ARIADNE

D16.4: Final Report on Natural Language Processing

Authors:

Andreas Vlachidis, USW
Douglas Tudhope, USW
Milco Wansleeben, LU
Jeremy Azzopardi, SND
Katie Green, ADS
Lei Xia, ADS
Holly Wright, ADS

SEVENTH FRAMEWORK PROGRAMME

Ariadne is funded by the European Commission's 7th Framework Programme.

Version: 1.0 *(final*) **16th January 2017**

Authors: **Andreas Vlachidis, University of South Wales, USW**

Douglas Tudhope, University of South Wales, USW

Milco Wansleeben, Universiteit Leiden, LU

Jeremy Azzopardi, SND

Katie Green, ADS

Lei Xia, ADS

Holly Wright, ADS

Quality Review **Achille Felicetti, PIN**

Versions	Nr.	Authors & changes made	Date
Draft	.2	Douglas Tudhope and Andreas Vlachidis (USW)	7.11.2016
Draft	.3	Holly Wright (ADS)	10.11.2016
Draft	.4	Douglas Tudhope, Andreas Vlachidis (USW), Jeremy Azzopardi (SND)	30.11.2016
Final	1.0	Holly Wright (ADS)	16.01.2017

The views and opinions expressed in this report are the sole responsibility of the author(s) and do not necessarily reflect the views of the European Commission.

Table of Contents

Glossary

ADS	Archaeology Data Service
Archaeotools	NLP project to create tools for archaeologists to allow archaeologists to discover, share and analyse datasets
CIDOC-CRM	The CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation
CRF	Conditional Random Field
F-Measure	Measure of accuracy calculated from recall and precision measurements
GATE	A computer architecture framework for NLP
GATEfication	The design and transformation options for translating the original resources (of SKOSified RDF files) into GATE friendly OWL-Lite structures
Gold Standard	A test set of human annotated documents describing the desirable system outcome
Grey liturature	Unpublished reports
IE	Information Extraction
JAPE	Specially developed pattern matching language for GATE
Linked Open Data	A way of publishing structured data that allows metadata to be connected and enriched
NLP	Natural Language Processing
NER	Named Entity Recognition
OBIE	Ontology Based Information Extraction
OWL-Lite	Ontologies in GATE purely support the aims of information extraction and are not stand-alone formal ontologies for logic-based purposes
Polysemy	Multiple, related meanings
RCE	Rijksdienst Cultureel Erfgoed Thesauri
RDF	Resource Description Framework
SENESCHAL	Semantic ENrichment Enabling Sustainability of arCHAeological Links
SKOS	Simple Knowledge Organization System
STAR	Semantic Technologies for Archaeological Resources
STELLAR	Semantic Technologies Enhancing Links and Linked data for Archaeological Research.
String matching	Action of matching several strings (patterns) within a larger string or text
SVM	Linear Support Vector Machine
Synonymy	Similar meanings
Text Mining	The process of deriving information from text
Training data/documents	The annotated text used to train NLP classifiers
URI	Unique Resource Identifier
USW	University of South Wales
XML	Extensible Markup Language
XSL	Microsoft Excel format

Executive Summary

This document is a deliverable (D16.4) of the ARIADNE project ("Advanced Research Infrastructure for Archaeological Dataset Networking in Europe"), which is funded under the European Community's Seventh Framework Programme. It presents the final results of the work carried out in Tasks 16.2 "Natural Language Processing (NLP)". NLP is an interdisciplinary field of computer science, linguistics and artificial intelligence that uses many different techniques to explore the interaction between human (natural) and computer languages.

The partners continued to focus on one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature", exploring both rule-based and machine learning NLP methods, the use of archaeological thesauri in NLP, and various Information Extraction (IE) methods in their own language.

USW extended their English language rule based methods using the GATE toolkit for NER (Named Entity Recognition) to Dutch and Swedish language grey literature reports, in collaboration with LU and DANS (Dutch reports) and SND (Swedish reports). This made use of glossaries and thesauri from the partners, including the Dutch Rijksdienst Cultureel Erfgoed (RCE) Thesauri. The process of importing the thesauri resources into a specific framework (GATE), and the suitability and performance of the selected resources when used for the purposes of Named Entity Recognition (NER) were analysed.

The NER techniques were focused on the general archaeological entities of Archaeological Context, Material, Physical Object (Finds), Monument, Place, and Temporal (Time Appellation). The methods proved capable of extracting CIDOC CRM element and in some case studies Getty Art and Architecture Thesaurus concepts, in addition to the native vocabularies. General archaeological NLP (GATE) pipelines for English, Dutch and Swedish have been developed. In addition experimental pipelines were developed for two exploratory thematic case studies on data integration, where the output is expressed as RDF Linked Data via a CRM based data model. An English language pipeline is available for a numismatic case study. English, Dutch and Swedish pipeline are available for a case study of item level data/NLP integration on a loose theme based around archaeological interest in wooden objects and their dating, as expressed in different kinds of datasets and reports. Both case studies have resulted in interactive demonstrators operating over the ARIADNE Linked Data Cloud. All 7 pipelines are freely available as open source ARIADNE outcomes.

The Archaeology Data Service (ADS) at the University of York continued developing a machine learning-based NLP technique which has now been integrated it into a new metadata extraction web API, which takes previously unseen English language text as input, and identifies and classifies named entities within the text. The outputs can then be used to enrich resource discovery metadata for existing and future resources. This API can be incorporated into existing interfaces and used by archaeological practitioners to automatically generate metadata related to text-based content uploaded on a per-file basis, or by using batch creation of metadata for multiple files.

This report presents the final results of the work carried out to date, and presents the issues to be addressed during the remainder of the ARIADNE Project.

1. Introduction

This document is a deliverable (D16.4) of the ARIADNE project ("Advanced Research Infrastructure for Archaeological Dataset Networking in Europe"), which is funded under the European Community's Seventh Framework Programme. It presents the final results of the work carried out in Task 16.2 "Natural Language Processing (NLP)". NLP is an interdisciplinary field of computer science, linguistics and artificial intelligence that uses many different techniques to explore the interaction between human (natural) and computer languages.

1.1 Background

In D16.2, the ARIADNE partners explored NLP with the aim of making text-based resources more discoverable and useful. The partners specifically focused on one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature".

The partners explored aspects of rule-based and machine learning approaches, the use of archaeological thesauri in NLP, and various Information Extraction (IE) methods. The rule-based work was carried out by the University of South Wales, in partnership with Leiden University, on archaeology thesauri for NLP, which applied Named Entity Recognition (NER) to the Dutch Rijksdienst Cultureel Erfgoed (RCE) Thesauri. The process of importing a subset of RCE thesauri resources into a specific framework (GATE), and the suitability and performance of the selected resources when used for the purposes of Named Entity Recognition (NER) were discussed. This revealed issues relating to the role of the RCE thesauri in NER and further development of techniques for the annotation of Dutch compound noun forms. Some of these issues are addressed in the current deliverable, including adapting the ontology resource to the requirements of the NLP task.

As reported in D16.2, University of South Wales also undertook a study for a Dutch NER pipeline, which included the results of the early pilot evaluation based on the input of a single, manually annotated document. The report also presented the results of the vocabulary transformation task from spreadsheets to RDF/XML hierarchical structures, expressed as an OWL-Lite (ontology). Observations related to the vocabulary transformation process and pipeline results, and revealed initial issues that affect vocabulary usage and focus of the NER exercise. The document was annotated with respect to the following entities; Actor, Place, Monument, Archaeological Context, Artefact, Material, Period. The NER pipeline is configured to identify the following entities: Place, Physical Thing (i.e. Monument), Physical Object (i.e. Artefact), Time Appellation (i.e. Period), Material, Context. Each entity produced differing levels of results, which in some cases were good, but others needed to be explored further for improvement.

Work was also carried out by the Archaeology Data Service (ADS) at the University of York, to develop and evaluate machine learning-based NLP techniques and integrate them into a new metadata extraction web application, which takes previously unseen English language text as input, and identifies and classifies named entities within the text. The outputs were used to enrich the resource discovery metadata for existing and future resources. The intention was to create a final application with a web-based, user friendly interface that can be used by archaeological practitioners to automatically generate metadata related to uploaded text-based content on a per-file basis or using batch creation of metadata for multiple files.

The work described in D16.2 revealed the NER module worked successfully and produced correct entities for the classes it has been trained to identify. It was useful for extracting resource discovery metadata from unstructured archaeological data, particularly grey literature reports, for resource discovery indexing, where little or no metadata currently exists. From a data management perspective however, the large quantities of entities extracted by the NER module were felt to be too large to effectively manage. A prototype annotation tool built into the web application was created

to allow users to produce more training data to better train the module. The intention of ADS was to continue to work to refine the tool, especially with regard to the interface to make it easier and more intuitive to use, exploring crowdsourcing for processing large quantities of unstructured data, improvement of the text extraction module, and development of a module to export the selected metadata in a variety of formats. Integration of the web application and techniques were explored to "tidy", group and rank the entities output from the NER module using text clustering, and generating cluster labels based on the content in respective clusters.

2 Rule-based NLP for Dutch language grey literature

2.1 Introduction

As stated in D16.2, Information Extraction (IE) is a specific NLP technique which extracts targeted information from textual content. It is a process whereby textual input is analysed to form a textual output capable of further manipulation. Rule-based IE systems consist of a pipeline of cascaded software elements that process input in successive stages. Hand-crafted rules make use of domain knowledge and vocabularies, together with domain-independent linguistic syntax, in order to negotiate semantics in context.

