


# Named Entity Recognition

Evelyn Gius

 <https://orcid.org/0000-0001-8888-8419>

**Abstract** This chapter introduces the automatic recognition of entities in texts using the method of *Named Entity Recognition*. After defining *Named Entities*, initial considerations regarding their recognition are presented. Then, the chapter outlines the development of *Named Entity* systems in language processing and the most important associated models. The applicability of *Named Entity Recognition* in theology is then examined and practical tips for testing *Named Entity* systems are provided. The chapter concludes with references to tools and resources for *Named Entity Recognition*.\*

**Keywords** Named Entity Recognition, Entities, Proper Names, Language Processing

## 1. What is *Named Entity Recognition*?

In language processing, words or expressions in a text referring to specific entities in the world are referred to as *Named Entities* and their automatic recognition as *Named Entity Recognition*.<sup>1</sup> *Named Entities* specifically include expressions for specific persons, places, or organizations. In principle, *Named Entities* have a clearly defined identity and can be identified by a name or a specific term, i. e., they can be named. In addition to proper names, *Named Entities* also include other designations. For example, both the personal proper name “Hildegard,” and the specific term “The Master of Rupertsberg,” are *Named Entities* of the personal entity type.

The recognition of *Named Entities*, also known as *Named Entity Recognition* (NER), is an important method in text processing and analysis. NER systems identify entities and classify them into predefined categories such as “persons,” “places,” or “organizations.” To do so, the systems use machine learning methods based on linguistic information and semantic correlations in texts. Some systems also use lists of known proper names, which are called *Gazetteers* in language processing. These are particularly helpful for places and other geopolitical entities, whereby the recognition results – possibly counterintuitively – are better when using fewer, highly

\* This chapter, including quotations in foreign languages, was translated from German by Brandon Watson.

1 For many disciplines in the humanities, one must of course add: in addition to real world entities, these entities can also be present in narrated worlds.

frequent proper names. Extensive lists of less frequent proper names, on the other hand, diminish the results.

Like all computational methods in language processing, NER systems were initially developed based on rules, whereas nowadays machine learning methods achieve better results, although the phenomena previously described in the rules do play a role (see the section “The development of NER systems”). Regardless of the technology used, *Named Entity Recognition* systems use sequence tagging approaches, in which each element of a sequence is assigned a value. For example, each word in a text is assigned the information as to whether it is a *Named Entity* and, if so, which class. Strictly speaking, *Named Entity Recognition* consists of two tasks: Recognizing *Named Entities* (identification) and classifying the recognized *Named Entities* into the predefined classes (classification).

The *Named Entity* classes used differ depending on the system. Many NER systems recognize the classes persons, locations, and organization (Tab. 1), which are typically designated as PER (cf. *person*), LOC (cf. *location*), ORG (cf. *organization*). Most systems also have a fourth class, comprising either of geopolitical entities (GPE, for *geo-political entity*) or a residual class (MISC, for *miscellaneous*). In addition to persons, places, and organizations, other classes and corresponding expressions are also considered *Named Entities*, e.g., dates (“September 17, 1179”), quantities (“five kilograms”), abstract terms (“religion”) and general classes (“monastery”).

**Tab. 1** The three most common *named entity* classes in NER systems.

| Class        | Tag | Example for Entities                          | Example ( <i>Named Entity</i> in bold)                            |
|--------------|-----|---|---|
| Person       | PER | human, figures, saints                        | <b>Abigajil</b> prevents further violence.                        |
| Location     | LOC | cities, mountains, countries, bodies of water | There are numerous regions in <b>South Asia</b> .                 |
| Organization | ORG | companies, associations, institutions         | The <b>Roman Catholic Church</b> is the largest Christian church. |

Understood within the context of language processing, *Named Entity Recognition* is a well-established and widely used technique. Along with other basic operations – such as the segmentation of text into word and sentence units (*Tokenization*, *Sentence Splitting*) and the tagging of word types and syntactic units (*Part-of-Speech Tagging*, *Dependency Parsing*) – NER is a pre-processing step in most language processing pipelines.<sup>2</sup> In machine learning processes, *Named Entities* are used as a *feature*, i.e., an aspect

2 On the structure of language processing pipelines, see Biemann et al. (2022, 85 ff.). The introduction is also suitable for deepening some of the other language processing methods mentioned here.

