# Comparing Annotated Datasets for Named Entity Recognition in English Literature

**Rositsa V. Ivanova,**[1] **Sabrina Kirrane,**[1] **and Marieke van Erp**[2]
[1]Vienna University of Economics and Business
Welthandelsplatz 1, 1020, Vienna, Austria
[2]KNAW Humanities Cluster
Oudezijds Achterburgwal 185, 1012 DK, Amsterdam, The Netherlands
{rivanova,skirrane}@wu.ac.at, marieke.van.erp@dh.huc.knaw.nl

## Abstract

The growing interest in named entity recognition (NER) in various domains has led to the creation of different benchmark datasets, often with slightly different annotation guidelines. To better understand the different NER benchmark datasets for the domain of English literature and their impact on the evaluation of NER tools, we analyse two existing annotated datasets and create two additional gold standard datasets. Following on from this, we evaluate the performance of two NER tools, one domain-specific and one general-purpose NER tool, using the four gold standards, and analyse the sources for the differences in the measured performance. Our results show that the performance of the two tools varies significantly depending on the gold standard used for the individual evaluations.

**Keywords:** named entity recognition, corpus annotation, annotated datasets, annotation guidelines, literature

## 1. Introduction

Over the years, researchers and engineers have investigated various approaches that could potentially improve the recognition and linking of named entities. These different approaches include using deep learning to extract entities (Li et al., 2020), providing tools with more annotated training data, for instance by making use of technologies such as linked databases to link entities despite diverse name variations (Zhu and Iglesias, 2018), enriching gazetteers (i.e. lists of aliases of entities) and using dictionaries (i.e. entity mappings) for better linking between entities (Hulpus et al., 2013). Most of these approaches have indeed led to significant improvements in the performance of named entity recognition (NER) tools. However, as we showcase in the following examples from the English literature domain, there is still room for improvement.

Generally speaking, the majority of NER tools struggle to perform well when the entities in the text contain specific characteristics. When it comes to English novels Dekker et al. (2019) observe the poor performance of off-the-shelf tools when names contain characters, which are used in an unusual manner for that particular language (e.g. d'Artagnan). In historical letters, Mac Kim and Cassidy (2015) describe the lack of cues to differentiate between names and ordinary words as one of the biggest challenges when it comes to automated analysis. Furthermore, Woldenga-Racine (2019) conclude that the most frequent cause for incorrect NER in old English novels is the use of capitalized words for non-named entities. Although, many tools rely on gazetteers and rules for NER, the gazetteers and machine learning methods used to enrich them are mostly based on modern English, making the analysis of old documents, including books, difficult and less accurate.

In this paper, we take a closer look at existing annotated NER datasets in the domain of English literature and compare the performance of NER tools using such annotated datasets as a means to detect the differences between the datasets. The remainder of this paper is organised as follows. Section 2 gives an overview of existing annotated datasets in the English literature domain and introduces two datasets in detail. Section 3 depicts the creation of a new annotated dataset for the purpose of this paper. Next, Section 4 introduces two tools used for our evaluation. Section 5 and Section 6 present the results of the evaluation and their analysis respectively. Following, in Section 7 we discuss four main topics related to the creation of annotated datasets and their use for NER tools. Lastly, Section 8 concludes the paper and discusses future work.

## 2. Existing Annotated Datasets

Focusing on dialogue-based extraction of characters from novels needed for the creation of a social network, Elson et al. (2010) introduce the Columbia Quoted Speech Attribution corpus of 60 annotated novels. Three annotators analysed conversational interactions and detected the corresponding characters. This work targets the domain of literary texts, however it addresses an analysis of the dialogue flow. Recently, Dekker et al. (2019) and Vala et al. (2015) analysed the challenges of detecting characters and their aliases (e.g. nicknames) for NER tools in the literary domain and created their own datasets for the purpose of their analysis. Simultaneously, a dataset for English novels called LitBank (Bamman et al., 2019) was created as a first step towards addressing the lack of labeled literary texts. Although there exist large annotated datasets in the literary domain in different languages such as German (Krug et al., 2017), the number of datasets and their scope in English is limited.

For the purpose of this paper we select two corpora. First, we use the currently most extensive dataset of annotated English novels, called Litbank[1], which was created with the intention of addressing the gap in relation to insufficient datasets specifically for the domain of English literary texts (Bamman et al., 2020). Second, we select the annotated dataset[2] proposed by Dekker et al. (2019), which was created to support the extraction of character networks. The authors evaluate the performance of various tools and analyse their shortcomings.

## 2.1. The LitBank Annotated Dataset

LitBank (Bamman et al., 2020; Bamman et al., 2019; Sims et al., 2019) is a dataset consisting of annotated sections of 100 novels. It follows the ACE 2005 guidelines and therefore contains six categories - people, facilities, locations, geo-political entities, organizations, and vehicles.[3] According to the description of the annotation process defined in Bamman et al. (2020) the annotation was done by three people. However, all but 10 novels were assigned to a single annotator each. The remaining 10 texts were used as a means to calculate the consistency between the annotations of the three people. It's worth noting that another publication about LitBank (Sims et al., 2019) states that the raw texts were annotated by only one person. In addition, five novels that were annotated by a second person served as control units to calculate an inter-annotator $F_1$ score. Both publications state that the scope of the collection is 210,532 tokens within 100 novels. The possibility of a single person vs. three people annotating all texts and producing the same number of tokens is very low. Furthermore, we inspected the commit history of the GitHub repository[4] cited by both papers and did not find an update in the existing annotations in the timespan between the two publications. Lastly, we compared a previously created GitHub repository for LitBank[5] to the latest above-mentioned repository and did not detect any changes in the files. Therefore, it remains unclear how many novels in the LitBank collection were annotated by more than one person.