The employment of rule-based IE and domain vocabulary resources distinguishes this approach from supervised machine learning work, which relies on the existence and quality of training data. The absence of a training corpus coupled with the availability of a significant volume of high quality domain-specific knowledge organisation resources, such as a conceptual model, thesauri and glossaries were contributing factors to the adoption of rule-based techniques in this study. Rules invoke input from gazetteers, lexicons, dictionaries and thesauri to support the purposes of Named Entity Recognition (NER). Such word classification systems contain specific terms of predefined groups, such as names of people or organisations, week days, months etc., which can be made available to the hand-crafted rules. In addition, rule-based IE techniques exploit a range of lexical, part of speech and syntactical attributes that describe word level features, such as word case, morphological features and grammar elements that support definition of rich extraction rules, which are employed by the NER process.

Rule-based techniques have previously been employed successfully with English language archaeology reports from the ADS Grey Literature Digital Library as part of the STAR Project (in collaboration with ADS), yielding promising evaluation results[1]. This took advantage of existing archaeological vocabularies from English Heritage (EH) and proved capable of semantic enrichment of grey literature reports conforming both to archaeological thesauri and corresponding CIDOC CRM ontology classes representing archaeological entities, such as Artefacts, Features, Monuments Types and Periods. This English language NLP work has been continued within ARIADNE, reported on below in Section 4, applied to annotations of Roman coins. The GATE framework[2] used for this work is the outcome of a 20 year old project established in 1995 at the University of Sheffield with a worldwide set of users; the GATE community has been involved in a plethora of European research projects.

A major development of the rule based approach within ARIADNE has been the generalisation of the previous rule based techniques to Dutch language grey literature. This faces the challenge of a different set of vocabularies available via the Rijksdienst Cultureel Erfgoed (RCE). It also faces the issue of differences in language characteristics, for example compound noun forms.

Initial work was reported in D16.2. Building on these outcomes, further work on Dutch language NLP is discussed below.

2.2 NER Pipelines for Dutch reports

This section discusses the latest version of the Dutch NER pipeline (version 3), which addressed the short term actions following the review of the earlier NER pipeline (version 2) as discussed in Deliverable 16.2. Such limitations concerned a) vocabulary coverage and suitability for NLP of the

[1] Vlachidis A, Tudhope D. 2016. A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology, 67 (5), 1138–1152, Wiley

[2] GATE (General Architecture for Text Engineering) https://gate.ac.uk

Rijksdienst Cultureel Erfgoed (RCE) Thesauri and b) plasticity of information extraction rules to address complex scenarios of compound noun forms and negated phrases. The latest version focused on adapting and enhancing the RCE resources to NLP and in particular to the task of NER with respect to Archaeological Context (Features), Material, Physical Object (Finds), Monument, Place, and Temporal (Time Appellation) entities. Some rule modifications and corrections were also made aimed at improving the performance of the pipeline but the latest version has not addressed the issues of compound noun forms extraction and negation detection.
The details of the latest version in terms of the applied improvements and problems encountered are discussed. First, we review the earlier NER versions of the pipeline. Three pipeline versions have been developed so far. All pipelines addressed the task of NER with respect to the following entities Archaeological Context (Features), Material, Physical Object (Finds), Monument, Place, and Temporal (Time Appellation). Their difference primarily resides on the size and origin of the vocabulary resources they engage. As the NER versions progressed the rules have been improved and enhanced informed by formative evaluation.

2.2.1 NER Pipeline Version 1

The NER pipeline developed in previous work for the STAR project was Gazetteer Based in contrast to the various pipelines developed for ARIADNE which are Ontology Based. The major difference between Gazetter and Ontology based pipelines relates to the construct of the vocabulary resources in GATE. It is a rather technical issue that affects the way vocabulary resources are made available to the information extraction rules and the flexibility of those resources to accommodate semantic features, synonyms, and to support partial matches. Gate ontologies provide better control for exploiting parts of the vocabulary through transitive Parent-Child relationships and feature matching. Gazetteers on the other hand, provide a flexible matching over word tokens and are easier to construct either as flat lists or featured lists from vocabulary resources of Excel spreadsheets or Word documents. Depending on the format of vocabulary resources and the aims of an information extraction task ontology based or gazetteer based or a combination of the two resources types can be engaged by the pipeline.

The first pipeline version utilised a range of vocabulary resources that were made available in Excel spreadsheet format. The vocabularies originated from the Archis database and partly consisted of subsets of the National Thesaurus RCE. In details, the vocabularies contained Artefact types, Monument types (complex types), Materials, Archaeological Context (aka Features / grondsporen), Periods, and Place names. A set of straightforward JAPE rules exploited the resources and delivered a set of named entities via mapping vocabulary to entity types.

2.2.2 NER Pipeline Version 2

The second version of the pipeline utilised the skosified version of the RCE thesauri (available from http://www.erfgoedthesaurus.nl/). The XML versions of the skosified resources were retrieved using a dedicated API key and the resources transformed to OWL-Lite format ontology using XSL template structures. The details of the GATEfication process (i.e. transformation of the original XML resources to OWL-Lite) is discussed in D16.2. Overall, five separate thesauri structures were transformed and joined under a unified OWL-Lite structure. These are, the Archaeological Types; Complextypen (monuments), Perioden (Periods), Artefacttypen (artefacts) and the Global Thesauri; Locaties (Locations), and Materialen (Materials). There is a lack of a dedicated thesaurus for Archaeological context terms as in the grondsporen.xsl resources used in version 1. However, the vast majority of archaeological context terms (grondsporen.xsl) are contained in the Artefacttypen structure. A set of

JAPE rules is used in version 3 of the pipeline for exploiting only the parts of the structure relevant to archaeological context.

The second version of the pipeline employed a range of JAPE rules that exploited the Ontology vocabulary and delivered matches with respect to the targeted entities (Archaeological context, Artefacts, Materials, Monuments, Places, Periods). In addition, a flat gazetteer resource containing period related suffixes, such as A.D, B.C, voor Christus etc. was constructed and used in the definition of JAPE rules targeted at matching numerical dates e.g. 1200 AD, 800 v.Chr. Similarly, a set of JAPE rules was defined for matching grid references and geographic coordinates (numerical places), such as 216.518/568.889. The pipeline also experimented with partial matching of vocabulary targeted at matching compound noun form cases. The available RCE thesuari, originally available in XML format, were transformed into gazetteer listings and engaged by JAPE rules targeting partial matching of terms. This was an experimental effort aimed at matching compound noun forms, which regularly occur in Dutch and affect the performance of the NER task. The initial results were encouraging and demonstrated the matching of compound noun form can be achieved. However, partial matching also delivered a significant amount of noise that affected precision (see discussion in D16.2).

2.2.3 NER Pipeline Version 3

The third version of the Dutch NER pipeline aimed at improving a range of vocabulary issues affecting the second version, primarily relating to the RCE thesauri and secondly to the quality of the gold standard. In particular, the third version addressed the problem of the overloaded ontology class labels which had resulted from the transformation of the original XML (RCE) thesauri to ontology and noted with version 2. Version 3 automatically and manually enhanced the ontology with a number of alternative labels, spelling variations and synonyms and also addressed major issues of Gold Standard quality that unnecessarily undermined the performance of the pipeline. Minor corrections and modifications were also made in JAPE rules aiming to improve the performance of the pipeline whenever that was possible.

2.2.4 Version 3 Performance Issues and Improvements

This section discusses in detail the three main strands (Vocabulary, Gold Standard and various Rule improvements) applied in the latest pipeline. The section also reveals ways and techniques that could be employed to overcome some of the performance problems that were encountered but not addressed by the third version. It ends with discussion of further work.

Vocabulary Related

The main effort of the vocabulary improvement was offered in breaking down the overloaded vocabulary entries into individual term constituents. Supplementary efforts enhanced and modified the vocabulary with spelling variations and synonyms informed by the Gold Standard input. In addition, a new set of thesauri structures were added into the ontology for complementing the vocabulary.

The RCE thesauri were not necessarily developed with Natural Language Processing in mind and as a result contain entries that are not suitable for automatic and algorithmic term matching due to their multi-term, sometimes descriptive and verbose punctuation structure. For example the vocabulary entry *amulet/talisman* and its child entry *amulet/talisman – kruisvormig* are not suitable for NLP. It is

very unlikely that such entries will be found in natural language text having this format. Most likely either *amulet* or *talisman* will be found as individual entries and if an adjective is used, such as *kruisvormig* (cruciform) this will follow a grammatically correct syntax form (i.e. *kruisvormig amulet* instead of *amulet kruisvormig).* Therefore, entries like the above should be enhanced with labels that are closer to what is likely to appear in text rather than containing descriptive and non natural language descriptions.

The aforementioned case of overloaded labels was often found in the ArtefactTypen Thesaurus and in the ComplexType thesaurus. Other thesauri resources of material, periods and places mostly contained single term entries. A set of XSL templates was developed for breaking down the overloaded entries. The overloaded entries followed a pattern for joining synonyms and specialisations together under a single label. The forward slash (/) character joins synonyms as in the case *amulet/talisman.* The hyphen (-) character adds specialisation as in the case *amulet/talisman – kruisvormig*. The comma (,) character adds a form of periphrastic description which can be treated as an alternative label, for example *amfoor, dikwandig aardewerk* (amphora, thick walled pottery). Some rare cases use the column (:) character to add a generalisation, for example *geelwitbakkend gedraaid:beker* (yellow white baked twisted : beker).

