glitters; what northern lights!

Here, Orion is just a light in the sky. Similarly, we would not annotate the sun or the moon as locations, unless they are used as such.

Metaphysical Locations Metaphysical entities can be very diverse in nature. Some good examples are the **L-Space** from the *Discworld* series, which is a place connecting all libraries across time and

space:

All libraries everywhere are connected in LOC L-space. All libraries. Everywhere.

The warrens from the *Malazan Book of the Fallen*, such as Omtose Phellack, are some sort of pocket worlds, and could also be considered as metaphysical locations. Those are simultaneously places that connect with the physical plane, and the source of magic in this lore:

My Warren touches LOC Omtose Phellack. I can reach it, CHR Adjunct. Any GRP T'lan Imass could.

4.2 Parts of Names

General Rule In accordance to our general principle from Section 2.4, we annotate parts of location names, under certain conditions. This is analogous to what we do with characters' titles (cf. Section 3.4) and societal roles (Section 3.6).

Noun Modifiers Certain location names are constituted of a common noun, acting as a noun modifier, and a proper noun. It is common for these locations to be referred to using only the former. For instance, the previously mentioned Vale of Arryn, from *A Song of Ice and Fire*, is frequently referred to simply as the Vale:

"A pity CHR Lysa carried them off to the LOC Vale," CHR Ned said dryly.

Similarly, in *Moby Dick*, the Massachusetts island named Martha's Vineyard is often simply called the Vineyard:

[...] once the bravest boat-header out of all LOC Nantucket and the LOC Vineyard; [...]

Exceptions It is important that the short form allows to uniquely identify the entity, and that it is frequently used. In the following example from *The Black Company*, avenue is used only thrice in the book, and to refer to two distinct avenues, so it should not be annotated when used by itself:

"We had come to the LOC Avenue of the Syndics', [...] There was a procession on the NOT Avenue."

Similarly, in this sentence from *A Song of Ice and Fire*, bay refers to Blackwater Bay:

[...] who would stand out in the NOT bay in case the GRP Lannisters had other ships hidden [...]

There are several other bays mentioned in the novels, so we do not annotate this word when used separately.

4.3 Definite Descriptions

General Rule As for other type of entities, we do not annotate definite descriptions, unless they have an important role in the story.

Examples Consider the following sentence:

I am going to NOT the lake, I'll be back late in the evening.

If this lake was the only lake mentioned in the novel, and if it was central to the story, then we would annotate it as LOC. A good example is the Wall in *A Song of Ice and Fire*, a monumental ice and rock structure spanning hundreds of kilometers:

There's not been a direwolf sighted south of the LOC Wall in two hundred years.

Numerical Expressions Certain expressions include numerical values: we annotate them too, due to the unicity they entail. For instance, from *Brave New World*:

- Told them of the test for sex carried out in the neighborhood of LOC Metre 200.
- Their wanderings [...] had brought them to the neighborhood of LOC Metre 170 on LOC Rack 9.

4.4 Nicknames

General Rule Although this is not as common as for characters, some locations are also referred to using nicknames. Like before, we annotate them if they are frequently used, and allow identifying the entity reasonably well.

Example Using the *Big Apple* instead of *New York* is a good example:

After years of dreaming, she finally arrived in the LOC Big Apple, ready to pursue her acting career.

5 Organization Entities (ORG)

We consider that an organization is an *institutional* entity: a state, a ministry, a guild... By comparison, informal groups such as families, demonyms, or ethnonyms, are annotated as groups instead (see Section 6).

5.1 Proper Nouns

General Rule As for the other entity types, we annotate all proper nouns referring to organizations.

Examples For organizations, proper nouns are not as common in novels as for characters and locations. For instance:

[...] **ORG Canonical** announced the release of their latest **MSC Ubuntu** update, promising new features.

Or the RAMJAC corporation, taken from Kurt Vonnegut's *Jailbird*:

That agency [...] is now a wholly-owned subsidiary of The **ORG RAMJAC Corporation**.

Here, we include Corporation in the annotation, similarly to what we do with honorific titles and qualifiers for other entity types (see Sections 3.4 and 4.2, for instance).

The Hogwarts houses from *Harry Potter* are also a good example of organizations possessing a proper noun:

He took off the hat and walked shakily towards the **ORG Gryffindor** table.