Lastly, LitBank uses multiple layers of annotations. This means that tokens can be a part of multiple entities simultaneously. The authors argue that a flat structure does not cover the annotated needs of literary texts, because for example *"The cook's sister ate lunch"* contains two PER entities *"[The cook]"* and *"[The cook's sister]"* (Bamman et al., 2019). In addition to the categorisation of the words as belonging to an entity type (e.g. PER), the first word of an entity is prefixed by a "B" (i.e. beginning) and the following words are prefixed by an "I" (i.e. inside).

This feature of multi-layered annotations does not have a corresponding flat annotation file representation such as the ones produced by most NER tools (e.g. BookNLP). Therefore, to be able to evaluate the performance of such tools and to compare this dataset to others, we produce a flat version of the file. The flattening of the layers inevitably leads to a loss of information. To address the multiple layers of people entities for the same token we chose to use the people entity with the longest scope. We acknowledge that while a specific approach needs to be chosen for the flattening of a multi-layered gold standard, this inevitably leads to a certain bias in the flattened gold standard. In this regard, the user extracting such flat files needs to make a decision, which based on the chosen approach may result in different gold standards. Unfortunately, to a certain degree this defeats the purpose of an "objective truth" targeted by gold standards. One possible way to reduce this bias would be to include each named entity type at most once per entity. This means that for example "Sofia" in "Sofia's friends" could have the layers B-LOC (i.e. location) and B-PER, but not B-LOC, B-PER and I-PER simultaneously. However, this approach would reduce the granularity of the annotated dataset. Alternatively, an additional version, which is flattened could be provided for tools with one-layered output files. In this case, it is essential not to mix the format (i.e. number of layers used) of the gold standards, if multiple tools output different number of entity layers.

## 2.2. The OWTO Annotated Dataset

The second dataset consists of 20 modern and 20 old novels only including the entity type person (Dekker et al., 2019). This dataset was used to study whether NER tools perform better with modern or old novels. The novels were annotated by two people, both of whom were assigned 20 novel fragments with an average length of 300 sentences. We further refer to this gold standard as the OWTO (Out with the old) annotated dataset.

## 2.3. Comparison of General Characteristics

The following subsection compares the general characteristics of the LitBank and OWTO datasets in terms of dataset size, source of raw texts, annotating approach, purpose, followed guideline, whether or not an initial automated annotation was done, the covered entity types, and the annotation layers. An overview of those can be found in Table 1. The size of the datasets varies both in terms of the number of books and the length

[1] `https://github.com/dbamman/litbank`
Commit: a371cd678701fc98371355b328a1a6c4b58508a3

[2] `https://github.com/Niels-Dekker/Out-with-the-Old-and-in-with-the-Novel`
Commit: ad31ce1fa515dceabb8febbaa7aa235f3de47ebd

[3] The Automatic content extraction (ACE) programme (1999-2008) was aimed at developing advanced information extraction technologies.

[4] `https://github.com/dbamman/litbank`

[5] `https://github.com/dbamman/NAACL2019-literary-entities`

of the annotated text per book. While LitBank covers more novels, OWTO provides longer sections of annotated text.

Both datasets use Project Gutenberg[6] as a source for the raw texts. In addition, Dekker et al. (2019) purchased certain books online, as newer novels were not available in the Project Gutenberg's collection. Both datasets contain annotations created by only one person. In the case of LitBank, according to (Bamman et al., 2019; Sims et al., 2019), five novels are annotated by two people and used as control units to calculate an inter-annotator $F_1$ score. The OWTO collection was annotated by two people, yet those were assigned different novels.

LitBank follows *"the guidelines set forth by the ACE 2005 entity tagging task"* (Linguistic Data Consortium, 2006) and OWTO used BookNLP (Bamman et al., 2014) to create an initial annotation as a means to speed up the actual annotation process. Lastly, LitBank offers annotation for the entity types *people, facilities, geo-political entities, locations, vehicles,* and *organizations*, while OWTO focuses on the entity type *person*. To allow for the same token in a sentence to be recognised as multiple entity types, LitBank uses multiple layers.

## 3. Dataset creation

To better understand the impact that the selection of annotation guidelines has on the tool performance, we create a third dataset. For this purpose we annotated the novel sections which are annotated in both gold standards. The overarching goal is not to create a new gold standard but rather to study the impact of annotation guidelines on the annotation process, the challenges faced by the annotators, and how different annotations effect tool performance.

