*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4573–4577
Marseille, 11–16 May 2020


# Creating a Dataset for Named Entity Recognition in the Archaeology Domain

**Alex Brandsen, Suzan Verberne, Milco Wansleeben, Karsten Lambers**
Leiden University
Einsteinweg 2, 2333CC, Leiden, The Netherlands
{a.brandsen, m.wansleeben, k.lambers}@arch.leidenuniv.nl
s.verberne@liacs.leidenuniv.nl

**Abstract**

In this paper, we present the development of a training dataset for Dutch Named Entity Recognition (NER) in the archaeology domain. This dataset was created as there is a dire need for semantic search within archaeology, in order to allow archaeologists to find structured information in collections of Dutch excavation reports, currently totalling around 60,000 (658 million words) and growing rapidly. To guide this search task, NER is needed. We created rigorous annotation guidelines in an iterative process, then instructed five archaeology students to annotate a number of documents. The resulting dataset contains ~31k annotations between six entity types (artefact, time period, place, context, species & material). The inter-annotator agreement is 0.95, and when we used this data for machine learning, we observed an increase in F1 score from 0.51 to 0.70 in comparison to a machine learning model trained on a dataset created in prior work. This indicates that the data is of high quality, and can confidently be used to train NER classifiers.

**Keywords:** Corpus Creation, Named Entity Recognition, Text Mining

## 1. Introduction

The archaeology domain, like other scientific fields, produces large amounts of textual data. Specifically, a large amount of excavation reports are available, which are created whenever an excavation is completed, detailing everything that has been found together with an interpretation of the site (Richards et al., 2015). In the Netherlands, this corpus is estimated at 70,000 documents, and is growing by 4000 each year (Rijksdienst voor het Cultureel Erfgoed, 2019). Most of these reports are created and published by individual commercial archaeology companies after they excavate, in low numbers and not widely shared.

This so-called grey literature is currently underused, even though most scholars agree that the information hidden in these reports is of immense value (Evans, 2015). The systems currently available to explore this corpus are metadata search engines that simply do not offer enough granularity for archaeologists to easily find what they are looking for. An example might be a single find from the Bronze Age which was not included in the temporal metadata as it is too specific. Currently, there is no way of finding this so called 'by-catch'; single finds of a different type than the rest of the excavation. Users of the currently available search engines report they download whole portions of the available data and manually search through PDF files one by one to find the information they are looking for (Brandsen et al., 2019).

Free text search across the entire corpus would already be a vast improvement, however this does not account for polysemy and synonymy, which occur often in archaeological texts. An example of polysemy could be the time period of the Neolithic, which can also be expressed as the Late Stone Age, 11,000 - 2000 BC, 13,000 BP, etc. And the other way around, there are terms like 'Swifterbant' that can mean a time period, an excavation, a specific type of pottery or a town in the Netherlands. To alleviate this problem, we have applied Named Entity Recognition (NER) to the dataset, to automatically extract and distinguish between these entity types. We are building an online search system that allows archaeologists to search through these entities, as well as full text search, using an intuitive interface. The system is called AGNES (**A**rchaeological **G**rey-literature **N**amed **E**ntity **S**earch)[1]. The overall goals of the project, a description of the first version of AGNES, and a user requirement solicitation study can be found in a previous publication (Brandsen et al., 2019).

As we are using machine learning for the Named Entity Recognition, a labelled dataset is needed as training data. A Dutch dataset created in the ARIADNE project (Vlachidis et al., 2017) was used initially in this project, but after some experiments we found that the data was of insufficient quality, with some entities being annotated incorrectly and some having inconsistent and inaccurate span lengths. For example, often (but not always) a quantifier was included in the span for time periods, e.g. "roughly around 200BC", where the correct entity would be just "200BC". When using this dataset as training data for a sequence labelling classifier with Conditional Random Fields (CRF) (Lafferty et al., 2001), we only managed to reach an F1 score of 51% (Brandsen et al., 2019). To see if we could alleviate these problems, we created a new training dataset.