Although in this case, there is also a metonymy, as Godric Gryffindor is the founder of this house.

Groups The difference with GRP entities is not always obvious: the annotator must take into account the informal vs. institutional nature of the entity, as explained in Section 8.6.

5.2 Definite Descriptions

General Rule Very often, organizations in novels are referred to only with definite descriptions. Of course, we annotate these expressions, otherwise we would miss completely the corresponding entities.

Examples Here are two examples taken from *A Song of Ice and Fire*:

- [...] I could sweep the **ORG Seven Kingdoms** with ten thousand **GRP Dothraki** screamers.
- **CHR Theon** is the rightful heir, unless he's dead... but **CHR Victarion** commands the **ORG Iron Fleet**.

And another example from George Orwell's *1984*:

The **ORG Ministry of Peace** concerns itself with war [...], the **ORG Ministry of Love** with torture [...]

5.3 Disjointed Entities

General Rule Like for characters (Section 3.8), disjointed names are annotated as organizations if each individual entity mention is self-sufficient.

Example Here is an example from *The Blade Itself*:

[...] from [...] high-born nobodies to the great magnates of the **NOT Open** and **ORG Closed Councils**.

The sentence mentions two institutions: the Open Council and the Closed Council. Only the latter can be recognized in the text: Open is not enough to identify the former.

Counterexamples Unlike for characters, it is rarely the case that each individual entity mention is self-sufficient, because the common portion of the entity name is often necessary to recognize the organization.

See for instance this example from *Brave New World*:

Then came the ORG Bureaux of Propaganda by Television, NOT by Feeling Picture, and NOT by Synthetic Voice and Music respectively—twenty-two floors of them.

This sentence lists three distinct organizations: the Bureau of Propaganda by Television, the Bureau of Propaganda by Feeling Picture, and the Bureau of Propaganda by Synthetic Voice and Music. However, only the first mention is recognizable by itself, hence our single annotation.

Another example, this time from *The Blade Itself*:

He established the NOT Councils, NOT Closed and NOT Open, he formed the ORG Inquisition.

This sentence refers to character Bayaz creating three institutions: the Closed Council, the Open Council, and the Inquisition. The councils are not recognizable by using only the Closed and Open parts of their full names.

6 Group Entities (GRP)

We define group entities as informal gathering or sets of characters, that do not have any proper institutional existence. They are used when the concerned name does not refer to an individual character but several ones at once, while still providing sufficient information to be able to identify them relatively well.

The rationale for annotating groups is that some authors extract character networks that contain vertices representing such groups. For instance, when studying Homer’s *Iliad*, Venturini *et al.* [19] model certain Greek tribes using a single vertex (e.g. Myrmidons), while in [8], Falk represents bystanders collectively, using a specific vertex.

6.1 Family Names

General Rule We annotate family names as GRP entities when they refer to several members of that family.

Example In the below example, **Baggins** is a family name from *The Lord of the Rings*, and it is used to refer to the family as a whole:

But there you are: GRP **Hobbits** must stick together, and especially GRP **Bagginses**.

6.2 Demonyms & Ethnonyms

General Rule We annotate ethnonyms (names referring to ethnic groups) and demonyms (names referring to the inhabitants of a place) as GRP.

In the following example, **Chyurda** is the name of people living in a specific kingdom, in the *The Farseer Trilogy*:

That was the first year the GRP **Chyurda** tried to close the pass.

The novel *Brave New World* by Aldous Huxley provides another good example, with the caste system it describes:

We decant our babies as socialized human beings, as GRP **Alphas** or GRP **Epsilons**.

Adjectives We also annotate demonyms and ethnonyms when used as adjectives. For instance, **Dothraki** are an ethnic group of nomadic warrior in *A Song of Ice and Fire*:

- “If the CHR **beggar king** crosses with a GRP **Dothraki** horde at his back, the traitors will join him.”
- A GRP **Dothraki** wedding without at least three deaths is deemed a dull affair.

In the first sentence, the word **Dothraki** explicitly refers to people. It is not the case in the second one, where it rather refers to the **Dothraki culture**. To keep our annotation rules simple, we annotate it nevertheless.

Here is a (extreme) limit case from *Brave New World*:

In a little grassy bay between tall clumps of GRP **Mediterranean** heather [...]