The annotation of the overlapping sections of the 12 novels was done using Doccano v1.2.2[7]. First, we gave an introduction to the tool to two annotators and gave them the chance to freely test a practice project, for which we used a text section from one of the 100 raw texts provided by LitBank. After confirming that the annotators were familiar with the process and that their questions were covered, we proceeded to the actual annotation process. We ensured that the two annotators did not communicate during the individual annotation process. After they annotated all texts, we detected the differences and let the annotators agree on a shared final version.

### 3.1. Annotation Guidelines

The annotators were provided with annotation guidelines to follow throughout the process. They were available on every page containing text that needed to be annotated. Overall, we focused on the entity type person. In our guidelines we differentiate between the labels PERSON and PERX. The PERSON label follows annotation guidelines extracted from the MUC-7[8] (Chinchor et al., 1999), which the CoNLL-2003 task is based on. We use the guidelines from the CoNLL-2003 task due to the fact that its datasets are amongst the most commonly used ones for the evaluation of tools.

We avoided changing the formulations of the rules and the examples as much as possible, to reduce unintended bias. This means that the majority of the guidelines are literal extractions of the sections relevant to the person entity type from the original guidelines, which included all entity types. The original guidelines consist of individual rules for the groups of entity types, presenting the taggable and non-taggable instances of entities. In contrast to the original guidelines, we clearly differentiate between the tokens that need to be marked as PERSON (i.e. include) and those that need to be ignored, by separating them in two categories. This helped the annotators to clearly identify, whether the entity is to be tagged or ignored.

The PERX label extends the PERSON label by accepting more tokens as the person entity type. The PERX label is based on the differences between the CoNLL-2003 guidelines and the annotation guidelines used by Bamman et al. for the creation of the LitBank corpora (Bamman et al., 2020; Bamman et al., 2019; Sims et al., 2019). We selected those guidelines for the extension of the PERSON label due to the fact that they were chosen with the purpose of creating an annotated dataset for the domain of English literature.

According to the authors of LitBank, their *"annotation style largely follows that of OntoNotes, in defining the boundaries for markable mentions that can be involved in coreference and in defining the criteria for establishing coreference between them"* (Bamman et al., 2020). With the purpose of coreference, OntoNotes aims to link all mentions of entities in the text to the correct entities. By containing those links, annotated datasets should provide examples that can be used to train computers to automatically extract information through the recognised entities (BBN Technologies, 2007). The main deviations of the LitBank annotation guidelines from those of OntoNotes described by Bamman et al. (2020) are the inclusion in LitBank of (i) *"noun phrases that are not involved in coreference"* (Bamman et al., 2020) (i.e. singletons) and (ii) quantified and negated noun phrases. OntoNotes generally does not treat negated noun phrases as taggable, however some exceptions do exist (e.g. "the students" in "none of the students") (BBN Technologies, 2007).

To evaluate the inter-annotator agreement we use Cohen's kappa (McHugh, 2012). Kappa values above 0.60 and below 0.80 are viewed as representing a *moderate* level of agreement. Those between 0.80 and 0.90

[6] https://www.gutenberg.org

[7] https://github.com/doccano/doccano

[8] https://web.archive.org/web/20060211040221/https://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4.pdf

| Dataset characteristics | LitBank | OWTO |
|---|---|---|
| Dataset size | 100 novels, ca. 2000 words each | 40 novels, 300 sentences each |
| Source | Project Gutenberg | Project Gutenberg<br>or purchased online |
| Annotators | 95 novels by one person<br>5 novels by two people[a] | two annotators<br>20 novels each |
| Inter-annotating $F_1$ score | 86.0 | NA |
| Purpose | coreference | coreference resolution,<br>creation of social networks |
| Guideline followed | ACE 2005, OntoNotes[b] | NA |
| Initial annotation | NA | using BookNLP |
| Entity types | people, facilities, geo-political entities, locations, vehicles, organizations | people |
| Annotation layers | multiple | one |

[a] Or 90 novels by one person each, remaining 10 by two people each (see 2.1).

[b] The annotation process describes certain deviations from the ACE 2005 annotation guidelines (e.g. excluding the entity type weapon). The authors also state that they followed the OntoNotes guidelines with certain deviations.

Table 1: Dataset Characteristics for LitBank and OWTO

are considered *strong*, while values above 0.90 are already viewed as *almost perfect*. In our case, the overall achieved Cohen's kappa scores are within these levels. An overview of the calculated scores[9] is displayed in Table 2. The inter-annotator agreement for the PERSON label (i.e. following the CoNLL-2003 guidelines) is in all but one case equal or above 0.90, which puts it in the level of an almost perfect agreement. The range in the Cohen's kappa for the PERX label is from 0.69 to 0.90, meaning that it is distributed in the moderate and strong levels. These results confirm the feedback received by the two annotators, who indicated that the more detailed annotator guidelines for the PERSON label made it easier for them to know which entities were to be tagged.

| Novel | PERSON label | PERX label |
|---|---|---|
| Alice in Wonderland | 0.88 | 0.90 |
| David Copperfield | 1.00 | **0.69** |
| Dracula | 0.90 | 0.74 |
| Emma | 0.97 | 0.77 |
| Frankenstein | 1.00 | 0.76 |
| Huckleberry Finn | 0.97 | 0.85 |
| Moby Dick | 0.95 | 0.78 |
| Oliver Twist | 1.00 | 0.76 |
| Pride and Prejudice | 0.98 | 0.90 |
| The Call of the Wild | 0.92 | 0.70 |
| Ulysses | 0.96 | 0.79 |
| Vanity Fair | 0.90 | 0.74 |

