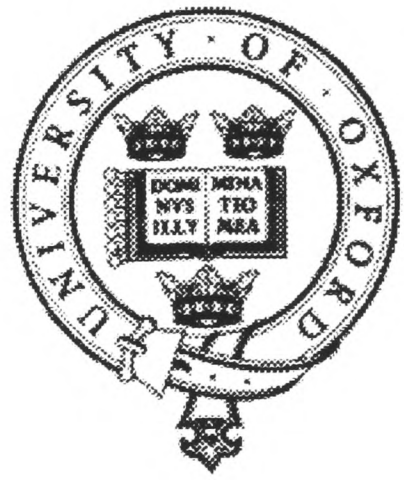
# Image to Interpretation:

## Towards an Intelligent System to Aid Historians in the Reading of the Vindolanda Texts

Phd

Melissa M. Terras

Christ Church




Robotics Research Group,

Department of Engineering Science,

University of Oxford.

Michaelmas 2002

A thesis submitted to the Department of Engineering Science, University of Oxford, in partial fulfilment of the requirements for the degree of Doctor of Philosophy. It is substantially the result of my own work except where explicitly indicated in the text.

# Abstract

## Image to Interpretation:

### Towards an Intelligent System to Aid Historians in the Reading of the Vindolanda Texts

Melissa Terras
Christ Church

Doctor of Philosophy
Michaelmas 2002

The ink and stylus tablets discovered at the Roman Fort of Vindolanda have provided a unique resource for scholars of ancient history. However, the stylus tablets in particular have proved extremely difficult to read. The aim of this thesis is to explore the extent to which techniques from Artificial Intelligence can be used to develop a system that could aid historians in reading the stylus texts. This system would utilise image processing techniques that have been developed in Engineering Science to analyse the stylus tablets, whilst incorporating knowledge elicited from experts working on the texts, to propagate possible suggestions of the text contained within the tablets.

This thesis reports on what appears to be the first system developed to aid experts in the process of reading an ancient document. There has been little previous research carried out to see how papyrologists actually carry out their task. This thesis studies closely how experts working with primary sources, such as the Vindolanda Texts, operate. Using Knowledge Elicitation Techniques, a model is proposed for how they read a text. Information regarding the letter forms and language used at Vindolanda is collated. A corpus of annotated images is built up, to provide a data set regarding the letter forms used in the ink and stylus texts.

In order to relate this information to the work done on image processing, a stochastic Minimum Description Length (MDL) architecture is adopted, and adapted, to form the basis of a system that can propagate interpretations of the Vindolanda texts. In doing so a system is constructed that can read in image data and output textual interpretations of the writing that appears on the documents.

It is demonstrated that knowledge elicitation techniques can be used to capture and mobilise expert information. The process of reading ancient, and ambiguous texts, is made explicit. It is also shown that MDL can be used as a basis to build large systems that reason about complex information effectively. This research presents the first stages towards developing a cognitive visual system that can propagate realistic interpretations from image data, and so aid the papyrologists in their task.

# Acknowledgements

First and foremost, I would like to thank my supervisors Professor Mike Brady and Professor Alan Bowman for their support and guidance throughout this project. Although coming from different sides of the academic divide their mutual guidance has ensured that such interdisciplinary research did not flounder, and for that I am grateful. Mike Brady has always provided the reassurance, backup, and the odd plane fare when necessary to ensure success. I like to think this thesis has surprised us both. Alan Bowman's door was always open, and I thank him for his support and encouragement.

The thesis would also be a very different entity without the input, advice, and support of Dr. Paul Robertson from the AI Lab at MIT. I would like to thank him for his patience and the contribution he made to this work, particularly regarding Chapter 4. I would also like to thank Paul and his wife Grace for their generosity and hospitality during my visit to MIT.

The research in Chapter 2 would not have been possible without the consent and enthusiasm of my guinea pigs (in no particular order): Professor Alan Bowman, Dr Roger Tomlin, and Professor David Thomas. All welcomed me into their offices and homes whilst I brandished a thousand questions and a tape recorder. This research could not have taken place without access to their expertise.

I would also like to thank Dr Charles Crowther and Dr John Pearce, both from the Centre for the Study of Ancient Documents at Oxford. They provided me with various information, files, and pointers along the way, and were always helpful, however bizarre my requests.

Dr Xiabo Pan and Dr Veit Schenk answered my many questions about the image processing angle of the project. Xiabo was very kind in helping me to prepare various illustrations for this thesis.

It was Dr Seamus Ross from the Humanities Advanced Technology and Information Institute, University of Glasgow, who encouraged me along this path in the first place, and I am grateful for the faith he has shown in me over the past few years.

Edward Vanhoutte, from the Centrum voor Teksteditie en Bronnenstudie at Ghent, has been a good friend and dancing partner as well as a close academic ally. I appreciate the time and effort he has spent assisting me through this, and other, projects.

This thesis was proof read by Margaret Terras, and I thank her for her patience and effort.