As before, for the sake of simplicity, we consider that **Mediterranean** is an adjective derived from a demonym.

In the following example from *The Black Company*, **Arctic** is a bit tricky:

NOT **Arctic** imps giggled and blew their frigid breath through chinks in the walls of my quarters.

Here, **Arctic** means *northern*: it is not a demonym, as there is no Arctic continent or people in this fantasy world.

Languages The same word is often used to refer to a social group and to the language of its people. However, note that we annotate languages as cultural objects, see Section 7.2.

6.3 Definite Descriptions

General Rule A number of groups of characters are described using definite descriptions. We annotate them as **GRP** provided they exhibit the usual properties of capitalization, frequency and unicity. This type of groups is sometimes difficult to distinguish from organizations: see Section 8.6 for more detail on this topic.

Example In the following example from Suzanne Collins’s *The Hunger Games* series, the expression **Career Tributes** denotes a set of characters that are grouped because of one of their attribute, without any institutional existence or structure. Therefore, we annotate it as **GRP**:

In **LOC** district 12, we call them the **GRP** Career Tributes [...]

Counterexample In *Brave New World*, Aldous Huxley likes to refer to groups of people through their job or role in the society, and capitalizes the expression:

Bent over their instruments, three hundred **NOT** Fertilizers were plunged [...]

We do not annotate such expressions, unless they are frequently mentioned as a group. Here, **Fertilizers** appear only once in the whole novel.

6.4 Enumerations

General Rule We annotate enumerations of multiple entities as groups if these entities are not self-sufficient.

Example In the following example, **Mr** is not self-sufficient, so we annotate the whole expression as **GRP**:

GRP Mr and Mrs Bennet plan to go to **LOC** London soon.

Example On the contrary, in the following example from *The Three Musketeers*, each individual character is clearly identified, and therefore annotated separately:

[...] highly applauded, except by **CHR** MM. Grimaud, **CHR** Bazin, **CHR** Mousqueton, and **CHR** Planchet.

7 Miscellaneous Entities (MSC)

This category gathers various types of entities likely to be of interest.

7.1 Temporal Entities

General Rule We annotate named temporal entities, also called praxonyms, as MSC. This encompasses events such as revolutions, crises, festivals, etc., and as well as historical periods.

Holidays & Festivals We annotate holidays and festivals, provided they have a name. For instance, in *Moby Dick*:

Now, it being MSC Christmas when the ship shot from out her harbor [...]

Events Many historical events are rather punctual, compared to historical periods that are discussed below. See this excerpt of *A Song of Ice and Fire*:

The MSC Red Wedding was my father's work, and CHR Ryman's and CHR Lord Bolton's.

The expression Red Wedding refers to an event that lasted a few hours, and constitutes a turning point in the story.

Alternatively, the event can be hypothetical, as in this example from *The Three Musketeers*:

at the day of the MSC Last Judgment CHR God will separate blind executioners from iniquitous judges?

Periods This type of annotation also concerns historical periods. For instance, from the *A Song of Ice and Fire*:

- “There was one knight,” said CHR Meera, “in the MSC Year of the False Spring [...]
- The signing of the MSC Pact ended the MSC Dawn Age, and began the MSC Age of Heroes.

Both Dawn Age and Age of Heroes refer to periods in the ancient history of this lore. Similarly, in *The Black Company*:

The ORG Company was in service to the CHR Archon of Bone, during the MSC Revolt of the Chiliarchs.

Sometimes, the distinction between punctual event and period is not obvious. For instance, in *A Song of Ice and Fire*:

“They date from before MSC Aegon's Conquest,” CHR Cersei explained to her.

The conquest of Westeros by Aegon took some time, so technically it is a period. But here, the expression Aegon's Conquest actually refers to his Aegon's coronation, which is used as a reference date by the historians in this lore (akin to BC/AD in the real world).

Dates We do not annotate dates in general:

- D'you know what that little girl of mine did last NOT Saturday, when her troop was on a hike [...]
- Last NOT Monday (NOT July 31st) we were nearly surrounded by ice [...]
- [...] the question didn't arise; in this year of stability, NOT A.F. 632, it didn't occur to you

The first and second examples are from *1984* and Mary Shelley's *Frankenstein*, whose worlds and calendars are similar to ours. The third one is from *Brave New World*, in which dates are expressed relative to the production of the first Ford T automobile, hence the A.F. (*Anno Ford*).