included in the analysis of texts in a wide variety of tasks, whereby the systems calculate an appropriate consideration (the so-called weighting) of the *feature* in the course of the learning process.<sup>3</sup>

In addition to the recognition of *Named Entities* in texts, typical applications of NER in language processing include methods that are based on the results and analyze further semantic information. These methods include the extraction of relationships between the entities (e.g., family relationships between persons, spatial relationships between places or persons and places, etc.), the creation of knowledge graphs in databases used for search engines, and the recognition of events, enabling further semantic textual analysis. The fields in which NER is used are correspondingly diverse. They range from scientific research to government institutions and companies. NER is also used to create market analyses, track customer feedback, and gain intelligence on potential threats, as well as to analyze historical texts and examine cultural developments.

## 2. An Initial Approach to Entity Recognition

A few examples of possible textual features that could be used to identify *Named Entities* systematically will suffice for illustrating the approach.

Example 1: In those days, a decree went out from **Caesar Augustus** that all the world should be registered. This registration was the very first and took place at the time when **Quirinius** was governor of **Syria**.

In this example, two personal entities (“Caesar Augustus” and “Quirinus”) and one location entity (“Syria”) are mentioned. Textual features used to recognize these entities could be the spelling. In German, as some other languages, proper names are capitalized.<sup>4</sup> Another characteristic of personal entities is also that proper names do not normally have an article, distinguishing them from other nouns. The term “Caesar Augustus” also includes the title “Caesar.” One can thus formulate a rule that titles and subsequent capitalized words denote personal nouns. “Syria” is also a proper noun recognizable by its capitalization. Moreover, certain prepositions such as “in,” “of,” etc. can refer to a location entity.

3 On the use of features in machine learning methods, see Jurafsky & Martin (2023, 59; 60ff.). The introduction is also suitable for in-depth study of *Named Entity Recognition* and all other language processing methods mentioned.

4 In German, however, capitalization applies not only to proper names, but also to nouns, which means that many other words also have this feature and makes it less easy to recognize *Named Entities* than in English, for example.

Example 2: **Saint Teresa of Ávila** was born in **Spain** in the **16<sup>th</sup> century**. Her mystical experiences led to important writings.

In this example, there are three mentioned entities: a person entity (“St. Teresa of Ávila”), a date entity (“16<sup>th</sup> century”), and a location entity (“Spain”). The mentioned features can be used for the person and place entities. The “saint” is also a kind of title, although one must conjure up rules for saints, such as the combination of the adjective “saint” preceded by “of” and a place name. For the date, one might define a series of formats that typically combine numbers, punctuation marks, and words, thus distinguishing them from other expressions.

Example 3: **Francis of Assisi** founded the **Franciscan order** in **Italy**.

The bold words in the third example are: a person entity (“Francis of Assisi”), an organization entity (“Franciscan order”), and a place entity (“Italy”). For “Francis of Assisi,” a partial rule of the rule of saints could be used, particularly, the scheme [proper name] “of” [place entity]. The same rules apply to “Italy” as to “Syria” and “Spain” in the previous examples. The “Franciscan Order,” on the other hand, can be recognized by the fact that the expression consists of a capitalized but not very frequent word introduced with a definite article. Presumably, a rule can also be derived from the composition, since “Franciscan” is a name derived from a proper name and “order” is a general organizational term.

The considerations on the three examples are intended to show that *Named Entities* can be distinguished from other expressions based on textual features. These features include spelling, the use of certain prepositions, or other combinations of word type sequences, as well as features on the character level such as capitalization, letter sequences, or the use of characters atypical for other word types such as numerals or punctuation, the syntactic structure (where in the sentence might one expect *Named Entities*?), or even typical contexts or occurrence frequencies of *Named Entities*. These features were initially used in the recognition of *Named Entities* based on corresponding rules.

### 3. The Development of NER Systems

The emergence of *Named Entity Recognition* goes back to the beginnings of computational processing of natural language in the 1950s and 60s. During this time-period, word processing systems were developed for analyzing basic linguistic information. Overall, the history of NER corresponds to the development of many language processing applications, ranging from rule-based recognition of phenomena to machine learning methods and *Deep Learning* approaches.