The research questions for this paper are:

- How high is the inter-annotator agreement, and by proxy, the reliability of the newly created dataset?
- To what extent will creating a more rigorous dataset yield higher accuracy in Named Entity Recognition?

The training dataset is available for download[2] (Brandsen, 2019).

## 2. Related Work

The go-to benchmark for Dutch Named Entity Recognition is the CONLL-2002 shared task, for language independent

[1] Which can be found at http://agnessearch.nl
[2] doi.org/10.5281/zenodo.3544544

| Entity | Description | Examples |
|---|---|---|
| Artefact | An archaeological object found in the ground. | Axe, pot, stake, arrow head, coin |
| Time Period | A defined (archaeological) period in time. | Middle Ages, Neolithic, 500 BC, 4000 BP |
| Location | A placename or (part of) an address. | Amsterdam, Steenstraat 1, Lutjebroek |
| Context | An anthropogenic, definable part of a stratigraphy. Something that can contain Artefacts | Rubbish pit, burial mound, stake hole |
| Material | The material an Artefact is made of. | Bronze, wood, flint, glass |
| Species | A species' name (in Latin or Dutch) | Cow, *Corvus Corax*, oak |

Table 1: Descriptions and examples for each entity type. Examples are translated from Dutch.

NER, which includes a Dutch dataset. But this task only looks at common, general-domain entities and is not comparable to our dataset (Tjong Kim Sang, 2002).
In the archaeology domain, NER datasets exist in other languages (English and Swedish), created in the ARIADNE project (Vlachidis et al., 2017). To our knowledge, the only directly related dataset that deals with both Dutch and archaeological texts is another dataset created in the same ARIADNE project, as briefly described in the introduction. As we are going to show in this paper, the dataset we have created is of better quality and much larger than the ARIADNE data.

## 3. Dataset Collection

From the total available corpus (70k documents), we currently have access to ~60,000 excavation reports and related documents, such as appendices, drawings and maps. These texts have been gathered by DANS (Digital Archiving and Networked Services) in the Netherlands, over the past 20 years. We received the documents from DANS as PDF files, and have used the pdftotext tool (Glyph & Cog LLC, 1996) to convert these to plain text. This dataset contains 30,152,318 lines and 657,808,600 words (as counted by the command line tool "wc").
The texts are quite diverse; the dates of publication span decades with the earlier ones having been scanned and OCRd from hardcopies created in the 80s. The other temporal variation is in how old the found artefacts are, ranging from 200,000 BC to the present. Also, the type of research can be very different between reports, some might describe a short desk evaluation of a small area without any fieldwork, while others detail huge excavations over multiple years with detailed analysis by a team of specialists. To get a representative sample across all these ranges, a random sampling strategy would not be ideal, and we instead opted to manually select documents, taking into account the variation described above. We selected a total of 15 documents as annotation candidates (~42,000 tokens).
For the purposes of calculating the inter-annotator agreement and evaluating the annotation guidelines, we manually selected roughly 100 sentences from these documents containing all the entity types (Table 1, explained below) and specific difficult cases as validation set, annotated by all annotators.

## 4. Annotation Setup

As an annotation tool, we used Doccano (Nakayama, 2019), an open source and intuitive system. After comparing the system to other available entity tagging tools, we found this was the easiest to use and most efficient tool for our purposes. The system was set up on a web server, data was uploaded for each user and entity types defined within the system.

### 4.1. Annotation Guidelines

The annotation guidelines were created in an iterative process. A first draft was created, containing general guidelines as well as specific examples of difficult situations. Two archaeologists used the guidelines to annotate around 100 sentences, and these annotations were compared to our own desired annotations to see where problems and inconsistencies were encountered. This information was then used to update the guidelines, after which they were tested again. This led to an inter-annotator agreement (F1 score, further explained in section 5.1.) of 0.94 between the two testers, which we consider sufficient for this task.
During the annotation process itself, whenever one of the annotators ran into a situation that was unclear, this was added as an example to the guidelines.
The annotation guidelines (in Dutch) can be downloaded as part of the dataset (Brandsen, 2019).