One justification for not annotating dates is that they are usually considered as a separate specific entity for NER. Moreover, there are tools specifically designed to handle them, such as HeidelTime [16].

7.2 Cultural Assets

General Rule We annotate as MSC important artistic and intellectual works, as well as cultural objects.

Intellectual and Artistic Works Intellectual works include books, songs, movies, paintings, etc. We annotate their names or titles when they appear explicitly in the text, for instance:

- We will take this book, the MSC Book of Mazarbul, and look at it more closely later.
- [...] like the men singing the MSC Corn Song, beautiful, beautiful, so that you cried [...]
- [...] to which CHR Helrnholtz had recently been elected under MSC Rule Two.

The first example come from *The Lord of the Rings*, and both others from *Brave New World*.

Cultural Objects Cultural objects encompass dishes and wines, as in this example from *A Song of Ice and Fire*:

There is a flagon of good MSC Arbor gold on the sideboard, CHR Sansa.

We can also mention spirits, for instance, in *Moby Dick*:

[...] and with a benevolent, consolatory glance hands him—what? Some hot MSC Cognac?

There are also games, such as this fictional card game from *The Black Company*:

We were playing head-to-head MSC Tonk, a dull time-killer of a game.

And sports, like in this excerpt of *Brave New World*:

The crowds that daily left LOC London, left it only to play MSC Electromagnetic Golf or MSC Tennis.

Cultural objects also include a wide array of similar concepts, e.g. a scientific technique in *Brave New World*:

But MSC Podsnap's Technique had immensely accelerated the process of ripening.

Or a commercial brand in *Harry Potter*:

He had patched up his wand with some borrowed MSC Spellotape [...]

Or the motto of the noble houses in *A Song of Ice and Fire*:

All but the GRP Starks. 'Winter is coming,' said the MSC Stark words.

Languages We include languages in this category. This is a bit far-fetched, but it allows distinguishing languages from demonyms, which often take the exact same form in English.

Here is an example from *The Three Musketeers*:

CHR D'Artagnan did not know LOC London; he did not know a word of MSC English [...]

We annotate fictional languages too, such as in this example from *A Song of Ice and Fire*:

They had no common language. MSC Dothraki was incomprehensible to her [...]

7.3 Named Artefacts

General Rule We annotate named items, also called ergonyms, as MSC, provided they are *not* sentient.

Example For instance, King Arthur's sword Excalibur is magical, but does not act independently nor communicate:

CHR Arthur drew his sword MSC Excalibur that he had gained by CHR Merlin from CHR Vivian .

Many vehicles are also named, and are consequently annotated as MSC, e.g.:

- CHR Batman raced through LOC Gotham City streets in the MSC Batmobile, ready for action.
- MSC Black Betha rode the flood tide, her sail cracking and snapping at each shift of wind.

The first example is invented, the second comes from *A Song of Ice and Fire*.

Counterexample However, as explained in Section 3.7, named items that are sentient are annotated as characters. This is the case, for instance, of Elric’s sword Stormbringer, or of the Sorting Hat in *Harry Potter*:

The MSC Sorting Hat chose you for ORG Gryffindor, didn’t it? And where’s CHR Malfoy?

7.4 Other Entities

A number of other kinds of entities are referred to by proper names in novels, in which case we annotate them as miscellaneous entities.

Meteorological Events These events, also called *phenonyms*, include tempests, cyclones, etc. Here is an example from *Moby Dick* that mentions the name of a wind:

[...] where that tempestuous wind MSC Euroclydon kept up a worse howling than ever it did [...]

Awards & Decorations We annotate the names of awards and decoration, as in this example from George Orwell’s *1984*:

CHR Comrade Withers [...] had been [...] awarded a decoration, the MSC Order of Conspicuous Merit

8 Type Confusion

In certain cases, it is not clear what the type of an entity is. In this section, we focus on this issue, and provide some example that aim at helping to make such distinction.

8.1 Characters vs. Locations

Characters as Locations Some organizations locations are named after a person's name. For instance, a commercial building such as **Moroggo's**, an inn in *A Song of Ice and Fire*, is named after its owner:

CHR Sam began his search at [...] **LOC** Moroggo's, places where **CHR** Dareon had played before.