The XSL templates incorporated the above patterns to generate the new vocabulary labels. For example :

- *amulet/talisman* → two labels (amulet, talisman)
- *amulet/talisman – kruisvormig* → two labels (kruisvormig amulet, kruisvormig talisman*)*
- *amfoor, dikwandig aardewerk* → two labels (amfoor, dikwandig aardewerk*)*
- *geelwitbakkend gedraaid:beker* → one label geelwitbakkend gedraaid beker

The employment of XSLT transformation and the automatic enhancement of the vocabulary with alternative labels undoubtedly has some trade-offs. In most cases special characters for joining synonyms and expressing specialisations or generalisations are standard across the thesauri and the transformation delivers useful alternative labels. However, there are cases that do not follow the standard use of special characters or are very verbose (eg *huisplattegrond:4-schepig - type St. Oedenrode*) . Such cases due to their complexity are not matched by the transformation templates and are ignored. There might be a possibility that the automatic transformation has delivered some erroneous or false results. However, such results will not reflect actual uses of the natural language and it is very unlikely that they will deliver any matches in the NER process.

Manual Vocabulary Enhancement

Analysis of a Gold Standard (human annotated documents) has suggested several alternative labels and synonyms that were used to enhance existing vocabulary entries. The extent of the Gold Standard is not sufficient to suggest a fully comprehensive list of synonyms and alternative labels for the range of the vocabulary terms. However, whenever possible the Gold Standard input was used to enhance existing vocabulary terms particularly with frequently used spelling variations. For example, *midden*, *laat* (mid, late) and other period related prefixes can appear with bracket as *(Midden) Mesolithicum* or with the acronym *MESOL,* instead of the original label, *Midden Mesolithicum.* Such alternative labels are applicable to many terms and not just those contained in the gold standard were added to the vocabulary. Other cases of manual vocabulary enhancement concern groups of synonyms or specialisations which are frequently used within the Gold Standard and fall under an

existing more generic vocabulary entry. For example the entry *zaad/vrucht/noot/pit* (seed/fruit/nut/kernel) together with the automatically generated labels (from splitting on the forward slash) was enhanced with the Gold Standard derived terms *graan* (grain), *vlas* (flax), *macroresten* (micro-remains), *gerst* (barley) and *tawe* (wheat). Other less important modifications concern descriptive labels, such as *overig* (other) and *onbekend* (unknown) which were renamed to <*overig*> and <*onbekend*> to refrain from matching.

Gold Standard Related

Several Gold Standard related issues that already appeared in the former versions of the pipeline have been addressed by the latest NER effort. The Gold Standard consists of 7 long (some are up to 300 pages) grey literature reports containing thousands of annotated text instances. However the large number of annotations has in some cases affected the quality of the annotations. Three separate issues relating to the gold standard definition have been identified.

1. **Missing annotations**: These are cases of valid annotations that the human annotator has clearly failed to identify. The large volume of the documents dictates a cumbersome task of manual annotation to identify where such cases of missed annotations might happen. For example *kuil* (pit) is a frequently occurring word and which in some instances might be overlooked by a human annotator. In such missing annotation cases, the NER pipeline may produce a (correct) annotation which, unfortunately, is recorded as a false positive match, ie it is not false but a true match that has not been identified by the human annotator. Such cases when possible to identify were corrected in the Gold Standard so the precision rates of the pipeline would not be penalised unnecessarily.

2. **Out of scope:** Due to potential ambiguity in the manual annotation guidelines (or misinterpretation by the annotators), the human annotators sometimes marked entities in the Gold Standard that are out of the scope of the NER task. Such cases included annotation of contemporary dates (eg 20 April 1987) and place names outside Netherlands (e.g. Belgium). Such cases were removed from the Gold Standard to avoid harming the recall rates of the pipeline.

3. **Non–considered**: These cases are different from the missing annotations on the basis that the former (non–considered) are cases that have not been overlooked (as in the case of kuil) but have not be taken into consideration as relevant terms to the manual annotation task. Most likely the manual annotation guidelines did not make clear that such entities are within the scope of the task. Such cases include terms like *gebouw (building), akker (field), weg (road), etc,* which are not necessarily monuments but are included in the monuments (complex types) thesaurus. No action was taken with regards to such cases but the terms have been identified and can be resolved in future versions of the pipeline.

4. **Material or Object**: This particular case appears to be a real problem with the semantics of language use in the archaeology domain regardless of language (the same behaviour has been observed in English and in Dutch). The problem is summarised under the notion that material finds can constitute small finds e.g. *pottery (aardewerk* in Dutch) and as such are annotated as objects (finds). To the human annotator dinstiction between the material sense and the object sense of the pottery terms may be apparent in context (taking into account the intended use of the annotations). However, this form of distinction is hard to address with computational methods.

JAPE Rule Related

JAPE rule-related issues describe cases where the performance of the pipeline is affected due to the limitations or under-performance of information extraction rules. The third version of the NER pipeline improved many cases of underspecified rules that were identified during evaluation of version 2 and also identified some new cases of under-performance that require further improvement. The following cases of JAPE rule-related issues have been addressed and identified for further improvement.

Improved Cases (JAPE rules)

1. **Period and Numerical dates**: The period suffix gazetteer list has been enhanced with entries, such as *eeuw (century)* and *e (th)* to enable matching of e.g. *8e eeuw* (8th century). It has also been enhanced with improved rules for matching range of dates e.g. *tussen 1600 en 1900* (between 1600 and 1900)
2. **Place Grid Reference**: New rules have been added to address alternative grid reference patterns which were not matched from previous version e.g. 216581 / 568889
3. **Additional Lookups:** New rules have been added for exploiting input from the newly added thesauri structures in the ontology, such as the Erfgoedthesaurus material and the Landscape elements of the Objecttypen thesaurus.
4. **Place, upper case restriction**: The restriction that any Place entity match must commence with an upper case letter has been lifted for those cases commencing with *s'*, such as *'s-Heerenberg, 's-Graveland* etc.

Further Improvement Cases (further work on JAPE rules)

1. **Compound noun forms**: Compound noun forms appear in Dutch regularly joining period terms with objects, object terms with material, material terms with archaeological contexts etc. A way forward of tackling such cases is to employ partial matching over words instead of the whole word matching. Partial matching is possible in GATE but should be planned and executed carefully due to the significant amount of noise generated and the complications for entity type assignment (i.e. will the compound form carry a single type, or two entity types).
2. **Negated entities**: The current version does not exclude the matching of negated phrases (e.g. geen vondsten). Such negated phrases might in fact be a comment in the report that *no* evidence has been found for a potential finding and thus should not be annotated (at least as a simple NER instance). A *negation detection* module would enable detection of the negated entities and annotate accordingly.
3. **Non-Single word matching**: The current pipeline imposes restrictions on the part of speech type of the matched terms, requiring all matches to be nouns except for period/time appellations. The noun validation is currently achieved using the part of speech input that is assigned on single words. Thus, two word vocabulary entries (eg *Metamorfe gesteente)* are not ignored from matching. In order to enable matching on non-single word, the noun

validation should happen over a noun phrase not just on a single word token. The English Noun Phrase module of GATE could be adapted in Dutch to enable this kind of validation.

4. **Adjective matching**: The noun validation restriction excluded adjectives from matching. However, the Gold Standard contains many material entities of adjectival form, such as *bronzen* (bronze), *stenen* (stone) etc. The restriction can be easily lifted but careful planning is required in order to conclude such cases either as individual material entities or as moderators of object or monument entities.

5. **Place names as organisations or surnames**: Several false positive matches of place names occur as part of an organisation's name or surname. In order to improve matching on such cases, actor and organisation entities should be identified first, in order to exclude them matching as Places. The Actors Organizations and Actors Person thesauri have already added in the ontology. A future version of the pipeline could address such entities and improve matching over place names.

6. **Erroneous Part of Speech and Tokenizer input**. JAPE rules and vocabulary Lookup matching relies on input from the POS and Tokenizer modules. In the cases where such input is erroneous (eg a verb tagged as noun, a wrong word root) Lookup matching and JAPE rules deliver erroneous matches. However, such cases are few and do not significantly affect the performance of the pipeline.

2.2.5 Longer term actions (requiring input from Dutch archaeology domain experts)

- Identify the most frequent cases of terms that contribute to compound noun forms. It will not be efficient to produce part-matches via gazetteer from the totality of the "Archeologische artefacttypen" thesaurus, as this will have an impact on precision (generating too many part matches). Instead a selected set of terms that frequently appear in compound forms should be identified and exposed as gazetteer list. For example "aardewerk" has much more chance of appearing as a compound noun than other terms, so it should be prioritised for part-matching.
- Identify entity-type combinations that deliver compound noun forms. Based on the GS results, it appears there are three main combination types a) material+artefact, b) period+artefact and c) artefact+artefact. It would be helpful to discuss these combination forms with Dutch archaeologists before resolving on any NLP matching rules.
- In addition to the above, the annotation approach towards compound entity forms should be discussed and finalised. At this stage it is not clear the number and type of annotations that should be delivered from a compound entity form. For example consider the case of "aardewerkfragment" (pottery fragment). Will it deliver:
    - a single span annotation (aardewerkfragment) associated with two SKOS references one for "aardewerk" and another for "fragment";
    - two separate annotations each associated with a SKOS reference;
    - three annotations, two separate annotations (as above) and a third for the whole span annotated as "P45.consists_of" property.
- Similarly, some thought should be given towards the annotation of compound entity forms of part-known constituents, where only one of the parts is "known" to the ontology. For

example “aardewerkmagering”, or how “magering”, which is not a known (available) term, will be treated. Will it just be ignored and only a single annotation will be delivered i.e. “aardewerk”, or will it be included in the annotation span (which is also possible).