The first NER approaches focused mainly on identifying the names of people and places. For these identifications, they defined rules or patterns to target specific properties of proper names as in the examples discussed above. These rule-based methods enabled the identification of names in texts based on certain characteristics such as capitalization or special characters. However, heuristic approaches are limited. They did not achieve satisfactory results due to the variety of named entities and contexts in which they occur.

The use of machine learning techniques in NER, which emerged in the 1990s, led to an improvement in the systems.<sup>5</sup> Statistical models and machine learning algorithms were used to recognize and classify named entities based on previously manually annotated training data. *Hidden Markov* models (HMM) and *Maximum Entropy* models were used, which can take context information and statistical probabilities into account in the NER. *Hidden Markov* models can analyze the sequence of words in a text and calculate the probability of a word being a *Named Entity*. They assume hidden states unknown at the beginning (in the case of NER: the entities), as well as observable states consisting of the words in the text. The model is trained to optimize the transition probabilities between these states and the output probabilities for each word. *Entity Recognition* is based on the most probable state transitions determined in this way, which establish a link between the unknown (or hidden) entity classes and the observable words. *Maximum Entropy* models (MaxEnt) are also probabilistic models. They are based on maximum entropy principles that optimize probabilities for a set of classes or categories. In NER, maximum entropy models predict the association of *Named Entity* categories by using trained weightings of appropriate text features. These features could be words, contextual information, capitalization, etc. The aim is to adjust the weights of the features so that the probability for each entity category is calculated in terms of maximum entropy.

The performance of NER systems has been further increased by the establishment of *Deep Learning*. Artificial neural networks, which “learn” phenomena in a large amount of data using several layers, are now being used to recognize *Named Entities*. The first *Deep Learning* systems to be used were *recurrent neural networks* (RNNs), which have been around since the 1980s but were only used in NER in the 2000s. RNNs are neural networks that have been specifically developed for processing sequential data. They can be used to process the sequence of words in a text and calculate the probability of each word belonging to a particular class. Unlike the models mentioned above (HMMs and MaxEnt), the RNN also considers the context of the previous words, thus enabling a more precise recognition of entities. However, RNNs have difficulties in processing long sequences. The next development, *Long Short-Term Memory* networks (LSTM), then included the capturing of long-term dependencies in sequences. In the NER, LSTMs enable more precise modeling of relationships

5 For a brief overview of the systems developed from the early stages to the current *Transformer*-based approaches, cf. Jurafsky & Martin (2023, 183).

between words and the recognition of entities that can vary over longer sections. LSTMs can effectively utilize both local and global contextual information, advancing the capabilities of NER. A further improvement of NER systems in the early 2010s, combined bi-directional LSTM models with *Conditional Random Fields* (CRFs). Bi-directional LSTMs not only capture the context of the sequence before a particular word, but also the context after the word, which increases the quality of NER. The use of CRFs helps to model dependencies between adjacent words and their classification, which enables a more coherent assignment of entity labels. The recognition of nested entities has also been significantly improved by *Deep Learning* methods.

The current *State of the Art* systems for NER are based on pre-trained *Transformer* models such as BERT or GPT, which have been under development since the mid-2010s. *Transformers* are a further development of recurrent networks in which context-dependent information can be calculated simultaneously through *self-attention* and *memory* layers. These networks can therefore be trained on a large amount of text data and generate even better language models. Using *Transfer Learning*, these general models can then be adapted for specialized tasks such as NER.

Regardless of the models, BIO annotation (Ramshaw & Marcus 1995) for NER has become the standard approach for sequence labeling in a span recognition problem. The approach provides three labels that also account for the boundaries of the *Named Entities*. This method captures each word (or *token*) of a *Named Entity* expression as follows: The first word is given the label B (for *begin*), all following words are labeled I (for *inside*) and all words outside the *Named Entity* are labeled O (for *outside*). There are separate B and I labels for each entity class to map these. For the beginning of example 1, a BIO annotation would look like this:

|     |       |        |       |       |     |    |     |
|-----|-------|--------|-------|-------|-----|----|-----|
| The | Saint | Teresa | of    | Avila | was | in | ... |
| O   | B-PER | I-PER  | I-PER | I-PER | O   | O  |     |

**Fig. 1** Sequence encoding of a *Named Entity* of Person (PER) with BIO labels

#### 4. Challenges of Automatic *Entity Recognition*

While NER is one of the basic methods of language processing and the current NER systems achieve good results, there are still some persistent challenges in the recognition of *Named Entities*.