Here, including the genitive 's in the annotation is consistent with our decision to annotate the outmost entity in nested entities (cf. Section 2.1). Here is another example, also from *A Song of Ice and Fire*:

[...] some stranger from the **LOC** Vale of Arryn whose name she had forgotten [...]

In this case, **Arryn** is the name of the person that claimed this territory after a battle. When the name contains a qualifier, such as **Vale** here, the distinction between character and location is much clearer.

The name of the person may be used without any modification, e.g.

LOC London Zoo is approximately a 30 minute walk from **LOC** Saint Pancras.

Where **Saint Pancras** is a train station, and not the person **Pancras of Rome**.

Locations as Characters It happens that a location strongly associated with a character is used in place of their name. We annotate the location name as a **CHR**. See, for instance, this excerpt from *Moby Dick*:

It drags hard; I guess he's holding on. Jerk him, **CHR** Tahiti! Jerk him off; we haul in no cowards here.

One of the unnamed seamen is from **Tahiti**, and **Captain Ahab** uses the name of this location instead of the proper character's name.

Double Meaning Sometimes, the confusion between character and location is more conceptual, as a large entity can be both a character and a place. For instance, in *The Adventures of Pinocchio*, **Pinocchio** and **Geppetto** are swallowed by a giant **shark**, which appears first as a character in the story, before becoming a place where **Pinocchio** can have a walk:

— "Is this **CHR** Shark that has swallowed us very long?" asked the **CHR** Marionette.

— "His body, not counting the tail, is almost a mile long." [...]

CHR Pinocchio [...] began to walk as well as he could inside the **LOC** Shark, toward the faint light which glowed in the distance.

Another similar example is **Erythro**, the sentient planet at the center of Isaac Asimov's novel *Nemesis*:

- [...] her thoughts were often on **LOC** Erythro, the planet they had been orbiting almost all her life.
- **CHR** Erythro had knowledge of only one kind of mind—its own.

8.2 Characters vs. Organizations

Characters as Organizations Metonymy is quite frequent when referring to organizations, which can lead to a certain confusion. In the below examples from *The Black Company*, it is the case between the person nicknamed **White Rose** and her armed group:

- I told him about the **CHR** White Rose, the lady general who had brought the **ORG** Domination down
- The **CHR** Lady is no exception. The **ORG** Sons of the White Rose are everywhere.
- **ORG** White Rose prophets and **ORG** Rebel mainforcers were the least of our troubles.

The full name of this organization is **Sons of the White Rose**, but the expression **White Rose** is more frequently used as a shortened form.

8.3 Characters vs. Miscellaneous

Characters as Items Metonymy between persons and the various objects covered by the MSC type are quite frequent. Consider these examples from *Moby Dick*:

- **CHR** Charity, his sister, had placed a small choice copy of **MSC** Watts in each seaman's berth.
- [...] seemed made of solid bronze, [...] like **CHR** Cellini's cast **MSC** Perseus.

Isaac Watts is an author of religious hymns, and in the first sentence, **Watts** refers to one of his books. In the second sentence, **Cellini** is the sculptor Benvenuto Cellini, and **Perseus** does not refer directly to the Greek hero, but rather to one of Cellini's work representing the eponymous mythological figure, and titled *Perseus with the Head of Medusa*.

8.4 Locations vs. Organizations

Metonymy is frequently used between locations and organizations. This can make it difficult to distinguish between **LOC** and **ORG** entities, as one name can be used as a location as well as an organization. In both case, it is important to take the context into account in order to decide of the entity type.

Locations as Organizations On the one hand, some location name can be used to refer to an organization. In *A Song of Ice and Fire*, the **Citadel** is the name of a neighborhood hosting the headquarters for the **maesters**, a group of scholars. In the following example, this name is used to refer to the organization instead of the place:

Last year when he took ill, the **ORG** Citadel had sent **CHR** Pylos out from **LOC** Oldtown [...]

Another example using a country name (**England**) in *Harry Potter*, when the author actually means an organization (the national Quidditch team):

[...] he could have played for **ORG** England if he hadn't gone off chasing dragons.