Table 2: Inter-annotator Agreement for New Datasets

## 4. Named Entity Recognition Tools

In terms of tool selection we chose one tool created specifically for the domain of English novels and another that does not target a specific domain but instead is deemed one of the best performing tools, in general. To narrow down the decision for the second tool, we considered the following specific criteria:

- The source code of the tool should be published and be free to use, for us to better understand and apply it.
- The tool should not have any specific requirements that we might not be able to meet (e.g. requirement of a GPU).
- It should be possible to use the tool off-the-shelf, without too many changes (e.g. rewriting parts of the code, having to set up parameters for machine learning algorithms).

### 4.1. BookNLP

Currently, BookNLP is the one of the few NER tools targeting English novels (Bamman et al., 2014). It is a tool created for the detection of characters in the literary domain – in particular novels from the 18th and 19th centuries. The main goal of the authors is to create a model, which *"account(s) for the influence of extralinguistic information (such as author)"* (Bamman et al., 2014). As such, it relies on the different styles of writing that authors have, as those affect the way characters are portrayed.

The BookNLP model is trained on data originating from online sources, such as Project Gutenberg[10] and HathiTrust[11], and further scanned and OCRed texts. The pipeline of the tool requires the following external software: Stanford POS tagger (Toutanova et al., 2003), linear-time MaltParser (Nivre et al., 2007) for

[9] For the calculation of the results we used the approach provided by `https://github.com/o-P-o/disagree`, which we updated to cover all issues reported in the repository by previous users. The final version of the used code can be find in the repository of this project.

[10] `https://www.gutenberg.org/`

[11] `https://www.hathitrust.org`

dependency parsing, and Stanford named entity recognizer (Finkel et al., 2005). For coreference resolution the authors differentiate between a character "mention" and an "entity". First, they define a set of existing characters and then map all indirect mentions (e.g. through proper nouns) to those initial characters. For this purpose a Bayesian approach is used.

### 4.2. Flair

The Flair framework (Akbik et al., 2019; Schweter and Akbik, 2020) aims to offer all word embedding types in an easy to use manner by abstracting from the details of their implementation. For the training of the model, Flair caters for accessing publicly available datasets for NLP. Based on the annotation guidelines, the task and the targeted language(s) one can select from nine corpora. The dataset is then downloaded and automatically split into training, testing and development sections. This enables the tool to be usable and comparable in various domains and based on different standards. For the purpose of NER in English the currently best performing pre-trained model is "ner-large"[12], which scores an $F_1$ score of 94.09 on the CoNLL-2003 benchmark dataset.

## 5. Results

We evaluate the performance of BookNLP and Flair using the CoNLL evaluation script and the four datasets without considering prefixes - LitBank, OWTO, the new dataset following the CoNLL-2003 guidelines for the entity type PEOPLE, and the new dataset extended by the annotation guidelines of LitBank targeting the domain of English novels. The source code, data and raw results of this experiment can be found at `https://github.com/therosko/annotated_datasets_en_comparisson`[13]. Table 3 and Table 4 depict the precision, recall and $F_1$ scores for BookNLP and Flair respectively. First, the scores per novel are presented. Then, at the bottom of both tables, we summarise the scores using the mean, standard deviation, and the median for each evaluation metric and annotated corpora.

Our main observation is that the results for both tools vary heavily based on the annotation dataset used as a gold standard. For BookNLP, we observe an $F_1$ score range from 0.00 for "David Copperfield" (using both new annotations) up to 95.65 for "Ulysses" (using the new annotation following the CoNLL-2003 guidelines). In the case of Flair this range is from 0.00 for "Dracula" (using OWTO) up to 96.97 for "Pride and Prejudice" (using the new annotation following the CoNLL-2003 guidelines). To better understand the discrepancy, we compare the annotation guidelines of the individual datasets.

## 6. Analysis

If we examine the $F_1$ score alone, both tools perform best when evaluated using the new annotation following the CoNLL-2003 guidelines (i.e. PERSON) and using OWTO, and they perform the worst using the new extended annotation (i.e. PERX). One of the main reasons for the poor performance of the tools using the new extended dataset is the fact that the extended dataset considers personal pronoun references (e.g. you, her) to be entities. In novels, such as Ulysses, Pride and Prejudice, and Frankenstein personal pronouns make up around 200 entities per novel, which is around 10% of all tokens in the annotated sections. When tools do not tag those pronouns as correct, their recall drops drastically, even if the precision of the tool is otherwise relatively high. In the case of BookNLP the median of the precision using the new extended dataset is 77.5%, while the recall is only 6.93%. In the case of Flair, the gap is even bigger with a precision of 90.54% and a recall of merely 9.65%. Both cases show a very low $F_1$ score on average.

Considering that the annotation guidelines of LitBank also include personal pronouns as entities, whenever they refer to (in our case) the entity type person, we would expect the results of the evaluation with LitBank to depict the same shortcomings of the tools as the results of the new extended dataset. However, despite the same formulation of the annotation rule, the LitBank gold standard contains only occurrences of personal pronouns in conjunction with other tokens (e.g. my mother) and not as single token entities (e.g. you).