### 4.2. Entity Types

Table 1 lists the targeted entities and provides a brief explanation of each type with some examples. With the exception of location, these are all uncommon entity types, not occurring in general-domain Named Entity Recognition tasks. The entity types have been chosen based on a user requirement study, where archaeologists indicated which entities they would like to search on.

### 4.3. Annotation Process

To carry out the annotation work, we recruited five Dutch archaeology students at the Bachelor level. We specifically selected students in their second and third year, as some basic knowledge of archaeology is extremely helpful in determining whether a word is a specific entity or not.
The students were asked to annotate a total of 16 hours each, over a two week period, during which they could come and work at times that suited them, a few hours at a time. We opted not to have the students work a whole day on this task, as the annotation process is tedious and monotonous, which makes it hard to keep concentration. Loss in concentration can cause mislabelling, and so having them work for only small amounts of time might help prevent this.
The students were first asked to thoughtfully read the guidelines and ask any questions. During annotation, we were

| | |
|---|---|
| Documents | 15 |
| Sentences | 33,505 |
| Avg. sentences per document | 2,234 |
| Tokens | 439,375 |
| Avg. tokens per sentence | 13.1 |
| Annotation spans | 31,151 |
| Annotated tokens | 42,948 |
| Avg. tokens per annotation | 1.38 |

Table 2: Annotated corpus statistics.

always present to resolve difficult sentences and entities and explain to the students how to handle these. The students reported this to be very helpful, and learned from each other's problems. Most of these issues were relatively rare edge case though, and the original annotation guidelines covered most encountered entities sufficiently.

## 5. Annotated Corpus Statistics and Results

Table 2 lists general statistics on the annotated corpus, including number of documents, sentences, tokens, annotations and averages over these categories.

Over a total of 90 hours, the students annotated ~31,000 entities, setting the average annotation rate at 346 per hour, or 5.7 per minute, which is higher than we expected. The previous dataset we used contained only around 11,000 annotations, so we almost tripled the amount of available training data. While this seems like a large amount of entities, the amount of tokens seen by annotators is but a fraction (0.066%) of the total number of words in the dataset. The breakdown per entity type is shown in Table 3.

### 5.1. Inter-annotator Agreement

For most tasks, Cohen's Kappa is reported as a measure of inter-annotator agreement (IAA), and is considered the standard measure (McHugh, 2012). But for Named Entity Recognition, Kappa is not the most relevant measure, as noted in multiple studies (Hripcsak and Rothschild, 2005; Grouin et al., 2011). This is because Kappa needs the number of negative cases, which isn't known for named entities. There is no known number of items to consider when annotating entities, as they are a sequence of tokens. A solution is to calculate the Kappa on the token level, but this has two associated problems. Firstly, annotators do not annotate words individually, but look at sequences of one or more tokens, so this method does not reflect the annotation task very well. Secondly, the data is extremely unbalanced, with

| Entity Type | Quantity |
|---|---|
| Artefact (ART) | 8,987 |
| Time Period (PER) | 8,358 |
| Location (LOC) | 4,436 |
| Context (CON) | 5,302 |
| Material (MAT) | 1,225 |
| Species (SPE) | 2,843 |
| TOTAL | 31,151 |

Table 3: Number of annotations per entity type in the dataset

| | |
|---|---|
| Cohen's Kappa on all tokens | 0.82 |
| Cohen's Kappa on annotated tokens only | 0.67 |
| F1 score | 0.95 |

Table 4: Inter-annotator agreement measures on 100 sentence test document. Calculated by doing pairwise comparisons between all combinations of annotators and averaging the results.

the un-annotated tokens (labelled "O") vastly outnumbering the actual entities, unfairly increasing the Kappa score. A solution is to only calculate the Kappa for tokens where at least one annotator has made an annotation, but this tends to underestimate the IAA. Because of these issues, the pairwise F1 score calculated without the O label is usually seen as a better measure for IAA in Named Entity Recognition (Deleger et al., 2012). However, as the token level Kappa scores can also provide some insight, we provide all three measures but focus on the F1 score. The scores are provided in Table 4. These scores are calculated by averaging the results of pairwise comparisons across all annotators. We also calculated these scores by comparing all the annotators against the annotations we did ourselves, and obtained the same F1 score and slightly lower Kappa (-0.02).