Organizations as Locations On the other hand, sometimes the name of an organization is used to refer to its location. For instance, in *Harry Potter*, **Hogwarts** is the name of an institution, but it is often used to denote the school grounds:

[...] he had ten minutes left to get on the train to **LOC** Hogwarts and he had no idea how to do it

Another example, this time from George Orwell's *1984*:

A kilometre away the **LOC** Ministry of Truth, his place of work, towered vast and white [...]

The expression **Ministry of Truth** does not refer to the actual organization here, but rather to the building hosting it. Similarly, from the same novel:

CHR Winston has never been inside the **LOC** Ministry of Love, nor within half a kilometer of it.

Undetermined In cases where the context is unclear about the type of the entity (it could be **LOC** as well as **ORG**), we annotate it as **ORG**. For instance, in *A Song of Ice and Fire*, the **Citadel** sometimes means a place, and sometimes the scholar organization sitting at this place:

That is so, my lady. The white ravens fly only from the **ORG** Citadel.

In the above example, it is not clear whether the **Citadel** is the place from which the ravens fly, or the organization that send them, or even both of them. In *The Black Company*, we have a similar case:

[...] from one of several nearby dives frequented by the **ORG** Bastion garrison.

The **Bastion** could as well be the organization localized in this building and commanding the garrison, as the building hosting this garrison.

Also in *The Black Company*, the **Jewel Cities** are a group of geographically and culturally close cities. Since the name does not refer to a group of people, it cannot be annotated as **GRP**. Depending on the context, the expression can be a location or an organization, but it is not always obvious, e.g.:

CHR Soulcatcher **ORG** commanded the Guard and allies from the **ORG** Jewel Cities.

Here, we apply the principle mentioned earlier and opt for **ORG**.

8.5 Locations vs. Miscellaneous

Locations as Events By metonymy, the name of a location can be used to refer to an important event that took place in this location. In the following example from *The Black Company*, **Beryl** is a city where a battle took place:

But **CHR** Soulcatcher is in high favor since **MSC** Beryl, and the **CHR** Limper isn't because of his failures.

8.6 Organizations vs. Groups

Principle The distinction between organizations and groups is not always obvious. Size can be a clue, as organizations tend to be larger. Possessing a proper noun would also shift the balance for being an organization. But the main difference is the institutional nature of the entity, i.e. how official it is.

Examples Here are some limit cases. In *Harry Potter*, the **Marauders** are a group of four students that gathered mainly to make mischief:

"Maybe the **ORG** Marauders never knew the room was there," said **CHR** Ron.

In this case, the fact that they have their own name would suggest a certain form of recognition. On the one hand, there is no proper structure within the group, but on the other hand, the secret nature of the **Marauders** hints at a certain level of organization. For these reasons, and although this decision is debatable, we annotate it as an **ORG**.

Also in *Harry Potter*, we annotate **Dudley Dursley's gang** as a group and not an organization, for the exact opposite reasons:

CHR Harry was glad school was over, but there was no escaping **GRP** Dudley's gang [...]

It is just a group of children informally gathered by **Harry's** cousin **Dudley**. Its name is fixed and appears relatively frequently, which is why we consider it as a proper group, and do not annotate only **Dudley** as a character.

In the *Lord of the Rings*, despite the title of the first volume in the trilogy, the **Fellowship of the Ring** is actually called **Company of the Ring** in the text:

The **ORG** Company of the Ring stood silent beside the tomb of **CHR** Balin.

Whatever its name, this fellowship is formed at a council, and it is constituted of nine members selected to represent specific races from the lore. This is all very organized, which is why we annotate it as an **ORG**.

9 Concluding Remarks

It is worth stressing that having to deal with novels has a significant influence on the annotation process. It is necessary to have a global vision of the whole document to perform this task correctly. This helps to identify elements that are specific to the considered story and to its environment: Who are the characters? What can be considered as frequent in this book? Which honorifics are used in the word of the novel? What are the nicknames? The places? Which conventions does the author use, for instance concerning capitalization?

Concretely, this means that it is more efficient and reliable for one person to annotate the whole book than a few chapters. Also, it is preferable to have a single person annotating the whole book than dividing up the chapters among several persons. Moreover, the annotator should not hesitate to use additional resources. This includes tools such as the *Calibre* ebook viewer², that allow them to assess how frequent some expression is, in order to determine whether it should be annotated as an entity or not. Such tools can also help to determine whether some expression uniquely identifies some entity.