In terms of precision, the main difference between the results using LitBank and the new extended dataset comes from the different approach to honorifics. LitBank includes honorifics as a part of the entity. Due to the fact that we follow the CoNLL-2003 honorifics are excluded from the list of taggable tokens in the new dataset. The effect of not including them in the new gold standards is clearest in the precision results of the novels Emma, David Copperfield, Pride and Prejudice, and Vanity Fair in the case of BookNLP, and of the novels Emma, David Copperfield, and Vanity Fair in the case of Flair. Interestingly, the effects differ as both tools handle honorifics differently. BookNLP appears to tag the majority of honorifics, yet Flair mostly tags unabbreviated honorifics (e.g. Miss) and excludes abbreviated ones (e.g. Mr.). Due to the fact that most honorifics in Pride and Prejudice are abbreviated, we see that Flair clearly scored higher in terms of precision compared to BookNLP. The same effect of the annotation rule about honorifics can be seen in the precision values using the new annotated dataset following CoNLL-2003.

In terms of the precision values observed using LitBank as the gold standard, we noticed that the gold standard does not consider Alice's cat "Dinah" to be a person

[12] `https://huggingface.co/flair/ner-english-large`

[13] A permanent link to the most recent location of the repository can be found at `https://rivanova.org/lrec2022`

| Novel | LitBank | | | OWTO | | | New (CoNLL) | | | New (Ext) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Alice in Wonderland | 80.00 | 54.05 | 64.52 | 92.00 | 74.19 | 82.14 | 100.00 | 80.65 | **89.29** | 100.00 | 12.89 | 22.83 |
| David Copperfield | 100.00 | 18.06 | 30.59 | 44.44 | 85.71 | **58.54** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Dracula | 45.45 | 10.87 | 17.54 | 14.29 | 33.33 | 20.00 | 45.45 | 35.71 | **40.00** | 45.45 | 2.16 | 4.12 |
| Emma | 86.67 | 38.24 | 53.06 | 83.10 | 98.33 | **90.08** | 40.00 | 36.73 | 38.30 | 37.78 | 6.59 | 11.22 |
| Frankenstein | 77.78 | 9.33 | 16.67 | 41.67 | 100.00 | 58.82 | 77.78 | 70.00 | **73.68** | 77.78 | 2.47 | 4.79 |
| Huckleberry Finn | 82.61 | 33.33 | 47.50 | 73.53 | 78.12 | **75.76** | 60.87 | 50.00 | 54.90 | 56.52 | 4.91 | 9.03 |
| Moby Dick | 71.43 | 6.58 | 12.05 | 37.50 | 100.00 | 54.55 | 71.43 | 62.50 | **66.67** | 42.86 | 1.42 | 2.75 |
| Oliver Twist | 73.33 | 11.96 | 20.56 | 70.00 | 100.00 | 82.35 | 93.33 | 87.50 | **90.32** | 86.67 | 6.81 | 12.62 |
| Pride and Prejudice | 95.74 | 42.06 | 58.44 | 73.08 | 98.28 | **83.82** | 31.91 | 31.25 | 31.58 | 29.79 | 4.58 | 7.93 |
| The Call of the Wild | 84.21 | 14.81 | 25.20 | 94.74 | 41.86 | **58.06** | 84.21 | 37.21 | 51.61 | 78.95 | 6.52 | 12.05 |
| Ulysses | 92.98 | 50.48 | 65.43 | 81.58 | 98.41 | 89.21 | 96.49 | 94.83 | **95.65** | 92.98 | 18.21 | 30.46 |
| Vanity Fair | 70.15 | 31.33 | 43.32 | 74.59 | 88.35 | **80.89** | 22.39 | 18.99 | 20.55 | 14.93 | 4.50 | 6.92 |
| Mean | 80.03 | 26.76 | 37.91 | 65.04 | 83.05 | **69.52** | 60.32 | 50.45 | 54.38 | 55.31 | 5.92 | 10.39 |
| Standard deviation | 14.46 | 16.89 | 19.75 | 24.8 | 23.1 | 20.34 | 32.32 | 29.02 | 29.9 | 32.14 | 5.11 | 8.65 |
| Median | 81.31 | 24.7 | 36.96 | 73.31 | 93.32 | **78.33** | 66.15 | 43.61 | 53.26 | 50.99 | 4.75 | 8.48 |