The linguistic form of *Named Entities* is very diverse. The wide range of inflections, derivations, morphs or syntactic rules, and word order in a language increase the complexity of recognition. NER is thus particularly difficult in morphological languages such as Hebrew. There is also a practical problem: NER systems are based on

extensive training data, so the performance of the systems depends on the availability of sufficient suitable data in the relevant language. The development of universally applicable NER systems is even more difficult by the linguistically and culturally dependent differences in grammar, syntax, and nomenclature, i.e., the way in which entities are named.

*Named Entities* are not only multiword phrases – such as “University of Tübingen” or “Mary Magdalene” – they are also sometimes nested. For the correct recognition of entities such as “Apostle Paul” or “Hildegard of Bingen,” the words belonging to the proper name must not only be recognized as the title (“Apostle” or “Bingen”); they must also be determined as belonging to the personal entity, which requires deeper semantic processing and better modeling of contexts in the text.

The fact that *Named Entities* can also be multiword phrases also makes the evaluation of NER systems more complex than with uniform segments. In contrast to *Part of Speech Tagging*, for example, where a value is assigned to each individual word, or to classification tasks that are performed for entire texts, the textual span that the respective entity covers must be determined for the NER. Given that words are typically the training unit for the NER and the output unit is entities – potentially multiword expressions – there is an incongruity. Accordingly, in the BIO annotation system, only partially recognized multiword entities are evaluated incorrectly several times because the annotations are incorrect due to the missing words (the B annotation comes one or more words too late or the O annotation too early, with corresponding consequences for the I annotations). This problem concerns the non-recognition of the same entity being evaluated in the same way and thus considered equally good or bad. However, this problem can be mitigated by a corresponding error weighting in the evaluation.

Languages considered to be data-poor, e.g., pre-modern languages, present a particular evaluative problem. There is often no further annotated data in these types of languages that can be used as *benchmarks* to check whether the evaluated NER system also achieves similarly good results with unknown texts or whether there is an *overfitting* on the training data, where only these instances are recognized.

There are several challenges that are more prerequisites. For example, a NER system only recognizes the entities in the text that are relevant to a question if they are named explicitly and with clearly defined names or expressions. NER is therefore not suitable for recognizing pronouns, generic expressions, unspecific terms, and indirect references to *Named Entities*. Moreover, there are difficulties in recognizing entities like abstract concepts and entities if named by infrequently used technical terms or local names.

While the latter difficulties are certainly addressed by NER systems, the recognition of pronouns or the like is not part of NER, for which are many pragmatic reasons. One reason is that the additional challenges of the closely related task of coreference resolution would also have to be solved. Coreference resolution is determining when pronouns, demonstrative expressions, or other referential elements in the text



refer to previously mentioned entities. Coreference resolution requires an in-depth understanding of the context and semantic relationships of a text. Additionally – esp. in the field of theology – there are relevant questions of identity, since all expressions referring to the same entity must be identified. This problem is often difficult in less obvious cases than the Trinity because the identity of many entities is difficult to ascertain, e.g., in the case of temporal or other changes. For example, a school of thought can be perceived as the same over decades or divided into certain sections, several organizational entities, such as a family, can be a single entity or the addition and removal of family members can each be perceived as new families, or the life phases of a person with very different views and actions can also be perceived as separate personal entities.<sup>6</sup>

## 5. NER in Theology?

Since the techniques of automatic language processing can be used in any science focusing on text analysis, NER can be used in theological research.<sup>7</sup> In principle, applying these methods is possible and useful in all areas in which entities such as persons, places, dates, or concepts or their relationship to each other are relevant to a research interest. Potential fields of application range from the identification of specific phenomena in individual texts to the analysis of large volumes of text or corpora. In addition to identifying the corresponding *Named Entities*, the NER is also suitable for analyzing the distribution, interrelationship, and developments over time, of the entities. Developments can also be compared with different text groups or grouping of texts based on *Named Entities*. An analysis based on *Named Entities* can be aimed at the question of the most frequent mentions of actors or places in religious texts or the quantitative comparison of the respective proportions of mentions between different texts or text groups. Questions about the first mention and subsequent development of the frequency of mentions of persons, locations, or concepts in a corpus of diachronic texts, i.e., texts that cover a longer period, can also be analyzed. An NER can be used to carry out stylistic analyses – such as in homiletics – or to identify texts in a corpus that relate to a specific topic recognizable via *Named Entities*.