Concerning novels that take place in imaginary worlds, especially Fantasy and Science-Fiction realms, it is particularly important to leverage the available online wikis. These are generally elaborated by fans and are very complete, exploring the lore in detail, providing a Web page for each entity, even minor. A number of concepts in such novels are completely foreign to an unfamiliar reader: a made-up name could as well be a character, a location, an organization, or even a honorific title. Such encyclopedic resources help a lot in alleviating ambiguities, and more generally, making certain annotation decisions.

Finally, our last piece of advice is to keep notes of the decisions one makes while annotating a book. For instance, was this group of people annotated as a **GRP** or an **ORG**? Was this expression annotated at all? Indeed, a significant amount of text can separate two mentions of the same entity, and one may forget how they previously handled it. Keeping notes helps keeping the annotation process consistent.

²<https://calibre-ebook.com/>

A Version History

We use three-part version numbers of the form major–minor–patch for both these guidelines and the *Novelties* corpus. Concretely, *major* changes correspond to very significant modifications of the rules, such as the introduction of a new type of entity. *Minor* changes are modifications of the existing rules (through their edition, addition, deletion). For instance, we could decide to include the determiners in the annotations. Finally, the *patch* level concerns the correction of errors or the clarification of existing rules, for instance by adding new examples to the guidelines, that illustrate cases never met before.

Version	Date	Changes	Corpus
0.1.0	25/01/2023	Beta version based on the work by Dekker et al. [6], with some corrections and additional entity types LOC and ORG.	0.1.0, 0.2.0
1.0.0	07/03/2024	Significant revision, based on our feedback from the first annotation pass: more precise definition of a named entity; entity type PER becomes CHR; new entity types GRP and MSC added.	1.0.0
1.0.1	11/06/2024	Minor revision: languages now considered as MSC entities, instead of being ignored.	1.0.0

Version 0.1.0 This is the beta version of our guidelines. We kick-started our own corpus by leveraging the OWTO corpus (*Out With The Old*) proposed by Dekker et al. [6]. Consequently, our very first guidelines are exactly the same as theirs. However, the original annotation guidelines of Dekker et al. [6] are extremely minimal, and the OWTO corpus exhibits encoding, tokenization, quoting and annotation problems. We leveraged the experience gained from correcting these errors to modify the guidelines, making them slightly more precise, and adding a few examples. We also extended the scope of entity types, adding *locations* (LOC) and *organizations* (ORG). We followed this version of the guidelines to produce version 0.1.0 of *Novelties*, which is used in [2], and version 0.2.0, which was not used in any published work.

Version 1.0.0 This is a major revision of our guidelines. It is based on our experience in annotating new chapters and even full novels, in an attempt to expand *Novelties*. We released an extensive annotation guide, including a general explanation of our concept of named entity, a list of different annotation cases for each entity type along with positive and negative examples, and a specific part dedicated to possible confusions between types. In terms of major changes in the guidelines themselves, we specifically extended the PER class to include all characters (such as animated weapons, sentient magical creatures, robots...), and thus renamed it CHR for *character*. We also introduced two new classes of entities: *groups* (GRP) and *miscellaneous* (MSC). MSC entities are common in other corpora and allow us to annotate additional entities of interest, while GRP entities allow us to distinguish between single characters or groups of them.

Version 1.0.1 In version 1.0.0 of the guidelines, we do not annotate languages (French, Dothraki, etc.) at all. This can be a bit confusing for the annotator, considering that the same word is often used in English to refer to a people and to its language. To make things clearer, in this minor revision we change this and annotate languages as cultural assets, using tag MSC.

B Todo List

Here is a list of examples missing from this document:

- Is there an example of explicit date (numbers) that is significant for the novel? (Section 7.1)
- Are there some situations where disjoint names are used for locations? (Section ??)
- Is the use of noun modifiers the only situation where only a part of a location name is mentioned? (Section 4.2)
- Is there an example of disjointed organization names where the mentions are self-sufficient? (Section 5.3)

Some questions still open:

- How to annotate adjectives derived from the proper noun of persons? e.g. Marxist, Circean, etc.

