TITLE: TOWARDS AN INTERPRETATION SUPPORT SYSTEM FOR READING ANCIENT DOCUMENTS

Authors: Roued Olsen, H.[1], Tarte, S. M.[2], Terras, M.[3], Brady, J. M.[4], Bowman, A. K.[1]

[1] Centre for the Study of Ancient Documents, University of Oxford
[2] Oxford e-Research Centre, University of Oxford
[3] School of Library Archive and Information Studies, University College London
[4] Department of Engineering, University of Oxford

Contact Author: Roued Olsen, H.

Email: henriette.olsen@classics.ox.ac.uk

ABSTRACT

Constructing readings of damaged and abraded ancient documents is a difficult, complex, and time-consuming task, often involving reference to a variety of linguistic and archaeological data sets, and the integration of previous knowledge of similar documentary material. Due to the involved and lengthy reading process, it is often difficult to record and recall how the final interpretation of the document was reached, and which competing hypotheses were presented, adopted, or discarded in the process of reading. This paper discusses the development of an Interpretation Support System (ISS), which aims to provide a system, which can aid the day-to-day reading of ancient documents, and in future other damaged documents, by keeping track of how these are interpreted and read. Such a system will facilitate the process of transcribing texts by providing a framework in which experts can record, track, and trace their progress when interpreting documentary material. Furthermore, it will allow continuity between working sessions, and the complete documentation of the reading process that has hitherto been implicit in published editions.

INTRODUCTION

The process of reading ancient documents is traditionally undertaken by an expert such as an epigrapher, palaeographer or papyrologist. The expert uses accumulated knowledge combined with external resources to piece together an interpretation of each ancient document. Such interpretation is often prolonged, and it can be difficult for experts to maintain a record of the interpretations made whilst undertaking their reading (Youtie 1963). This is important when disputing interpretations and sharing hypotheses with other experts, or pausing the reading of an ancient text and hoping to continue the same thought process at a later time.

The Image, Text, Interpretation: e-Science, Technology and Documents project (also known as eSAD: e-Science and Ancient Documents, http://esad.classics.ox.ac.uk), aims to use computing technologies to aid experts in reading ancient documents. The project is developing an Interpretation Support System (ISS) that can support the day-to-day reading and interpretation of ancient documents. This involves advanced IT tools that can aid the interpretation of damaged texts such as the stylus tablets from Vindolanda (http://vindolanda.csad.ox.ac.uk) and image processing algorithms to analyse detailed digital images of the documents (Tarte *et al*. 2008).

BACKGROUND

Although Classics as a subject has made much use of information technology (see Crane 2008 for an overview), the use of IT to aid in the actual reading process of ancient documents is in its infancy. Terras (2006) developed a prototype system which demonstrated that it was possible to propagate plausible and useful interpretations of ancient texts, in a realistic timeframe. This used linguistic and palaeographic datasets to provide the "knowledge base" which could inform a decision making system to aid experts in reading texts.

Decision Support Systems (DSS) have previously been developed in the Department of Engineering Science at the University of Oxford to aid multi-disciplinary teams working with cancer patients in making decisions about their treatment (Austin *et al*. 2008). This system is based on a set of rules and allows experts to analyse and interpret digital images while recording decisions made about diagnosis and treatment, and suggesting possible next action steps.

BUILDING THE INTERPRETATION SUPPORT SYSTEM

The research presented here, though inspired by the above-mentioned medical application, shifts the focus from a Decision Support System to an Interpretation Support System (ISS). In contrast with medical practitioners, experts transcribing ancient documents do not make decisions based on evidence but instead create interpretations of the texts based on their perception. The ISS relies upon the idea that an interpretation is made up of a network of minor perceptions (percepts) ranging from low level percepts such as "these three line fragments are an incised stroke" to higher level percepts such as "these five letters can make up the word *'legio'*".

We want to make this otherwise implicit network of percepts explicit in a human-readable format through a web browser based application. To build an explicit network of percepts leading to an interpretation, we define an elementary percept as a region of an image that contains what is perceived to be a grapheme. The image can then be divided into cells where each cell is expected to contain what is perceived as a character or a space. This division of the image constitutes a tessellation. A single document might be tessellated in various ways and each of the tessellations might yield either an interpretation or a dead-end, but in both cases, the explicit network of percepts will document this.