Table 3: Evaluation of BookNLP

| Novel | LitBank | | | OWTO | | | New (CoNLL) | | | New (Ext) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Alice in Wonderland | 76.92 | 54.05 | 63.49 | 92.31 | 82.76 | 82.27 | 100.00 | 83.87 | **91.23** | 100.00 | 13.40 | 23.64 |
| David Copperfield | 87.50 | 19.44 | **31.82** | 6.25 | 7.69 | 6.90 | 6.25 | 6.67 | 6.45 | 6.25 | 0.49 | 0.90 |
| Dracula | 57.14 | 8.70 | 15.09 | 0.00 | 0.00 | 0.00 | 57.14 | 28.57 | **38.10** | 57.14 | 1.72 | 3.35 |
| Emma | 60.00 | 26.47 | 36.73 | 57.78 | 68.42 | 62.65 | 71.11 | 65.31 | **68.09** | 68.89 | 12.02 | 20.46 |
| Frankenstein | 46.15 | 8.00 | 13.64 | 15.38 | 50.00 | 23.53 | 69.23 | 90.00 | **78.26** | 61.54 | 2.83 | 5.41 |
| Huckleberry Finn | 72.41 | 36.84 | 48.84 | 62.07 | 81.82 | 70.59 | 72.41 | 75.00 | **73.68** | 68.97 | 7.55 | 13.61 |
| Moby Dick | 88.89 | 10.53 | 18.82 | 33.33 | 100.00 | 50.00 | 88.89 | 100.00 | **94.12** | 66.67 | 2.84 | 5.45 |
| Oliver Twist | 6.67 | 10.87 | 18.69 | 66.67 | 90.91 | 76.92 | 86.67 | 81.25 | **83.87** | 80.00 | 6.28 | 11.65 |
| Pride and Prejudice | 29.41 | 14.02 | 18.99 | 21.57 | 31.43 | 25.58 | 94.12 | 100.00 | **96.97** | 88.24 | 14.71 | 25.21 |
| The Call of the Wild | 88.24 | 27.78 | 42.25 | 88.24 | 75.00 | **81.08** | 88.24 | 69.77 | 77.92 | 85.29 | 12.61 | 21.97 |
| Ulysses | 90.48 | 54.29 | 67.86 | 71.43 | 100.00 | 83.33 | 88.89 | 96.55 | **92.56** | 87.30 | 18.90 | 31.07 |
| Vanity Fair | 37.23 | 23.33 | 28.69 | 36.17 | 49.28 | 41.72 | 55.32 | 65.82 | **60.12** | 41.49 | 17.57 | 24.68 |
| Mean | 61.75 | 24.53 | 33.74 | 45.93 | 61.44 | 50.38 | 73.19 | 71.9 | **71.78** | 67.65 | 9.24 | 15.62 |
| Standard deviation | 27.33 | 16.42 | 18.6 | 31.46 | 34.1 | 30.37 | 25.47 | 28.56 | 26.5 | 25.09 | 6.44 | 10.16 |
| Median | 66.21 | 21.39 | 30.26 | 46.98 | 71.71 | 56.33 | 79.54 | 78.13 | **78.09** | 68.93 | 9.79 | 17.04 |

Table 4: Evaluation of Flair

entity in “Alice in Wonderland”, however the tools tag the cat as an entity.

When we look at the recall values achieved by the tools, it is surprising that both tools achieve 100% recall for some of the novels when evaluated using the OWTO gold standard. We manually confirmed that the values are correct and found out that based on the OWTO gold standard three of the novels - Frankenstein, Moby Dick, and Oliver Twist - have only 4, 3 and 11 person entities respectively. This explains why it is realistic to achieve 100% recall on all three of them. Surprisingly, Flair also correctly tagged all entities in Ulysses without scoring any false negatives despite the section of the novel having 45 entities.

Furthermore, Flair achieves 100% recall also when evaluated using the novels Moby Dick, as well as Pride and Prejudice from the new annotated dataset following CoNLL-2003. The lowest precision score by BookNLP using this gold standard is for David Copperfield, as all entities in the novel contain a honorific followed by a name and those are not treated as parts of an entity following the CoNLL-2003 guidelines. This is further the reason why the precision equals 0.00% also for the evaluation using the new extended gold standard. Those cases present two edge cases, in which we avoid the division by zero by setting the $F_1$ score to 0.

“Dracula” is the novel, on which Flair scores the lowest when the OWTO gold standard is used. The outlier score of 0.00 results from the the fact that the tool did not detect any entity entirely correctly. The tool tags “Mina” as a person twice and once as miscellaneous, while the annotator of the gold standard only treats one of the occurrences as a person. Furthermore, Flair does not tag “Count” as a part of Dracula’s name, however in the gold standard both “Count” and “Dracula” are labeled as a person entity type. Lastly, “Jonathan Harker” is also considered as a false positive, however we did not find an explanation for this. We view the tagging of "Mina" and "Jonathan Harker" with the label "O" as annotation mistakes.

Further, BookNLP and Flair score relatively low in terms of recall when evaluated using LitBank as a gold standard vs. when using OWTO or the new dataset following CoNLL-2003. The main reason for this is that the annotation guidelines of LitBank include common phrases such as "a boy" and require entities to include the entire noun phrases such as "the youngest of the two daughters of a most affectionate, indulgent father" (from the novel Emma). Those entities are tagged neither by BookNLP nor by Flair. As the new extended dataset also applies those rules, the recall values achieved by the tools using it as a gold standard are also low.

## 7. Discussion

In this section, we discuss the various limitations and problems we encountered related to data availability, standards, annotation guidelines, challenges of annotating, and evaluation.