### 5.2. New NER Results

We have used these entities as new training data, using the same CRF model as mentioned in the introduction (Brandsen, 2018), and have seen a large increase in the overall micro F1 score, from 0.51 to 0.70, showing that this data is of better quality than the previously used training data. The difference between this, and the F1 between five human annotators (0.95) indicates that there is also still room for improvement.

In Table 5 we show the difference in F1 score per entity type. Most types see a substantial increase, especially Locations, while the Material category sees a decrease in F1 score. We wondered if this could be explained by the fact that we have much fewer annotations for the Material category, only 1,078 while all other categories have at least double that amount.

To assess this, we divided the dataset into 10 chunks, and retrained the CRF model 10 times, every time adding one more chunk of data. In Figure 1 we have plotted the F1 score for individual entity types and the overall micro F1

| | Old | New | Difference |
|---|---|---|---|
| Artefact | 0.51 | 0.63 | +0.12 |
| Time Period | 0.57 | 0.69 | +0.12 |
| Location | 0.26 | 0.66 | +0.40 |
| Context | 0.58 | 0.84 | +0.26 |
| Material | 0.54 | 0.39 | -0.15 |
| Species | n/a | 0.49 | n/a |
| Overall Micro F1 | 0.51 | 0.70 | +0.19 |

Table 5: F1 scores for entity types and overall micro F1 compared between the previous and new dataset. Species wasn't included in old dataset, so we only present the score for the new dataset.

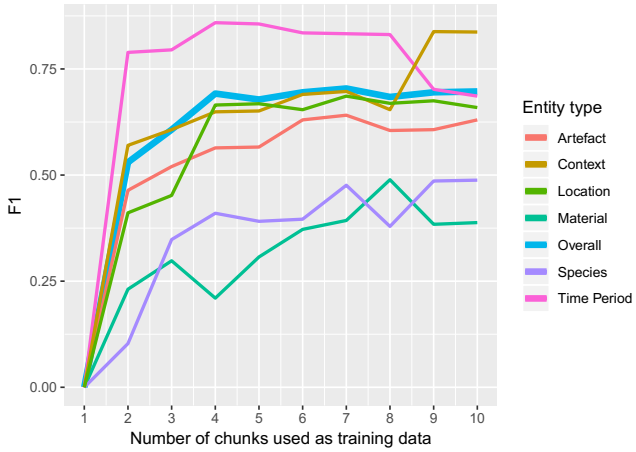


Figure 1: CRF F1 score for each entity type per 1/10th chunk of data added to the training set.

score for each model. Even though there are some fluctuations, it is evident that after adding a certain percentage of the data, the F1 scores for all the entity types plateau, even for the Material type. This probably indicates that the amount of annotations is sufficient and adding more data won't substantially increase the F1 scores, although redundancy and noise in the dataset could also potentially cause similar results. We will investigate this further in future research.

The Species category performs similarly as Material, at 0.49, this could possibly be explained by the fact that Species are written in both Dutch and Latin, but more work needs to be done to see if this is indeed the case. We also performed this analysis but instead of adding 10% of the data each time, we added a new document each time, which showed the same trend.

To see if there is another explanation for the under performance of the Material entity, we plotted a confusion matrix for all the different types, as seen in Figure 2. The diagonal

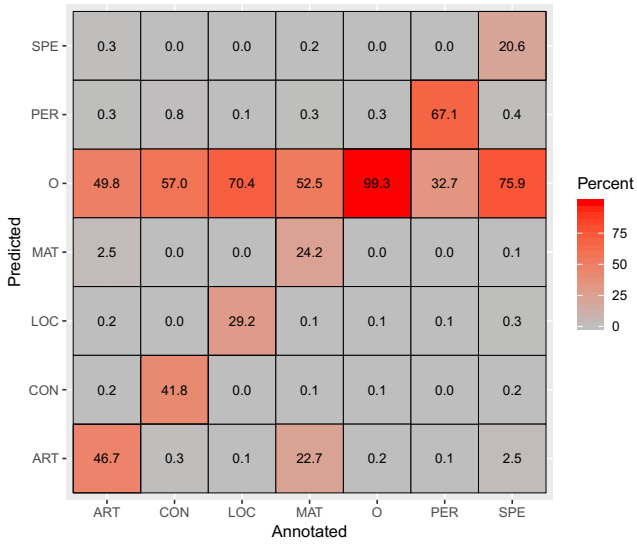


Figure 2: Confusion matrix showing percentages for each combination of predicted and annotated entity type.