These types of applications lead to interesting findings with the NER relevant for theology. Nevertheless, *Named Entity Recognition* is not widespread in theology and has not yet had any recognizable significance in publications relevant to digital

6 For an in-depth consideration, see the *Stanford Encyclopedia of Philosophy*. On the identity problem, see Noonan & Curtis (2022). On the problem of fictional entities, see Kroon & Voltolini (2023).

7 For an overview of language processing methods in the humanities, see Piotrowski (2012); Sporleder (2010); and the methodological introductions in the forTEXT portal at <https://fortext.net/routinen/methoden> (Accessed: 17 June 2024).



approaches.<sup>8</sup> There are several reasons why NER has not gained any traction in religious studies. One reason is that the application of language processing techniques in the humanities is generally still a relatively young branch of research beyond computer and corpus linguistics. In addition, there is a hesitancy towards the use of computational tools in theology as well as in other humanities that work more exemplarily or hermeneutically. Finally, the so-called operationalization of a question, i. e., the translation of the question into qualities that can be measured by *Named Entities*, is not a trivial task, which is methodologically contrary to the established practices of theological text analysis. However, if recent developments in the field of *Digital Theology* are examined, one can assume that some progress will also be made in the field of Computational Theology in the coming years and that NER methods will also be used. However, even if any reservations have been dispelled and the necessary skills for the implementation of NER are available, there are limitations to the quality of the analyses that must be considered.

## 6. Notes on the use of NER systems

Like most language processing methods, NER systems are typically developed for English and based on news articles or texts found on the internet. Therefore, for languages other than English, or for text types other than news and internet texts, the quality of available systems decreases. Moreover, the results are often different depending on the *Named Entity* class. While the classic categories for persons, locations, and organizations are usually well recognized and achieve recognition rates of over 90% in the better systems, the recognition quality for other categories is considerably lower. Nonetheless, the NER can also be used in cases where the systems do not work optimally if prepared and implemented accordingly.

Prior to using a NER system, one should assess the extent to which the quality of the recognition is sufficient to make reliable statements based on the results. In language processing, results with a F1 value of 0.8 or more are considered very good, while results of 0.95, which are now achieved in NER for English – and occasionally

8 For example, NER is only mentioned once in Heyden & Schröder (2020) or Sutinen & Cooper (2021). NER is not mentioned at all in the publication series *Introductions to Digital Humanities – Religion* (ed. by Claire Clivaz, Frederik Elwert, Kristian Petersen, Ortal-Paz Saar and Jeri Wieringa) nor in the *Digital Biblical Studies* (ed. by Claire Clivaz and Ken M. Penner). Searches in catalogs were also virtually fruitless: a search for NER in the Religious Studies Bibliography of the Specialized Information Service (FID) at <https://www.relibib.de> (Accessed: 17 June 2024) yields only one hit (Blouin 2021), which is potentially relevant but not pertinent. There are no theological titles among the results of the search in the University of Frankfurt catalog for “Named Entity Recognition.” Even if there are individual publications not found in these search attempts, the lack of results indicates at least a low relevance of NER in theology to date.

for other languages such as German – are considered (almost) perfect. The F1 value is made up of the values for the measures of recall and precision. Accordingly, an F1 value of 0.8 means that the average proportion of phenomena found in the text (recall) and of correctly identified passages among the passages found (precision) is 80%. Since the value is an average of the two values and these are in turn calculated for several subcategories (persons, locations, organizations, etc.), the F1 value does not indicate anything about the quality of recognition for specific aspects. The F1 value is therefore – like any evaluation measure – only a guide value for the actual quality of the application. The value usually indicates nothing about the suitability of the system for the specific research interest. One must first check the extent to which a system delivers suitable results for the research question and the text corpus used. A quality check is more important if further steps based on the NER are implemented automatically, such as the recognition of entity relations or the coreference resolution, in which all entity names and other possible references – such as pronouns – to one and the same entity are recognized.