The making of the tessellation, which in itself is an interpretive process, marks the boundary between lower and higher level percepts. The lower level percepts are based on physical identification of the features of the document (through the application of image analysis methods to detect features such as strokes); the higher level percepts (words, groups of words) work more towards gradually adding meaning to the transcription in progress. Ultimately, an interpretation can then be represented as a network of substantiated percepts, which will be made explicit through an ontology. Here an ontology is defined as a model of the concepts found in a text such as the concept of a word that contains several characters.

The ontology aims to make the rationale behind the network of percepts visible and thus expose both: (a) some of the cognitive processes involved in damaged texts interpretation; and (b) a set of arguments supporting the tentative interpretation. The system will use the ontology as a framework to assist the expert through the different levels of percepts ultimately yielding a final transcription. The transcription is a part of the overall edition of which there may be several and it will be formatted in EpiDoc style XML (http://epidoc.sourceforge.net/) allowing further interaction with other documents.

BUILDING THE KNOWLEDGE SETS

Much of the knowledge base that serves as justification for the commitment to a given percept during the interpretation process will come from the experts. However, letter frequency, word- and character-lists from documents such as the Vindolanda ink tablets will provide an invaluable source of information which can be used to generate the statistical likelihood of patterns in language and writing which may appear on the texts. We have taken a new approach to the XML encoding of the Vindolanda ink tablets based on contextual encoding (Hippisley 2005). The Vindolanda ink tablets have been encoded with EpiDoc standard XML to a very detailed granularity. The contextual encoding which is then imposed on the documents consists on encoding words, person names, geographical place names, calendar references and abbreviations. For example any instance of the word *pulli* (='chickens') in a document will be encoded <w lemma="*pullus*" n="1">*pulli*</w>. This encoding provides us with the information that the word *pulli* has the lemma *pullus* under which we can index this instance of the word and that this is the first instance of this lemma in the document. This information has been used to generate word frequency lists and is extremely useful as a part of a knowledge base to build the ISS on. Further knowledge bases will be generated from the marked up dataset, to provide uncertainty and character frequency lists. Additionally, further work will be undertaken with the experts to generate lists of common percepts and interpretation making processes. By encoding these in XML, the knowledge sets for the system will be in place.

CONCLUSION

The construction of an Interpretation Support System for ancient texts, although ambitious, will provide a useful tool for those experts who work on developing interpretations of damaged documents by facilitating and recording the evolving interpretation process. Additionally, by making explicit the percepts which trigger such transcriptions of ancient documents, we will further our knowledge of the reasoning process undertaken by experts in propagating readings of ancient documents. Furthermore, the successful development of an image and language based Interpretation Support System will provide a set of tools which can be adopted and adapted by other domains which rely on detailed analysis and interpretation of image based material.

REFERENCES

Austin, M., Kelly, M., Brady M. (2008), “The benefits of an ontological patient model in clinical decision-support”. In Fox, D. and Gomes, C. P. (eds), AAAI, pages 1774–1775. AAAI Press.

Crane, G. (2008) “Classics and the Computer: An End of the History”. In Schreibman, S., Siemens, R., and Unsworth, J. (eds). A Companion to Digital Humanities. Blackwell Companions to Literature and Culture. Blackwell.
http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-4&toc.depth=1&toc.id=ss1-2-4&brand=default

Hippisley, D. (2005) "Encoding the Vindolanda tablets: an investigation in contextual encoding using XML and the EpiDoc standards." MA Dissertation submitted for the MA in Electronic Communication and Publishing, School of Library, Archive and Information Studies, UCL.

Tarte, S. M., Brady, J. M., Roued Olsen, H., Terras, M., Bowman, A. K. (2008), "Image Acquisition and Analysis to Enhance the Legibility of Ancient Texts". UK e-Science Programme All Hands Meeting 2008 (AHM2008), Edinburgh, September 2008.

Terras, M. (2006)."Image to Interpretation: Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts". Oxford Studies in Ancient Documents. Oxford University Press.

Youtie, H. C. (1963): "The Papyrologist: Artificer of Fact". GRBS 4 (1963), p. 19-32.