### 7.1. Using Existing Gold Standards

Some of the best known and most frequently used corpora for English contain between 30,000 and 400,000 annotated tokens and consist of over 1 million words (Grishman and Sundheim, 1996; Tjong Kim Sang and De Meulder, 2003; Walker et al., 2006; Krug, 2020; Pradhan et al., 2011). The predominant types of texts annotated in all of these datasets are news or web articles, and social media conversations. Rösiger et al. (2018) explore the topic of coreference for literary text and state that *"literary texts differ from news texts and dialogues to a great extent, as their purpose is not to transfer information as it is the principle task of a newspaper, but rather to provide poetic descriptions and good storytelling"*. Overall literary language *"tends to use a larger set of syntactic constructions than the language of non-literary novels"* (van Cranenburgh and Bod, 2017). Furthermore, it applies a mix of direct and indirect speech, and uses rich vocabulary (Rösiger et al., 2018).

Taking into account these differences it is essential to consider the individual standards and annotation guidelines used for datasets. Choosing to use the annotated corpus of the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) might be suitable for the evaluation of a tool, written for NER in news or web articles. However, a tool such as BookNLP, which was tailor made for the literary domain may not be as good at handling such texts.

A greater focus on the differences and similarities between literary texts and other text types (e.g. historical letters) could be useful in detecting whether or not existing annotated datasets could be properly utilised for the purpose of evaluating or even training existing tools. Such analysis might help to create *"more consistent annotation of larger amounts of board coverage data for training better automatic techniques for entity and event identification"* (Pradhan et al., 2011).

### 7.2. Dataset maintenance

The number of entity types, the sizes of the datasets, and their formats that has already evolved in the 15 years between the MUC-6 (Grishman and Sundheim, 1996) and the CoNLL-2011 (Pradhan et al., 2011) tasks is large. Now, ten years later, the selection is even wider and includes examples such as LitBank, which aims to cover a new domain, yet do not follow the exact (i.e. unchanged) annotation guidelines of any of the bigger existing datasets. While projects such as LitBank could be beneficial for a specific purpose, examples from the past show that frequent changes in standards might also be harmful to the targeted progress (Pradhan et al., 2011).

Using old datasets such as CoNLL-2003 is a stable solution: no changes need to be considered, it makes evaluations comparable and easier to execute, and may reduce certain bias relating to the creation of new gold standards (e.g. interpretation of guidelines, different annotation formats). However, using a single dataset provides a very limited view on the problem of NER.

An alternative solution would be to analyse the differences and similarities between the guidelines and the purposes of said guidelines, and to create bigger clusters of corpora, consisting of similar enough datasets. Furthermore, one could analyse whether or not the individual collections of datasets could be shaped into individual homogeneous corpora. Our evaluation shows that a random combination of datasets for the evaluation of any tools is not an appropriate solution. Therefore, grouping existing datasets could only be done based on their precise characteristics. A detailed study of existing datasets would be beneficial for the creation of a better understanding of the state of the art, and for the exploration of the abovementioned ideas.

### 7.3. Evaluation Metrics

The CoNLL standard, which we followed in this paper, accepts only full matches between tool tagging and the gold standard as correct. As such, partial correctness is viewed as a mistake. There are a few aspects of this approach to be considered. Firstly, there are certain use cases (e.g. searching for a specific gene in bioinformatics), which may not require an exact match for a tag to be counted as correct. In such cases the span of the entity is less relevant than its presence. Secondly, there are some cases, in which it might be more objective to differentiate between "ambiguous" and "incorrect" tags. Ambiguity is resolved to a certain degree by the use of multiple layers. However, there are other cases of ambiguity such as in coreference resolution, which also need to be addressed. Poesio and Artstein (2005) give an example from anaphoric annotation, which shows a case, in which *"judgments may* *disagree - but this doesn't mean that the annotation scheme is faulty; only that what is being said is genuinely ambiguous"* (Poesio and Artstein, 2005). The authors propose to distinguish between cases, in which the annotators cannot come to an agreement over the correct version because of ambiguity, from cases, in which some annotators might have simply made a mistake in the tagging. Multiple correct answers could, for example, be represented by a set of answers (Poesio and Artstein, 2005). Lastly, our experiment shows how a small difference in the definition of an entity type (e.g. handling of pronouns) between the tool and the gold standard is amplified by the frequency of its occurrence. Thus, the frequent use of pronouns, for example, could drastically reduce the performance of certain tools in comparison to other tools.

### 7.4. Annotation and Training Challenges

As observed throughout our annotation process, there are many potential reasons for errors in the annotations. These may range from insufficient annotation guidelines to inexperienced annotators. An aspect relevant for the literary domain is the length of the texts used. Sometimes short datasets might have too few and repetitive entities (e.g. the novel Dracula in the section of OWTO used for our experiment). In such cases longer datasets might offer a better chance for a tool to be evaluated. From the perspective of an annotator, however, longer texts are difficult to annotate and introduce a higher risk of inconsistency (Pradhan et al., 2011). Annotators often make errors, which could be grouped together as occurring due to the lack of experience. In the process of annotating, there are two main types of knowledge that annotators apply - text knowledge and world knowledge (Rösiger et al., 2018). Text knowledge refers to the knowledge, which can be found within the text. Annotators could tag entities based on their knowledge as readers or based on the knowledge a character in the book has at the respective point of the story. In our experiment, both annotators used their readers' knowledge. The second type of knowledge - world knowledge - denotes the knowledge that a typical reader would have had at and about the time of writing of the novel. The lack of this knowledge could influence the labeling decisions of the annotators, as they might not be aware of the fact that certain tokens refer to an entity.