- 'Expert annotator' review of the existing GS for consistency and in light of the automatic output results.
- Actions concerning adding new thesauri concepts, and releasing respective SKOS references.
- Rearranging a thesaurus structure for adding new broader terms for a set of specialised terms already included in the resource e.g. “Nederzetting” (settlement).
- With regards to the above, a quick fix for NLP purposes which would not require restructuring the resource, could be adding an alternative label of the broad term to the existing specialised terms e.g. “Nederzetting met stedelijk karakter”. Or to use a general purpose thesaurus that contains the broad term (Nederzetting), such as Erfgoedthesaurus Objecttypen for delivering matches with SKOS reference respective to the broad term not to a specialised term.

3 Rule-based NLP for Swedish archaeological reports

Subsequent to the Dutch NER work, the Swedish National Data Service (SND) and USW collaborated on an initial investigation of the generalisation of the (USW) English language rule-based approach to NER on Swedish language archaeological reports. USW contributed on the technical side and SND contributed with archaeological reports, vocabularies, mappings to AAT, manual annotations and evaluation of the NLP outputs, amongst other work. Due to the limited time available, this is an exploratory investigation intended to lay the groundwork for further development.

3.1 Swedish language pilot general NLP pipeline

SND collaborated with USW in the creation of an NLP general pipeline as part of the metadata enrichment effort by contributing to the Swedish-language version of the general pipeline (GP). The steps involved were:

1) Initial manual annotations of first run – Nine archaeological reports in Swedish were selected from SND's corpus of published studies. These nine reports were annotated by a group of three archaeologists and data managers. Annotation consisted of marking keywords within each text by the categories used for the Dutch NLP work, following similar Instructions for Annotators as for the Dutch NER work (see Appendix 1):
    a. Context
    b. Material
    c. Monument
    d. Object
    e. Place
    f. Time appellations

2) Vocabularies – In parallel with manual annotation, vocabularies for each of the abovementioned categories were also produced. These Swedish-language vocabularies were based on SND's own vocabularies, which are used for metadata within its own online data catalogue.

Vocabulary mapping: In order to aid the mapping of keywords within the ARIADNE portal, the Swedish language vocabularies were revised and mapped to Getty's Arts and Architecture Thesaurus (AAT). This mapping was done using SKOS Mapping Properties to facilitate semantic web integration and metadata enrichment within the ARIADNE registry and portal. This is described in more detail in ARIADNE D15.1 - Report on Thesauri and Taxonomies.

The results of this effort were useful for evaluating the use of the NLP general pipeline for metadata enrichment, and produced a number of recommendations for future improvements. Future work includes exploration of an expanded general vocabulary for Swedish archaeology. This would be a useful resource for metadata enrichment, both automated and otherwise.

To provide a benchmark, SND annotated 9 documents following the Instructions for Annotators. These were archaeological reports from four different archaeological investigators (one county museum, one private company, one university funded organization, and the National Heritage Board's contract archaeology division), and from different phases of the archaeological process (desk-based assessment, field evaluation and final excavation). These were accompanied by the relevant SND vocabularies. One issue for future work flagged up in the manual annotation process is that the Instructions assume a single entity for any given word. It is not possible to annotate the texts in "layers", i.e. put one single word in multiple categories (e.g. "Vråkeramik", "Vrå" being both a place name and a pottery type, as well as a local term for the time period that covers the Pitted Ware culture, etc…).

A brief note on the vocabularies:

Places - There are two separate lists for the geographical information (Places). One is based on the modern administrative use of counties (Sw. län), municipalities (Sw. kommun) and parishes (Sw. församling), the other contains the names of historical/traditional provinces (landskap) and parishes (socken[3]). Only the first was used in the pilot pipeline.

Time Appellations - Time periods were provided along with some of the most common local sub-divisions of archaeological periods, as well as inflections. This list of geographical/local variations is not comprehensive.

Objects and Materials - The lists for Archaeological Objects and Materials are taken from the Swedish History Museum. They are not spellchecked and considered as controlled vocabularies.

Monuments - The Monuments vocabulary is the one used by the Swedish National Heritage Board, with some additional less common terms and architectural features, as well as inflections.

Context Types - This is a list created by the SND during the work of annotation - it is not comprehensive and is not controlled by any archaeological authority. In Swedish archaeology, the term feature is usually used, rather than context (except for urban and/or historical archaeology (ca. 1100-1800 AD), where the Single Context method is more commonly used).

Following the availability of the annotated benchmark corpus, USW built a pilot Swedish archaeological NER Pipeline targeting the six entity types (Context, Object, Material, Monument, Place, Time Appellation). The pipeline uses the OPEN NLP Tokeniser, Sentence Splitter, Part of Speech tagger for Swedish and a Gazetteer created from the vocabulary files provided.

This was evaluated against the benchmark following the standard GATE evaluation procedure and precision/recall metrics for the 9 documents were produced together with the annotations in context of the original document. The results were promising considering the early stage of the pipeline and the underpinning vocabularies.

The evaluation results were analysed by SND and also USW from a technical perspective according to the standard NLP evaluation dimensions. The evaluation is discussed below, together with suggested points for future work:

- Missed terms by NLP: The biggest miss was due to the lack of term variations, such as singular/plural, definite/indefinite, possessive, and compound terms. Examples include

Vocabulary term	Grey lit. term	Grey lit. form
Nedgrävning	nedgrävningar	Plural
Stolpe	Stolpar/stolppar	Plural(misspelled plural)

[3] Socken is an archaic name for the original rural church parishes, “kyrk-socken”. It also describes a secular area, a sockenkommun ("rural area locality") or a taxation area, a jordbokssocken. The socken system was in many ways the predecessor to modern municipalities. In 1862, the socken parishes in Sweden were abolished as administrative areas during municipality reforms. The jordbrukssocken term ("taxation area") remained in use until the "Reform for registration of real property" 1976–1995 was complete. No further alterations to the socken names or borders were made after this. Even though the term socken is no longer in use administratively, it is still used for cataloging and registering events, artefacts and archives within the research fields of history, archaeology, botany, and history of languages (such as toponymy and dialect research).

Lager	lagret	Definite
störhål	störhålets	Possessive
kulturlager	kulturlagerrest	Compound
stolphål	stolphålsbotten	compound
störhål	stör	stem

This confirmed a known limitation of the initial pilot pipeline, which did not attempt to match term variations.

- Missing vocabulary elements: Several missed terms were due to incomplete initial vocabulary coverage, especially for Monuments and Objects and some temporal terms.

- Compound terms are missed and thus pose challenges for recall: This is a similar issue to that encountered when working on the Dutch grey-literature. An example of a compound term is 'fornlämningsområdet'. Configuring the system for partial (part of word) matching could support matching of compound terms. However, this would bring a risk of noise and false positive matching. Rather than enabling partial matching on the whole range of vocabulary terms, partial matching could be restricted to a small sub-set of commonly occurring term in as compound parts. This might enable the capture a majority of compound cases without introducing too much noise.

- False Positives: Manual annotation is not consistent in some cases: some (not all) of the false positive results are due to inconsistency in the manual annotation (as encountered with other languages). One particular case is 'Torv' or peat. In this article, 'torv' is only present as a material within a deposit. However, it is used for C14 analyses. It is possible that it was not marked because it did not appear to be of direct interest. This could be the case for other terms. In future work, a revised and more specific set of Instructions for Annotators could be considered for the Swedish context with a more specific description of the entities to be annotated for Swedish archaeological practice.

  In other false positives, mostly not due to problems with manual annotation, some terms are very generic, e.g.: Vad, Väg, Byggnad, Hus, Struktur. 'Vad' also has multiple meanings, including 'what' and 'ford'. A better context awareness might help to catch false positives. In some cases the number of false positives could be reduced by looking at context marker e.g. Hydda could be counted as a monument unless the words 'I' or 'en' are present before the term itself. Another example is hög – it can mean 'mound'/'heap' or 'high'. If it has measurements such as 0,2m before it, these indicate height.

  More generally, context could be used as indicators of when a general term is being used in an archaeological sense; in/definite articles in from of some monument terms tend to indicate non-archaeological remains, while terms without definite articles are used for archaeological features - e.g. vägsträcking, väg. This is somewhat speculative currently but could be a topic for future work.

- Revised user guidelines: Some false positives in the NLP outcomes are the results of occasional lack of adherence to the guidelines in manual annotation, which can be improved with practice, so to speak. In some cases, the entities in the guidelines (and in the NER exercise generally) could be clarified for the Swedish context, eg does 'Monument (Complex) Types' indicate that it is (part of) a complex? It would be useful to create specific category descriptions for the Swedish situation (see below on Monument and Context annotation).

- Context entities performed well, although Monuments had low recall (more vocabulary needed)

- Material had good Recall, less good Precision: As with English and Dutch, there is the problem of ambiguity in whether a term is treated as Material vs Object sense. The problem includes the terms *Keramik, kvarts, flinta, kol, bränd lera, tegel, tall, ben, skärvsten, malm, porslin.* The term *'Kol'* could be removed from the objects vocabulary and replaced with *'kolbit'* – charcoal pieces, to avoid conflict with the material vocabulary. Again more context aware NER would also help. This can include adjacent terms and consideration of singular/plural. If a number/quantifier is present before the term (e.g. 'ben', 'sten'), this usually indicates objects rather than materials. Another indicator is when adjectives before the term are in the plural form (e.g. 'brända ben'), which also indicates quantity and thus (usually) objects. Articles are also another indicator of objects rather than materials.

  Additionally, for the use case for a given NLP pipeline should be considered; is the distinction between material and object actually relevant for the main intended uses?

- Many NLP outcomes marked as *partially correct* terms are in fact more specific instances of existing vocabulary terms and are thus correct.