If a NER system is used for texts that differ from the texts used and evaluated during the development of the system, a specific check of the recognition quality should therefore be carried out beforehand. Ideally, the system should be evaluated based on an annotated test data from the corpus used, i.e., a meaningful F1 value should be created for the specific research requirements. However, at least a sample check of the output results and individual text parts should be carried out regarding the phenomena found. The sample check can be used to assess whether a system correctly recognizes the phenomena searched for and to what extent it may include false phenomena. In addition, possible systematic errors can be recognized, e.g., whether a location name is incorrectly recognized as a personal name, whether certain multiword expressions are not or only partially recognized, or whether individual terms tend not to be recognized. Such errors can greatly distort the text analysis, depending on the type of error. For example, if one wishes to compare the relevance of certain concepts in texts, one should ensure that one of the concepts is not recognized significantly worse than the others and is therefore found less frequently in the texts.

If the quality of the system is unsatisfactory and cannot be used for automatic analysis, there are still two ways for it to be used, both involving a further manual check of the results and thus ensure an analysis based on them. First, each system can be used as a heuristic system and point out any interesting aspects of the analyzed texts. Even if the results of the NER are not evaluated quantitatively – which should not be done anyway if the results are not good enough – their results can be used as an indication of potentially interesting texts or text passages. Perhaps the NER can be used to find a text that has not previously been recognized as relevant in a certain context, or one comes across terms that have not yet been considered, although they were prominent and relevant at a certain time or in certain texts. The results may also reveal connections due to the common occurrence of entities that have not previously been considered.

Second, NER systems not suitable for automatic analysis can be used as a pre-processing step that provides data for subsequent manual processing. If a system has an acceptable *recall*, i. e., finds a good proportion of the phenomena searched for, the quality of the data can be improved significantly by manual processing. To do so, the incorrect results are sorted out. The remaining data can then be used in further – even manually supported – steps of analysis or for a quantitative evaluation. This approach is viable for coreference resolution in longer texts because checking and correcting the coreference chains requires comparatively little effort. Manual processing essentially consists of correcting the incorrect mentions of entities in the coreference chains, which contain all mentions of an entity in a text, and rejoining chains that may have been separated due to recognition errors. Depending on the knowledge gained from the data prepared, manual checks are to be considered.

## 7. Tools and Resources

There have been many types of NER systems developed in recent decades. When selecting systems and platforms, one should first note that rule-based NER methods are often suitable for simpler cases, while more complex scenarios may require machine learning. In addition, ideally several systems should be tested on the same data to identify the most suitable system. Currently, three *open source* systems are widely used in applications: the *Natural Language Toolkit* (NLTK),<sup>9</sup> *spaCy*,<sup>10</sup> and the *Stanford Named Entity Recognizer*.<sup>11</sup> All three achieve good results for various natural languages and are regularly updated. Their *Python* or *Java*-based models are comparatively easy to use. However, one might also benefit from a search for other language specific NER systems.<sup>12</sup> Platforms that allow the assembly of one's own processing pipeline are also particularly interesting for users who are not (yet) experienced. The German platform *WebLicht* is freely accessible to members of many scientific institutions and pro-

9 See <https://www.nltk.org>. For the use of NER, see <https://www.nltk.org/book/cho7.html>. All addresses mentioned in this section were accessed on 17 June 2024.

10 See <https://spacy.io/models>. For the use of NER, see <https://spacy.io/universe/project/video-spacysner-model-alt>.

11 See <https://stanfordnlp.github.io/CoreNLP/ner.html>. For the use of the Pipeline, see <https://stanfordnlp.github.io/CoreNLP/pipeline.html>.

12 There are good approaches for Latin (see, e.g., Erdmann et al. 2016), Ancient Greek (see, e.g., Yousef et al. 2022), Hebrew (see, e.g., Bareket & Tsarfaty 2021), and premodern or classical languages (see, e.g., Johnson et al. 2021 and Burns 2019).

vides various systems for both pre-processing and NER itself, which can be combined on a graphical interface and applied to provided texts in many languages.<sup>13</sup>

To develop NER systems, one must first select suitable data. There are several annotated corpora that can be reused depending on the field of application, such as the English corpus for literary texts by Bamman et al. (2019), or the German newspaper text corpora (among others) by Tjong Kim Sang & De Meulder (2003) and Benikova et al. (2014). Further annotation of pre-processed data for the NER may be particularly useful for languages with fewer resources (e.g., for Latin in the *EvaLatin* corpus by Sprugnoli et al. (2020), which is already enriched with information on lemmatization and *Part of Speech Tagging*).