References

- [1] M. Alrahabi, C. Brando, F. Frontini, A. Provenier, R. Jalabert, M. Bordry, C. Koskas, and J. Gawley. Guide d’annotation manuelle d’entités nommées dans des corpus littéraires. Technical report, Campagne d’annotation OBVIL 2019–2021, 2021. URL <https://hal.science/hal-03156278>.
- [2] A. Amalvy, V. Labatut, and R. Dufour. The role of global and local context in named entity recognition. In *61st Annual Meeting of the Association for Computational Linguistics*, pages 714–722, 2023. doi:10.18653/v1/2023.acl-short.62.
- [3] A. Amalvy, V. Labatut, and R. Dufour. Renard: A modular pipeline for extracting character networks from narrative texts. *Journal of Open Source Software*, 9(98):6574, 2024. doi:10.21105/joss.06574.
- [4] D. Bamman, S. Popat, and S. Shen. An annotated dataset of literary entities. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2138–2144, 2019. doi:10.18653/v1/N19-1220.
- [5] N. Chinchor and P. Robinson. Appendix E: MUC-7 named entity task definition (version 3.5). In *7th Message Understanding Conference*, 1998. URL <https://aclanthology.org/M98-1028>.
- [6] N. Dekker, T. Kuhn, and M. van Erp. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189, 2019. doi:10.7717/peerj-cs.189.
- [7] M. Ehrmann. *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Phd thesis, Université Paris Diderot, 2008. URL <https://theses.hal.science/te1-01639190>.
- [8] M. Falk. Making connections: Network analysis, the Bildungsroman and the world of The Absentee. *Journal of Language, Literature and Culture*, 63(2-3):107–122, 2016. doi:10.1080/20512856.2016.1244909.
- [9] J. R. Finkel and C. D. Manning. Nested named entity recognition. In *Conference on Empirical Methods in Natural Language Processing*, pages 141–150, 2009. URL <https://aclanthology.org/D09-1015/>.
- [10] R. V. Ivanova, S. Kirrane, and M. van Erp. Comparing annotated datasets for named entity recognition in english literature. In *13th Language Resources and Evaluation Conference*, pages 3788–3797, 2022. URL <https://aclanthology.org/2022.lrec-1.404/>.
- [11] V. Labatut and X. Bost. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52(5):89, 2019. doi:10.1145/3344548.
- [12] Linguistic Data Consortium. ACE (automatic content extraction) english annotation guidelines for entities. Technical report, Linguistic Data Consortium, 2008. URL <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf>.
- [13] David McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In *Acquisition of Lexical Knowledge from Text*, 1993. URL <https://aclanthology.org/W93-0104>.
- [14] S. Rosset, C. Grouin, and P. Zweigenbaum. Entites nommées structurées : guide d’annotation Quaero. Technical report, Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur, 2011. URL <https://perso.limsi.fr/rosset/quaero-guide-annotation-2011.pdf>.
- [15] A. Soudani, Y. Meherzi, A. Bouhafs, F. Frontini, C. Brando, Y. Dupont, and F. Mélanie-Becquet. Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle. In *SAGEO Atelier Humanités Numériques Spatialisées*, 2018. URL <https://github.com/DHNamedEntities/19thCenturyFrenchNovels/blob/master/paper-fr.pdf>.
- [16] J. Strötgen and M. Gertz. A baseline temporal tagger for all languages. In *Conference on Empirical Methods in Natural Language Processing*, pages 541–547, 2015. doi:10.18653/v1/d15-1063.
- [17] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *7th Conference on Natural Language Learning*, pages 142–147, 2003. doi:10.3115/1119176.1119195.
- [18] H. Vala, S. Dimitrov, D. Jurgens, A. Piper, and D. Ruths. Annotating characters in literary corpora: A scheme, the charles tool, and an annotated novel. In *10th Language Resources and Evaluation Conference*, pages 184–189, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/1130_Paper.pdf.
- [19] T. Venturini, L. Bounegru, M. Jacomy, and J. Gray. How to tell stories with networks: Exploring the narrative affordances of graphs with the Iliad. In *Datafied Society: Studying Culture Through Data*, chapter 11, pages 155–170. Amsterdam University Press, 2016. doi:10.1515/9789048531011-014.
- [20] R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, and N. Xue. OntoNotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation*. Springer, 2011. URL <https://www.cs.cmu.edu/~hovy/papers/09OntoNotes-GALEbook.pdf>.