- Monument and Context annotation: the definition of these categories could be clarified as there may be some overlap – **Anläggningar** (equated with archaeological context in this exercise) means a structure, building, installation, something which was constructed. **Archaeological context**, on the other hand, is tightly linked to a stratigraphic event, and may or may not include structures. Thus, a soil deposit is an archaeological context but not an anläggning, while a rubble wall foundation can be both an archaeological context and an anläggning. A stone oven may or may not be an archaeological context in itself, but it is an anläggning. An execution site is not and archaeological context but it is an anläggning. The distinction between these entities should be revisited for Swedish archaeology with updated vocabularies - it may be that the issue revolves around the treatment of grouping and phasing interpretation for NLP purposes versus the previous identification of stratigraphic contexts.

- The context vocabulary list needs to be reviewed for words such as 'långsida', 'gavel' and other architectural components, which should either be completely removed or else moved to the monuments/objects lists.

- Placenames - some placenames, like 'Mark', 'Ny' and 'Vara' are also very common words, and could thus lead to low precision. Recall seems to be low mostly due to many low-level or non-administrative place names present in the texts.

- Dating - Numerical dates should be catered for. A preliminary glossary of date context markers can be found in Appendix 2:

- Negation detection would be a useful addition.

- Tables pose challenges and merit a specific NLP module.

- Abbreviations and dating terminology (including ± symbol denoting C14 dates) would be useful additions to the vocabularies. Geological periods, minor place names, informal regional names, and pottery phase/typology names may be interesting vocabularies to add in the future.

3.2 Swedish language revised general NLP pipeline

Taking account of the evaluation of the pilot Swedish NLP system by SND and USW, a revised Swedish archaeological NER pipeline was produced with improved matching on term variation (a key issue brought up by the evaluation). This makes use of a stemmer[4] (morphological analyser) in order to address the pilot system's limitations on term variation discussed above. The stemmer enables matching on word root input rather than on the whole string, allowing matching of singular/plural and other term variations from a single vocabulary entry. Although the quality of the stemmer dictates the quality of term variation matching (with scope for some loss of recall), it is preferred, in terms of time scale and final result, to employ a stemmer than enriching a vocabulary with each term's variations. This general Swedish NLP pipeline is available as an ARIADNE outcome along with the other pipelines described here.

[4] http://snowball.tartarus.org/algorithms/swedish/stemmer.html

4 Rule-based NLP investigations on specific case studies

Two additional case studies conducted on specific application areas are briefly reported here.

4.1 Numismatic case study

Natural Language Processing techniques were employed by USW (Hypermedia Research Group) to extract numismatic information from a sample set of six English language reports from the ADS Grey Literature library to demonstrate the potential of NLP in data integration. The resulting data was expressed in the same CIDOC CRM, AAT and Nomisma form used for the numismatic item level integration case study investigated as part of WP14. An extract from this resulting CRM based RDF was integrated into the FORTH-ICS case study demonstrator and it was found that the NLP techniques had identified items from the report text not explicitly mentioned in the site record metadata. See ARIADNE D15.2[5] for a discussion of the item level investigation. The NLP techniques were slightly adapted from the information extraction pipeline used in the STAR Project's OPTIMA toolkit[6], including some grammatical patterns for Relation Extraction. This is capable of extracting 'rich phrases' combining CRM semantic entities, such as '*medieval silver coin*', '*late Roman coins*', '*coins dating to AD 350–53*', '*coins belonged to the second half of the 3rd century AD*'. The Nomisma with its numismatic vocabulary including coin denomination, was part of the information extraction, eg yielding '*radiate of Tetricus I or II dating to around the AD 270s*' (NER of Emperors was not included in this exercise but that would form one of the next priorities to incorporate).

4.2 Data/NLP multilingual case study on item level data integration

As a final integrative task within WP16, it was decided to investigate a specific case study of item level data/NLP integration with NLP output expressed as RDF and made available for exploration in an interactive demonstrator. Inspired by the DANS work on dendrochronological analysis, a loose theme to organise the study was chosen based around archaeological interest in wooden objects and their dating, as expressed in different kinds of datasets and reports. Accordingly, the NLP was focused on concepts relevant to this theme, such as samples, materials, objects and temporal information, together with their connections. The work was undertaken by USW on the technical side, in collaboration with DANS and SND as regards Dutch and Swedish archaeological datasets, reports and vocabularies. This is very much an exploratory prototype (with limited resources), one not intended to reveal new scientific findings but rather to show future possibilities of the semantic techniques for larger scale efforts on multilingual integration of datasets with reports. Grey literature reports are a vast but under-utilised resource which can be used together with datasets where they exist for meta research and large scale studies. NLP has the potential to extract more information from the reports than can be found in the metadata alone.

The multilingual demonstrator aims to investigate the potential for NLP information extraction techniques to achieve a degree of semantic interoperability between archaeological datasets and the textual content of grey literature reports. The case study has a broad theme relating to wooden material including shipwrecks, with a focus on indications of types of wooden material, samples taken, wooden objects with dating from dendrochronological analysis, etc.

The resources comprise English and Dutch language datasets and grey literature reports, together with Swedish archaeological reports. The ADS Grey literature archives were searched for reports relating to "dendrochronology" and 11 documents were retrieved. DANS provided a sample of 9

[5] ARIADNE D15.2 forthcoming at http://www.ariadne-infrastructure.eu/Resources

[6] Vlachidis A, Tudhope D. 2016. A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology, 67 (5), 1138–1152, Wiley

Dutch reports and SND provided 5 Swedish reports based on a focus on wood material and dendrochronological analysis. ADS datasets included two shipwreck datasets (Newport Medieval Ship and the Flower of Ugie) and the Vernacular Architecture Group database. DANS facilitated an extract from the database of the international Digital Collaboratory for Cultural Dendrochronology (DCCD[7]). A CRM based data model was designed to connect the data elements and the NLP entities, which include Object, Sample, Material, Place (in some cases), date ranges. A spine vocabulary was identified from the AAT hierarchies for material and objects. Corresponding Dutch terms for the AAT concepts mostly existed already from WP15 mapping work, while Swedish terms for the AAT concepts were produced by SND, as part of their WP16 effort.

An extract of relevant sections from the Swedish reports was produced manually for the case study. The Dutch pipeline explored the potential for automatic detection of dendrochronology related sections based on a bespoke glossary. The Dutch pipeline has a component that detects sections relevant to the case study. While this was a simple technique based on a fixed number of sentences surrounding a glossary lookup, it proved sufficient for the exploratory case study. More elaborate versions would involve additional rules and pattern detection. In future work, the automatic detection of sections to emphasise for NLP or conversely to avoid would be a useful component, in light of the length of some archaeological reports.

Following formative evaluation, certain English and Dutch terms were excluded from matching (ie acting as 'stop words') due to their high potential for producing false positives. Polysemous Swedish terms, such as *lager*, might also be good candidates for stop words, due to their ambiguity. As with the general NLP pipelines, improvements to the Dutch and Swedish Part of Speech taggers and Stemmers would be an immediate focus in future work, together with an improved glossary of date indicators and context markers. For example, context markers for single years (as opposed to other instances of integers) would be very valuable. Another priority would be a careful manual annotation set to drive systematic evaluation, which would also require an iterative approach to refining the guidelines for manual annotation. Vocabularies need be improved using terms from a larger corpus than the one used in this limited study. Resources such as glossaries, word lists, trade/thematic lexicons, and other such resources could be used to enlarge the vocabularies being used. Such enlargements would need to be evaluated against manually annotated texts so that precision is not negatively affected.

As with the general case, ambiguity between material and object senses proved challenging in some cases (for *both* machine NLP and human annotation). For example, in the Swedish reports, it was sometimes difficult to distinguish between a tree (e.g. *tall,* or *pinetree*) and material made from pine. If the distinction is considered important, it would be useful to make use of context markers in the NER, which might allow for an improvement in precision. The term 'tall' by itself can be ambiguous due to the interchangeable way that materials and objects are discussed in archaeological report. However, word sense disambiguation techniques could be used to resolve such ambiguities as, for example, in the phrase *'av tall'*, or *'of pine'*, where the term is more likely to indicate material than object. In addition, the definite form of a tree name (eg *tallen – the pine)* seems to be often used in Swedish to connote the material used, rather than a specific tree.

Future efforts will have the choice to focus on either thematic NLP (as in this case) or more generic archaeological NLP (or both cases). Following a theme has been useful in this exercise, as it helped contain the problem to a specific theme within archaeology. Such smaller themes could be developed by smaller groups, and could facilitate subsequent efforts made to create a more encompassing tool. A relatively broad theme, as in this exercise, arguably poses more challenges than more concrete topics, such as the exercise on coins reported in the previous section.

[7] DCCD database - http://dendro.dans.knaw.nl

Experimental NER pipelines were developed for the above entities and vocabulary in English, Dutch, Swedish, building on the work for the general NLP pipelines described above. The atomic entities resulting from the NER were combined into CRM properties, where considered appropriate. The work is still at an early stage and results are preliminary with need of further refinement to reduce false positives and extend the vocabularies used. More work is also needed on Relation Extraction (RE) algorithms that assert CRM properties between connections. The English language NLP output is based on grammatical patterns for Relation Extraction, building on previous work for the STAR project[8]. For the Dutch and Swedish reports, simpler NER techniques are used that do not attempt connections between entities extracted (other than occurrence within the same sentence). A priority in future work is to apply a pattern-based information extraction approach to Dutch and Swedish reports similar to the English language work. Improved NLP pipelines could be substituted into the data integration workflow developed for the case study.