Existing directories can often be reused for the creation of *gazetteers*. In principle, large directories, ideally freely available under appropriate licenses like the Creative Commons license, are suitable for this purpose. For example, corresponding *Wikipedia* categories can be used (such as man, woman, figure, saint for personal names or corresponding categories for locations, etc.) to obtain entity names.<sup>14</sup> Another source is the *Gemeinsame Normdatei* (GND), which provides authority data from catalog data on persons and other areas in a range of metadata and data services, which is also worth searching for specific data.<sup>15</sup> For historical texts, such as the ruling class of the Roman Empire in the early and high imperial period, the encyclopedia of persons of the Berlin-Brandenburg Academy of Sciences and Humanities might be used,<sup>16</sup> or even the lexicon of Greek personal names of the University of Oxford.<sup>17</sup> Institutions like the EU or individual states also provide numerous data relevant to the NER. The EU offers a large directory of names as well as a range of other information,<sup>18</sup> and the U.S. Geological Survey (USGS) provides various data on locations and other geological information.<sup>19</sup> Many internet directories are available.

There are two introductory texts recommended, one in German and one in English – the German-language exercise by Schumacher (2019) on adapting the *Stanford Named Entity Recognizer* for literary texts, which is also suitable for beginners, and the English-language introduction by Grunewald et al. (2022), which provides a low-threshold introduction to a *Python* analysis of locations in data on prisoners of war and explains how to integrate a *gazetteer*.

13 See <https://weblicht.sfs.uni-tuebingen.de/weblicht/>. For a description of the available NER models, see [https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tools\\_in\\_Detail#Named\\_Entity\\_Recognition](https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tools_in_Detail#Named_Entity_Recognition).

14 To perform these tasks structurally, among others, see <https://www.wikidata.org>.

15 Cf. the GND service *Entity Facts* at [https://www.dnb.de/DE/Professionell/Metadatendienste/Daten\\_bezug/Entity-Facts/entityFacts\\_node.html](https://www.dnb.de/DE/Professionell/Metadatendienste/Daten_bezug/Entity-Facts/entityFacts_node.html).

16 Cf. *Prosopographia Imperii Romani saec. I. II. III.*, available at <https://pir.bbaw.de>.

17 Cf. <https://www.lgpn.ox.ac.uk>.

18 For an overview, cf. <https://data.jrc.ec.europa.eu>, and for a list of names, see <https://data.jrc.ec.europa.eu/dataset/jrc-emm-jrc-names>.

19 Cf. <https://www.usgs.gov/products/data/all-data>.

## References

- Bamman, D., Popat, S., & Shen, Sh. (2019). An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North* (pp. 2138–2144). Minneapolis, Minnesota: Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/N19-1220> (Accessed: 17 June 2024).
- Bareket, D., & Tsarfaty, R. (2021). Neural Modeling for Named Entities and Morphology (NEMO2), *Transactions of the Association for Computational Linguistics*, 9, 909–928. DOI: [https://doi.org/10.1162/tacl\\_a\\_00404](https://doi.org/10.1162/tacl_a_00404) (Accessed: 17 June 2024).
- Benikova, D., Biemann, Ch., & Reznicek, M. (2014). NoSta-D Named Entity Annotation for German. Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 2524–2531). Reykjavik: European Language Resources Association. URL: [http://www.lrec-conf.org/proceedings/lrec2014/pdf/276\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf) (Accessed: 17 June 2024).
- Biemann, Ch., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text. Eine Einführung in das Text Mining*, 2. Wesentlich überarbeitete Auflage. Lehrbuch. Wiesbaden [Heidelberg]: Springer Vieweg. DOI: <https://doi.org/10.1007/978-3-658-35969-0> (Accessed: 17 June 2024).
- Blouin, B., Magistry, P., & Van Den Bosch, N. (2021). Creating Biographical Networks from Chinese and English Wikipedia, *Journal of Historical Network Research*, 5(1), 303–317. DOI: <https://doi.org/10.25517/JHNR.V5I1.120> (Accessed: 17 June 2024).
- Burns, P.J. (2019). Building a Text Analysis Pipeline for Classical Languages. In M. Berti (Ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution* (pp. 159–176). Berlin/Boston: De Gruyter Saur [= *Age of Access? Grundfragen der Informationsgesellschaft*, 10]. DOI: <https://doi.org/10.1515/9783110599572-010> (Accessed: 17 June 2024).
- Ehrmann, M., Hamdi, A., Pontes, E.L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents. A Survey, *ACM Computing Surveys*, 56(2), 1–47. DOI: <https://doi.org/10.1145/3604931> (Accessed: 17 June 2024).
- Grunewald, S., & Janco, A. (2022). Finding Places in Text with the World Historical Gazetteer, *Programming Historian*, 11, no. pag. DOI: <https://doi.org/10.46430/phen0096> (Accessed: 17 June 2024).
- Heyden, K., & Schröder, B. (Eds.) (2020), *Theologie Im Digitalen Raum*, Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung*, 65(2)].
- Johnson, K.P., Burns, P.J., Stewart, J., Cook, T., Besnier, C., & Mattingly, W.J.B. (2021). The Classical Language Toolkit. An NLP Framework for Pre-Modern Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. System Demonstrations* (pp. 20–29). Online: Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2021.acl-demo.3> (Accessed: 17 June 2024).