The importance of annotated datasets is relevant not only for the evaluation of tools, but also for training purposes. Due to the complexity of the annotation process and the lack of availability of large annotated datasets in the field of English literature, it is likely that most tools are developed or improved using existing data from other domains. This fact once again underlines the importance of correct labeling.

Over the years, there have been different approaches for the handling of datasets containing erroneous labels such as using algorithms less sensitive to noise, and improving data quality prior to using it (Frénay and Verleysen, 2013). Despite their success in certain aspects, some strategies might negatively influence other aspects of the training sets. For example, filtering labels that seem noisy based on robust loss[14] might unintentionally also filter out labels, which are more difficult to detect. Simultaneously, some wrong labels might also be similar enough to correct ones, making them indistinguishable for automated tools (Cordeiro and Carneiro, 2020). More work could be done in targeting recognised shortcomings of existing datasets instead of creating new ones to derive more precise annotated datasets. This could be done via the help of tools such as CrossWeight (Wang et al., 2019), but also by letting multiple human annotators find, discuss, and correct faulty labels.

## 8. Conclusions and Future Work

In this paper, we discuss the state of the NER tools and datasets used in the field of English literary novels. In particular, we focus on the named entity type person. We conclude that, due to the specific structures of literary texts as described by Cranenburgh et al. (van Cranenburgh and Bod, 2017), it is more reliable to use domain-specific gold standards for the evaluation of NER tools. We suggest that future work should look at the similarities to closely related domains (e.g. historical letters). A better understanding of the linguistic properties of related domains could help define how they can be used together to create more and better gold standards.

Further, we found out that when used for the evaluation of NER tools, the individual gold standards yield different and oftentimes opposite results in terms of precision, recall, and $F_1$ score. This makes the evaluation process biased even within the same domain (i.e. literary texts), as the intentional selection of a specific gold standard could lead to better evaluation results for a certain tool.

Lastly, we identify characteristics of a gold standard dataset, which should be considered when evaluating the performance of NER tools. Considering the different results yielded by the use of the four gold standards, we identify a need for agreed-upon annotation guidelines to be used for the annotation of literary novels. Lastly, we identify the annotation process as essential for the quality of the gold standard. By letting at least two people annotate the same texts and agree on one version in the end, we reduced the number of unintentional human errors in the process of the annotation.

## 9. Acknowledgements

We thank Martin Beno and Martin Trajkov for the annotation of the texts.

[14] *"Loss correction approaches usually add a regularization or modify the network probabilities to penalize less the low confident predictions, which may be related to noisy samples"* (Cordeiro and Carneiro, 2020).

## 10. Bibliographical References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Bamman, D., Underwood, T., and Smith, N. A. (2014). A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Bamman, D., Popat, S., and Shen, S. (2019). An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144. Association for Computational Linguistics, June.

Bamman, D., Lewke, O., and Mansoor, A. (2020). An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54. European Language Resources Association, May.

BBN Technologies. (2007). Co-reference guidelines for english ontonotes version 7.0. `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-coreference-guidelines.pdf`. Last accessed: 2021-04-18.

Chinchor, N., Brown, E., Ferro, L., and Robinson, P. (1999). Named entity recognition task definition. *Mitre and SAIC*.

Cordeiro, F. R. and Carneiro, G. (2020). A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 9–16. IEEE.

Dekker, N., Kuhn, T., and van Erp, M. (2019). Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189.

Elson, D., Dames, N., and McKeown, K. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 465–474.

Krug, M., Reger, I., Jannidis, F., Weimer, L., Madarász, N., and Puppe, F. (2017). Overcoming data sparsity for relation detection in german novels. In *DH*.

Krug, M. (2020). *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. Doctoral thesis, Universität Würzburg.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Linguistic Data Consortium. (2006). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities. `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf`. Last accessed: 2021-03-04.

Mac Kim, S. and Cassidy, S. (2015). Finding names in trove: named entity recognition for australian historical newspapers. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 57–65.

McHugh, M. L. (2012). Interrater reliability: The Kappa statistic. *Biochemia medica*, 22(3):276–282.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Rösiger, I., Schulz, S., and Reiter, N. (2018). Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138. Association for Computational Linguistics, August.

Schweter, S. and Akbik, A. (2020). Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*.

Sims, M., Park, J. H., and Bamman, D. (2019). Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

*Linguistics*, pages 3623–3634. Association for Computational Linguistics, July.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Vala, H., Jurgens, D., Piper, A., and Ruths, D. (2015). Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774. Association for Computational Linguistics, September.

van Cranenburgh, A. and Bod, R. (2017). A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1228–1238.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., and Han, J. (2019). Crossweigh: Training named entity tagger from imperfect annotations. In *Proc. 2019 Conferenc. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing, EMNLP-IJCNLP 2019,*, volume 1.

Woldenga-Racine, V. (2019). *Issues in Named Entity Recognition on Early Modern English Letters*. Master thesis, University of Washington.

Zhu, G. and Iglesias, C. A. (2018). Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, 101:8–24.