Illustrative examples of the various NLP output (with colour coding indicating the semantic entities identified) include the following:

The calculation of the common felling period for each dated timber from this floor suggests a construction date between AD 1682 and c AD 1699.

The felling date of AD 704/5 identified for these timbers indicates that the structure was in use during the early eighth century.

This sample has a calculated felling date range of AD 1609 to AD 1645.

Two timbers dated from the west wing roof produce felling dates in the winter of AD 1735/6 and the spring of AD 1736.

The results identified that one board was datable by tree-ring dating techniques, with this board felled in either the late-sixteenth century or early seventeenth century.

Many of the oak boards appeared from external examination to be timbers suitable for analysis;

Dendrochronologisch onderzoek door Stichting RING in Amersfoort wijst uit dat de eik waaruit de paal is vervaardigd, is geveld tussen 55 en 69 na Chr.

De dateringen op basis van dendrochronolo- gisch onderzoek van het hout uit de sporen 6 en 9 wijzen uit dat een eventuele de reparatie voor 62 na Chr.

Wel valt op dat het aantal dateerbare eiken toeneemt naarmate we dichter in de buurt van de 4600-4550 BC komen.

Een van de paal genomen dendro-monster leverde een kapdatum van 1516 ± 6.38

Två prover togs från åtelpålen och kunde genom en dendrokronologisk analys dateras till 1730-tal.

Samtliga prov dateras och täcker den mest exakta dateringen vinterhalvåret 1677/78.

Prov 1 som var bearbetat virke av ek daterades till fällningsår vinterhalvåren 1536/37.

Den större fartygslämningen , daterad till tidigt 1800-tal har troligen varit en skuta för kustseglation eller fiske.

[8] Vlachidis A, Tudhope D. 2016. A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology, 67 (5), 1138–1152, Wiley

The aim of this exploratory case study was to investigate the potential of the various semantic techniques employed in a multilingual demonstrator. The results suggest that the approach is worth continued investigation. The NLP techniques were able to generate CRM/AAT based output from English, Dutch and Swedish texts in the same format as the instance data extracted and mapped to the CRM/AAT. Of course, the NLP derived RDF statements do not carry the same degree of reliability as those derived from the datasets and false positives can be found. In future work, some indication of the provenance (and hence reliability) of the RDF data could be included in the CRM model. Nonetheless, the principle of semantic data integration from text documents and databases has been demonstrated.

The output from GATE is expressed as RDF using the CIDOC CRM model with connections also made to the AAT. This was achieved by means of a new mapping/extraction tool, STELETO, developed by USW for ARIADNE and available as open source[9]. STELETO is a 'lite' cross platform version of the STELLAR.CONSOLE application developed for the STELLAR Project[10][11] - a simpler version of STELLAR with an improved command line functionality. STELETO is a general delimited text data conversion utility that can be used for mapping and extracting instance data to the CRM via templates that hide the complexity of the CRM model. It was used for the CRM based RDF for both the datasets and text reports.

The multilingual demonstrator cross-searches over the datasets and text reports via SPARQL queries. Output is expressed as RDF using essentially the same CIDOC CRM model as used for the Coins Demonstrator with mappings made to the AAT. The outcome is a pilot demonstrator of the technical possibilities, operating over a Linked Data expression of the output, which offers cross search over both the datasets and text reports via an interactive, browser based SPARQL query builder. It demonstrates the potential for alternative user interfaces to a plain SPARQL endpoint building on the 'widget' techniques developed in the SENESCHAL project[12]. The work is ongoing and will be reported in the forthcoming ARIADNE deliverable D15.3 Report on Semantic Annotation and Linking[13].

4.3 NLP pipelines made available for further work

USW developed three separate general archaeology Named Entity Recognition pipelines for English, Dutch, and Swedish languages using the GATE platform. The pipelines are rule-based and driven by domain vocabulary expressed as OWL-Lite ontologies or flat gazetteer lists as in the case of Swedish pipeline. The vocabulary of the English pipeline originates from the Heritage Data vocabularies[14], the Dutch vocabulary from Erfgoed Thesaurus[15], and the Swedish from SND (in house resources). All three pipelines focus on the recognition of the following entities; Archaeological context (i.e. post-hole, ditch etc), Physical Objects, Materials, Temporal (as in Periods and as in Numerical Dates but not contemporary dates) and Monument types. In the case of Dutch and Swedish reports, Placenames and Grid references are also addressed.

In addition, as described above, experimental English, Dutch and Swedish language pipelines for the data/NLP data integration case study on the wood / dendrochronology theme were developed,

[9] STELETO https://github.com/cbinding/STELETO

[10] STELLAR Project http://hypermedia.research.southwales.ac.uk/kos/stellar/

[11] Binding C., Charno M., Jeffrey S., May K., Tudhope D.: Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data. International Journal on Semantic Web and Information Systems, 11(1), 1-29. IGI Global.

[12] SENESCHAL Project http://hypermedia.research.southwales.ac.uk/kos/SENESCHAL/

[13] ARIADNE D15.3 forthcoming at http://www.ariadne-infrastructure.eu/Resources

[14] http://www.heritagedata.org/blog/vocabularies-provided/

[15] http://www.erfgoedthesaurus.nl/

together with an English language pipeline for the numismatic data integration case study. All the NLP pipelines are freely available as open source ARIADNE outcomes of WP16[16].

The plan for future work is to use the AAT vocabulary as a spine for cross searching among different languages. An initial study on the Material entity showed that the AAT coverage for this particular entity type for English and Dutch is around 80%. A sub-set of the Swedish vocabulary has been manually mapped to AAT concepts, as part of the work for WP16.

[16] English, Dutch, Swedish rule-based NLP pipelines
https://github.com/avlachid/Multilingual-NLP-for-Archaeological-Reports-Ariadne-Infrastructure

5 Machine Learning API for the ADS Grey Literature Library

5.1 Introduction

As reported in D16.2, the ADS has built upon the work and lessons learned from the Archaeotools project, to further develop NLP tools and help the archaeological domain better access the vast resource of unstructured digital data available to archaeologists in the form of text. This text typically exists in PDF, MS Word, or plain text files within the ADS Library of Unpublished Fieldwork reports (also known as the Grey Literature Library), digitised journal collections, and reports deposited within project archives.

Training data developed by Archaeotools was applied to a classifier. A classifier is a machine learning tool that takes data items and places them into classes resulting in a statistical model, which is used to extract entities from entered text. After an evaluation of classifiers, the CRF classifier was chosen, as it was easier to implement into the web application and required less computing time to produce results. The models built by the classifier with gazetteers were then directly applied to the unseen data from grey literature reports. As there is currently no Gold Standard for archaeological grey literature, a group of reports from the North Yorkshire region (knowing there had not been previous training on grey literature from a North Yorkshire dataset) were chosen and manually scored. The gazetteers were especially useful for improving extraction performance, when applied to more unseen corpora. This confirmed there is substantial overlap of information from various corpora within the grey literature. To train the CRF classifier, a window size of five surrounding tokens and the following feature set was used:

- N-Grams with max length of six tokens (i.e. contiguous sequence of words)
- Exact token string
- Features from previous word class sequence
- Archaeological Gazetteer

A prototype web application interface was developed for testing and demonstration purposes and also reported in D16.2. The prototype allowed domain experts to annotate reports, generate resource discovery metadata where none exists, and generate metadata which can be used to further train the classifiers. The application was designed to allow text to be entered into an “input text area”, or a file (PDF or DOC) to be uploaded to the application. When using the latter option the prototype extracted text out of a PDF or DOC file automatically, and displayed it in the ‘input text area’. While only a prototype, the interface showed how the API might be visualised if implemented in an existing interface, which may be useful for producing more training data in the future, as it allowed users to correct results which can then be used by the training classifier.

To extract the possible metadata from the uploaded documents, an NER module was created and the prototype was built as a simple Java application written to utilised the CRF classifier. When text was entered into the “input text area” entities were extracted from the text using the NER module based on the CRF classifier. The extracted entities were displayed as suggested metadata to the right of the entered text, and users can assess the relevance of the extracted entities. The web application also detected and extracted UK grid references using manually crafted regular expressions. Extracted grid references were automatically verified using UK Geospatial data held within an Oracle Spatial database, where incorrect grid references can be filtered out from the result. By clicking on the magnifying glass icons beside each entity generated, users could jump directly to the word in the text from which the result was derived.