- Jurafsky, D., & Martin, J.H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed. [Draft]. URL: <https://web.stanford.edu/~jurafsky/slp3> (Accessed: 17 June 2024).
- Kroon, F., & Voltolini, A. (2023). Fictional Entities. In E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Stanford University: Metaphysics Research Lab. URL: <https://plato.stanford.edu/archives/fall2023/entries/fictional-entities> (Accessed: 17 June 2024).
- Noonan, H., & Curtis, B. (2022). Identity. In E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Stanford University: Metaphysics Research Lab. URL: <https://plato.stanford.edu/archives/fall2022/entries/identity> (Accessed: 17 June 2024).
- Piotrowski, M. (2012). NLP Tools for Historical Languages. In id. (Ed.), *Natural Language Processing for Historical Texts* (pp. 85–100). Cham: Springer International Publishing [= *Synthesis Lectures on Human Language Technologies*]. DOI: [https://doi.org/10.1007/978-3-031-02146-6\\_7](https://doi.org/10.1007/978-3-031-02146-6_7) (Accessed: 17 June 2024).
- Ramshaw, L., & Marcus, M. (1995). Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*. URL: <https://aclanthology.org/W95-0107> (Accessed: 17 June 2024).
- Schumacher, M. (2019). Named Entity Recognition mit dem Stanford Named Entity Recognizer, *forTEXT. Literatur digital erforschen*, 1–53. URL: <https://fortext.net/routinen/lerneinheiten/named-entity-recognition-mit-dem-stanford-named-entity-recognizer> (Accessed: 17 June 2024).
- Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains, *Language and Linguistics Compass*, 4(9), 750–768. DOI: <https://doi.org/10.1111/j.1749-818X.2010.00230.x> (Accessed: 17 June 2024).
- Sprugnoli, R., Passarotti, M., Cecchini, F.M., & Pellegrini, M. (2020). Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020. 1st Workshop on Language Technologies for Historical and Ancient Languages* (pp. 105–110). Marseille: European Language Resources Association (ELRA). URL: <https://aclanthology.org/2020.lt4hala-1.16> (Accessed: 17 June 2024).
- Sutinen, E., & Cooper, A.-P. (2021). *Digital Theology. A Computer Science Perspective*. Bingley: Emerald Publishing Limited. DOI: <https://doi.org/10.1108/9781839825347> (Accessed: 17 June 2024).
- Tjong Kim Sang, E.F., & 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*, 4, 142–147. Edmonton: Association for Computational Linguistics. DOI: <https://doi.org/10.3115/1119176.1119195> (Accessed: 17 June 2024).

Yousef, T., Palladino, Ch., & Jänicke, S. (2023). Transformer-Based Named Entity Recognition for Ancient Greek. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot & P. Helling (Eds.), *Digital Humanities 2023. Collaboration as Opportunity (DH2023)* (pp. 1–3). Graz: Zenodo. DOI: <https://doi.org/10.5281/zenodo.8107629> (Accessed: 17 June 2024).