and horizontal red lines are expected: the cells on the diagonal is when the algorithm predicts the correct entity, the horizontal red line is when the algorithm mistakes an entity for the O entity, the most common error in Named Entity Recognition. The only significant exception is the cell at the centre-bottom: this shows that in 22.7% of the cases, what has been annotated by humans as a Material, has been predicted by the algorithm to be an Artefact. There is also some confusion the other way around, but at a much lower rate of only 2.5%. Interestingly, from our experience supervising the annotators, this is something humans struggle with as well. The confusion is caused mainly by the words "pottery" and "flint", which depending on the context can be either a Material ("a flint axe") or an Artefact ("we found flint").

## 6. Conclusions

In this paper, we have presented a new corpus for Dutch Named Entity Recognition in the archaeology domain, annotated with six entity types. Many of the entity types are not available in standard corpora.

We trained a CRF model on the dataset, as a first experiment to assess the quality of NER with this data. The results with CRF show that using the new data substantially increases accuracy for the NER task compared to an earlier dataset. However, we only reach an F1 score of 0.70, while the IAA is 0.95. More research needs to be done to why this is the case and how we can increase the accuracy of the NER model(s).

In our current work we are using the recent advances in transfer learning to our advantage, and apply the BERT (Bidirectional Encoder Representations from Transformers) models to this task (Devlin et al., 2018). We will be using both Google's own multi-lingual model, and a model pretrained on a large Dutch corpus, to see which is more effective.

## 7. Acknowledgements

We would like to thank the Leiden University Centre for Digital Humanities for providing us with a grant to hire students to do the annotation work described in this paper.

## 8. References

Brandsen, A., Lambers, K., Verberne, S., and Wansleeben, M. (2019). User Requirement Solicitation for an Information Retrieval System Applied to Dutch Grey Literature in the Archaeology Domain. *Journal of Computer Applications in Archaeology*, 2(1):21–30, 3.

Brandsen, A. (2018). alexbrandsen/archaeo-CRF: Version 0.1 of Archaeo CRF. *Zenodo Repository*, 5.

Brandsen, A. (2019). alexbrandsen/dutch-archaeo-NER-dataset: First version. *Zenodo Repository*, 11.

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K., and Solti, I. (2012). Building gold standard corpora for medical natural language processing tasks. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:144–153.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 10.

Evans, T. N. L. (2015). A reassessment of archaeological grey literature: Semantics and paradoxes. *Internet Archaeology*, 40, 6.

Glyph & Cog LLC. (1996). pdftotext.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. Technical report, Association for Computational Linguistics.

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Lafferty, J., Mccallum, A., Pereira, F. C. N., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Carla E Brodley et al., editors, *Proc. 18th International Conf. on Machine Learning*, pages 282–289, San Fransisco. Morgan Kaufmann Publishers Inc.

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

Nakayama, H. (2019). chakki-works/doccano: Open source text annotation tool for machine learning practitioner.

Richards, J., Tudhope, D., and Vlachidis, A. (2015). Text Mining in Archaeology: Extracting Information from Archaeological Reports. In Juan A. Barcelo et al., editors, *Mathematics and Archaeology*, pages 240–254. CRC Press, Boca Raton, 6.

Rijksdienst voor het Cultureel Erfgoed. (2019). Archeologisch onderzoek - aantal onderzoeksmeldingen | Erfgoedmonitor.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Vlachidis, A., Tudhope, D., Wansleeben, M., Azzopardi, J., Green, K., Xia, L., and Wright, H. (2017). D16.4: Final Report on Natural Language Processing / Resources / Ariadne - Ariadne. Technical report, ARIADNE.