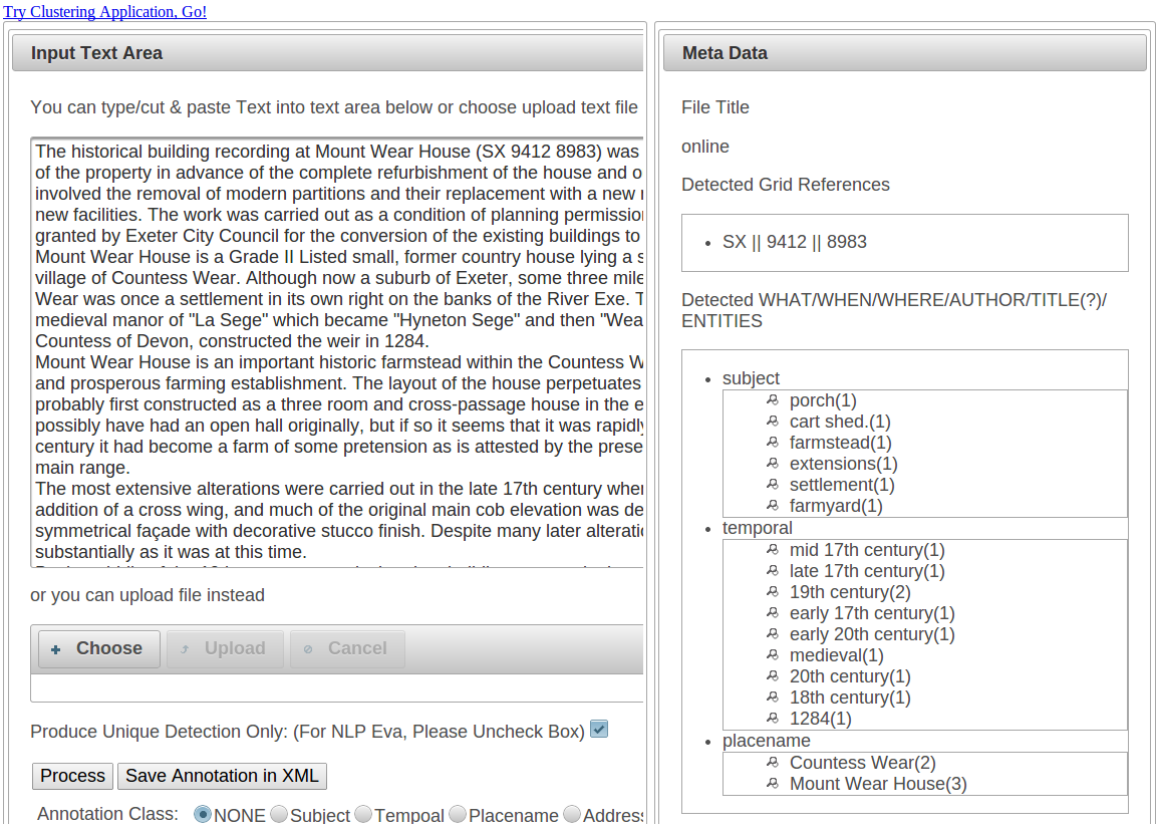

Figure 1: Screen shot of the prototype ADS web application showing entities extracted from the text.

The entities extracted by the NER module using this method (using a relatively short piece of text), specifically composed to provide an introductory overview of an archive), produced very successful results, and the relatively small number of entities were easy to view and manage within the web application by a user, although this became more complicated when tested with a larger body of text. For a full analysis and examples of the entities extracted using the NER module, please see D16.2.

5.2 Named Entity Recognition (NER) Web Service

In D16.2, it was stated ADS planned to continue development of the web application, but after additional consideration it was decided further refinement of the interface would be less useful than the creation of an NER Web Service API, which could be made freely available to the archaeological domain. Subsequently, development time was focussed on creating and refining an API that allows users to submit NER tasks to an ADS server, which then returns a set of terms, including their category and offsets, which developers can incorporate into their existing interfaces.

The API follows common practice for a RESTful HTTP web service. Users submit a task and clients POST JSON to an API endpoint. If successful, it will return HTTP status 200, and return JSON in the response. Depending on the complexity of the task and length of the content, the API may return the result asynchronously, in which case the results are not immediately available, and a delay must be implemented on the developer's end after submitting a task.

API endpoint and supported methods

All API URLs are relative to the root endpoint on the ADS server:

- POST http://ads.ac.uk/nlp/api/nse/analysis -- for simple text analysis synchronous operation, returns a set of named entities

- POST http://ads.ac.uk/nlp/api/nse/asyncAnalysis -- for lengthy analysis and asynchronous operation, returns a set of named entities

The API provides a basic HTML interface:

http://ads.ac.uk/nlp/demo.jsf

If an endpoint supports POST, it can be opened in a web browser, and raw JSON data can be submitted. The demo program implemented previously is now using these services.

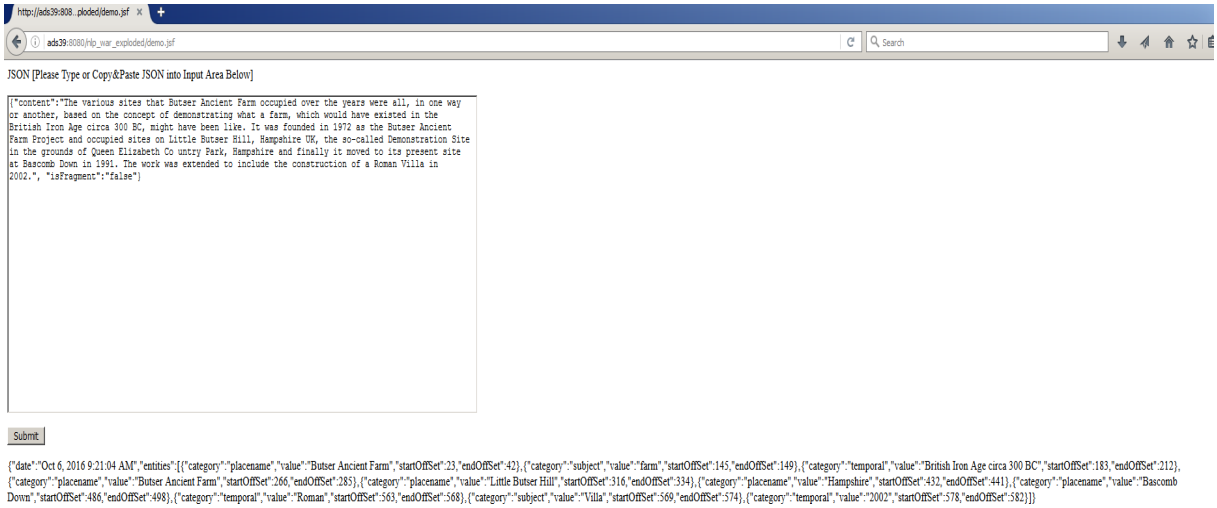

Figure 2: Screenshot of the simple API HTML interface.

Create a task

Below is a simple example of a test JSON request. It includes two parameters, both are mandatory.

content	string	Text content to be analysed
isFragment	boolean	Inform service to add string matching operation for extremely short content where context is too short to provide useful information for NER task

Task.json

{

"Content":

"The various sites that Butser Ancient Farm occupied over the years were all, in one way or another, based on the concept of demonstrating what a farm, which would have existed in the British Iron Age circa 300 BC, might have been like. It was founded in 1972 as the Butser Ancient Farm Project and occupied sites on Little Butser Hill, Hampshire UK, the so-called Demonstration Site in the grounds of Queen Elizabeth Country Park, Hampshire and finally it moved to its present site at Bascomb Down in 1991. The work was extended to include the construction of a Roman Villa in 2002.",

"isFragment":"false"

}

The following command can test the service:

curl -H "Content-Type: application/json" -X POST -d '{json}'
http://ads.ac.uk//nlp/api/nse/asyncAnalysis

Sample Response:

{"Entities":

{"term":[{"category":"placename","value":"Butser Ancient
Farm","startOffSet":23,"endOffSet":42},{"category":"subject","value":"farm","startOffSet":145,"end
OffSet":149},{"category":"temporal","value":"British Iron Age circa 300
BC","startOffSet":183,"endOffSet":212},{"category":"placename","value":"Butser Ancient
Farm","startOffSet":266,"endOffSet":285},{"category":"placename","value":"Little Butser
Hill","startOffSet":316,"endOffSet":334},{"category":"placename","value":"Hampshire","startOffSet":
431,"endOffSet":440},{"category":"placename","value":"Bascomb
Down","startOffSet":485,"endOffSet":497},{"category":"temporal","value":"Roman","startOffSet":56
2,"endOffSet":567},{"category":"subject","value":"Villa","startOffSet":568,"endOffSet":573},{"catego
ry":"temporal","value":"2002","startOffSet":577,"endOffSet":581}]},"

Date":"2016-10-04T08:52:47.263Z"

}

category	pre-defined categories such as the locations, subject, and period
value	Identified tokens
startOffSet/endOffSet	string start and end index

From the small paragraph of sample text, the service has successfully recognised:

Placename: Butser Ancient Farm

Placename: Little Butser Hill

Placename: Hampshire

Placename: Bascomb Down

Subject: farm

Subject: Villa

Temporal: British Iron Age circa 300 BC

Temporal: Roman

Temporal: 2002

As stated in D16.2, ADS planned to test the API as part of the redevelopment of the OASIS system (the online system for indexing archaeological grey literature in the UK). The aim was to allow an archaeologist to upload a report to OASIS, and by choosing to use the NER service, be able to automatically extract suggested metadata for the report. The metadata could then be accepted or rejected by the user and then automatically populated into the correct fields within OASIS. Unfortunately the timeframe for this major re-development project across several UK organisations

has been delayed, and this testing was not possible within the ARIADNE project. The API was circulated to ARIADNE partners for review however, and both the University of South Wales and INCIPIT CSIC tested the service and provided interim feedback. It was found that while the service did not return any false positives, it failed to return all potential positives. This would indicate that while the metadata generated by the service is reliable, it may not be complete. It was determined that this was likely due to a need for more training data, and/or an adjustment to the algorithm. ADS will continue to work on the service beyond the completion of the ARIADNE project, to see if further improvement is possible. The service is currently freely available for use by the archaeological community, and is one of the services developed though ARIADNE, but available to all.

6 Conclusion

The ARIADNE partners involved in this deliverable continued to explore NLP with the aim of making text-based resources more discoverable and useful. The partners have specifically focused on one of the most important, but traditionally difficult to access resources in archaeology; the largely unpublished reports generated by commercial or "rescue" archaeology, commonly known as "grey literature". The partners explored aspects of rule-based and machine learning approaches, the use of archaeological thesauri in NLP, and various Information Extraction (IE) methods.

USW extended their English language rule based methods using the GATE toolkit for NER (Named Entity Recognition) to Dutch and Swedish language grey literature reports, in collaboration with LU and DANS (Dutch reports) and SND (Swedish reports). This made use of glossaries and thesauri from the partners, including the Dutch Rijksdienst Cultureel Erfgoed (RCE) Thesauri. The process of importing the thesauri resources into a specific framework (GATE), and the suitability and performance of the selected resources when used for the purposes of Named Entity Recognition (NER) were analysed.

The NER techniques were focused on the general archaeological entities of Archaeological Context, Material, Physical Object (Finds), Monument, Place, and Temporal (Time Appellation). The methods proved capable of extracting CIDOC CRM element and in some case studies Getty Art and Architecture Thesaurus concepts, in addition to the native vocabularies. The English language NLP pipeline has been evaluated in previous work for the STAR Project and has gone through several iterations. The Dutch and Swedish pipelines were evaluated as part of the ARIADNE work and the findings are reported in this deliverable. In total, three versions of the Dutch pipeline and two versions of the Swedish pipeline were developed.

General archaeological NLP (GATE) pipelines for English, Dutch and Swedish have been developed. In addition experimental pipelines were developed for two exploratory thematic case studies on data integration, where the output is expressed as RDF Linked Data via a CRM based data model. An English language pipeline is available for a numismatic case study. English, Dutch and Swedish pipeline are available for a case study of item level data/NLP integration on a loose theme based around archaeological interest in wooden objects and their dating, as expressed in different kinds of datasets and reports. Both case studies have resulted in interactive demonstrators operating over the ARIADNE Linked Data Cloud. All seven pipelines are freely available as open source ARIADNE outcomes.

Within the constraints of the resources available for WP16, the rules for the Dutch and Swedish NER and RE pipelines were not as elaborate as the pattern-based English language rules. Nonetheless, the outcomes are promising and show the potential for the application of NLP methods to Dutch and Swedish language reports. Further work is needed before an operational capability is achieved, as discussed above. In particular, work on enlarging the vocabularies available for the NLP and structural modification of these resources would be helpful, including adapting the thesaurus terminology in some cases for NLP purposes.

Further development of techniques for the annotation of compound noun forms are important for extending the English language techniques to Dutch and Swedish. Tables pose challenges in all languages and merit a specialised NLP module. The ambiguity between material and object in natural language use should be revisited. This has proved a problematic issue in each language for both machine and human annotators. If the distinction is indeed important (it may not be depending on the use case) then further refinement of NER techniques is required. This could include identification of context markers for each case to inform pattern based rules. A more elaborate list of context markers for dates would be a cost effective addition in light of the archaeological concern with dating. An appropriate set of expert annotated reports is necessary for evaluating and improving NLP techniques. The set of entities for NER should be revisited for the intended use cases. Effort should

be devoted to creating annotation guidelines tailored to the context of each language and creating a gold standard set of annotated reports. This should itself be evaluated; it is sometimes the case that apparent false positives are in fact caused by omissions in the manual annotation.

The Archaeology Data Service (ADS) at the University of York, continued to develop and evaluate machine learning-based NLP techniques and integrate them into a new metadata extraction Named Entity Recognition module, which takes previously unseen English language text as input, and identifies and classifies named entities within the text. The outputs can then be used to enrich the resource discovery metadata for existing and future resources. The final output for this deliverable was intended to be a more refined Web application interface, but after additional consideration it was decided this would be less useful than the creation of an NER Web Service API, which could be implemented by anyone in the archaeological community.

Early work revealed the NER module worked successfully and produced correct entities for the classes it was trained to identify. The NLP tools were very useful for extracting resource discovery metadata from unstructured archaeological data, particularly grey literature reports, for resource discovery indexing, where little or no metadata currently exists. From a data management perspective however, the large quantities of entities extracted by the NER module may be too large to effectively manage and this will need further exploration. The Web Service API is currently available for use and integration into other interfaces, and will continue to be developed beyond the completion of the ARIADNE project.

The partners have successfully explored a variety of NLP techniques to make text-based archaeological resources more discoverable and useful. However, there are still areas requiring further work to fully achieve the potential of the techniques explored.

Negation in unstructured archaeological text content was observed during the evaluation. It is crucial for any practical implementation this is recognised. For example, if a named entity occurs inside the scope of a negation then that named entity should not be included in the output. Negation detection should be explored.

Some English language NLP research has begun to investigate the issue of negation detection in archaeological grey literature reports, with a view to distinguishing a finding of evidence, for example, of Roman activity from statements reporting a lack of evidence, or no sign of Roman remains. A technique previously used in the biomedical domain was adapted to archaeological vocabulary and writing style. Evaluation on rules targeted at identifying negated cases of four CIDOC-CRM entities gave promising results, Recall 80% and Precision 89%[17]. Further research is needed on negation detection (e.g. a negative finding) and the ability to discriminate in reports between important findings of archaeological evidence and mentions in passing of less important information. This is picked up below.

Feedback from domain experts suggests that although the entities detected by the system are valid terms, some are not considered important. Although, we do not think this is a problem from an NLP perspective, nevertheless, it is desirable for the system to be more selective. So far, work has been focused on NER, but it may be possible to solve this issue using techniques from Entity Linking (EL). The problem is distinct from the NER module, as it does not identify the occurrence of the “names”, but their reference. In order to build such a system, a knowledgebase is needed. It may be possible to develop this knowledgebase from available archaeological Linked Open Data. Entity linking can be defined as matching a textual entity detected by the NER module, to a knowledgebase entry, such as

[17] Vlachidis A, Tudhope D. 2015. Negation detection and word sense disambiguation in digital archaeology reports for the purposes of semantic annotation. Program: electronic library and information systems 49, 2 (2015), 118-134.

a Linked Data node that is a canonical term for that entity. However, entities are often detected by the NER module which have different surface forms, including abbreviations, shortened forms, or aliases. Therefore, EL must find an entry despite changes in the detected string by the NER module. Entity Ambiguity (EA) resolution is another problem that will need to be resolved when using this technique. For instance, “Roman”, can match multiple Linked Data entries as either “subject” or “temporal”. The last difficulty is the absence of the entity in the knowledgebase. Processing large text collections guarantees that many entities will not appear in the Linked Data, so the system may not be able to cope with this situation. Addressing a “negative sample” could be used when creating the training data, as opposed to the positive samples taken during the original training of the data used for this tool.

7 Appendix 1: Instructions for Annotating Grey Literature Documents

The manual annotation task aims to annotate grey literature documents with respect to archaeological concepts that relate to cultural and heritage data. The annotation is focused on the Named Entity Recognition (NER) task and particularly in the identification of the following concepts;

- Time Appellations,
- Archaeological Objects
- Materials
- Places
- Monument (**Complex)** Types
- Archaeological Context types (also known as features).

The annotation should target the above types in "isolation", following the aims of NER, thus activities, relationships and events are not in scope. Actors are not in the scope of the current task.

It is proposed to use the following highlight colours to mark the annotations.

- Time Appellations (Blue)
- Archaeological Objects (Brown)
- Materials (Purple)
- Places (Yellow)
- Monuments (Green)
- Archaeological Context (Light-Green)
- Negation (Red)

You may use any other highlight colours of your choice for marking the annotations. In this case please give a colour key definition at the top of the document.

Annotation Principles

Annotators should produce their manual annotations with the following principals in mind.

1. **Negation Detection**: Any of the above entities that are negated should be annotated as Negation. For example "No evidence of pottery" should be annotated as Negation. The span of this particular annotation types should cover the WHOLE sentence clause denoting negation.
2. To consider how relevant is the entity to the overall discourse. Topicality could affect cases of ambiguity e.g. Physical Objects or Materials. For example the term 'brick' can either refer to a material (a brick wall) or to a physical object (a brick found in context). Annotators should decide on the conceptual alignment of terms that can be either materials or physical objects. Other case of topicality might affect Place names which can also be as people surnames (common in English not so sure if this is the case in Dutch).
3. Annotators should consider, plural when applicable as well as spelling variations and acronyms common in the archaeology domain eg CBM (cERAMIC Building Material). Compound words containing any of the targeted entities should also be annotated. The annotation should span only on the part of compound word relevant to an entity type. If more than one entities are relevant then respective annotationhighlight colours should be used for distinguishing the parts of the compound word. Compound works are common in Dutch (not that much in English). For example the word "steentijd vindplaatsen" should deliver two annotation spans (steentijd as Time Appellation) and (vindplaatsen as Place).
4. Annotators should consider conjuncted phrases. For example annotators should consider conjunctions of the kind 'Early Roman to Late Roman', 'Pottery and brick', as well as 'worked flint', 'small finds' etc.

Entities Annotation

Time Appellations: All time appellations both numerical and lexical. eg Roman, 1045 AD, early-mid Iron age etc

Archaeological Objects: Objects of archaeological interest such as finds, small find, architectural elements etc.

Materials: Objects of archaeological interest, contemporary material of little archaeological interest such as plastic should be excluded from annotation.

Places: Places of archaeological interest and relevant Place names. Grid references may also be annotated.

Monument (Complex) Types: Such as building types and architectural features

Archaeological Context types: Contexts revealed during the excavation process, also knows in Dutch as features, such as pit, pit fill, deposit, and larger context groupings as post-hole, post-hole structures, circular pits etc.

8 Appendix 2: Initial glossary of Swedish date context markers

±

B.P.

BP

e.Kr.

e.v.t.

Efter Kristus

Efter vår tideräkning

evt

f.Kr.

f.v.t.

f.v.t.b

fvt

fvtb

Före Kristus

Före vår tideräkning

Medel

Medeltida

Sen

Sentida

v.t.

vt

Yngre

Ålder

Äldre