There are numerous people who have not been directly connected with this research but who have nevertheless helped me to keep my head above water. Timor Kadir, Veit Schenk, Steve Reece, Cathy Dolbear, and Vicky Mortimer made the Terrapin Hut an interesting place to work. Natalie Walker and Dr Fleur Taylor have always looked after my welfare. Additionally, I would like to thank the following: Justin Pniower, James Melville, Thom Falls, Rosalind Porter, Verity Platt, Alison Parkinson, Andy Fry, Duncan Robertson, Fiona Kimberley, David Beaumont, and Claire Easingwood. Caroline Davidson and Joyce Millar from Christ Church deserve some thanks for the practical help they have given me over the past few years.

Finally, I would like to thank my parents, Margaret and Robert Terras, and my grandparents, Helen and James Nelson, for their interest and support, particularly throughout the last year.

# Contents

# CHAPTER 1

# Introduction

"We can try a little experiment. Let us resort to the fiction of programming an information transducer, a machine to read [ancient] texts. While so far only human beings have learned it, it is equally possible, and may one day be tried, to teach this skill to a machine ..."
E. Reiner (Journal of Cuneiform Studies, 1973, p.6)

The ink and stylus texts discovered at Vindolanda are a unique resource for scholars of the Roman occupation of Britain. This thesis reports on the development of an appropriate knowledge based system, to aid papyrologists in the reading of the stylus texts which uses input from image processing techniques as well as additional procedural information elicited from the experts themselves.

The texts from Vindolanda, a Roman fort on the Stanegate near Hadrian's Wall and modern day Chesterholm, are an unparalleled source of information regarding the Roman Army for historians, linguists, palaeographers, and archaeologists. The hand-writing on the ink texts can be made visible through the use of infrared photography. However, due to their physical state, the stylus tablets (one of the forms of official documentation of the Roman Army) have proved almost impossible to read. This chapter provides an introduction to the ink and stylus texts, before discussing the developments in image processing that have allowed some features of the stylus tablets to be detected. The focus of this thesis is then introduced: the construction of a computer system to aid the papyrologists in the reading of the stylus texts. Background information is given regarding the use of

computing in the field of papyrology, and the use of Minimum Description Length as a means to compare and contrast complex information within the field of image processing. An MDL based system, GRAVA, is introduced, that allowed the integration of image and textual data into a reasoning system to generate possible interpretations of the text contained within images of the Vindolanda tablets. Finally, a thesis overview is provided.

## 1.1 The Vindolanda Texts

The two types of texts discovered at Vindolanda[1] are unparalleled resources for classical historians because textual sources for the period in British history from AD 90 to AD 120 are rare[2]. The ink and stylus tablets are a unique and extensive group of written documents from the Roman Army in Britain, and provide a personal, immediate, detailed record of the Roman fort at Vindolanda from around AD 92 onwards (Bowman and Thomas 1994; Bowman 1997). The texts

> now cast a flood of light – or at least a galaxy of pinpoints of light – upon a Dark Age in northern Britain, upon the Roman Army, its logistics, organisation and social structure, and upon the spoken and written language of the time and milieu (Tomlin 1996, p.463).

The ink tablets, carbon ink written on thin leaves of wood cut from the sapwood of young trees, have proved the easiest to decipher. In most cases, the faded ink can be seen clearly against the wood surface by the use of infra red photography, a technique used frequently in deciphering ancient documents (Bearman and Spiro 1996). The majority of the three hundred writing tablets that have been transcribed

---

[1] For further information regarding Vindolanda see Breeze and Dobson (1976), Birley (1977), Bidwell (1985), and Bidwell (1997). For an account of the discovery of the tablets see Birley, Birley et al. (1993), and Birley (1999).

[2] Aside from the Vindolanda documents, the main sources of textual information regarding Roman Britain are histories, inscriptions, and coins. Examples of these can be found in Ireland, Chapter IX (1986).

so far contain personal correspondence, accounts and lists, and military documents (Bowman and Thomas 1994).

**Figure 1.1:** Ink text 291, captured by digital infrared photography. This diptych contains a letter to Lepidina, the wife of Flavius Cerialis who was the prefect of the Ninth Cohort of Batavians stationed at Vindolanda, from Claudia Severa, the wife of Aelius Brocchus. The letter is a warm invitation to Severa's birthday celebration. Incidentally, this tablet almost certainly contains the earliest known example of writing in Latin done by a woman, in the closing comments, bottom right.

The two hundred stylus tablets found at Vindolanda appear to follow the form of official documentation of the Roman Army found throughout the Empire (Turner 1968; Fink 1971; Renner 1992). It is suspected that their subject matter will differ from the ink tablets as similar finds indicate that stylus tablets tended to be used for documentation of a more permanent nature, such as legal papers, records of loans, marriages, contracts of work, sales of slaves, etc (Renner 1992). However, the palaeographic and linguistic characteristics of the stylus tablets may reasonably be expected to be similar to the ink texts as they are contemporaneous documents from the same source.

**Figure 1.1:** Stylus tablet 836, one of the most complete stylus tablets unearthed at Vindolanda. This text is a letter from Albanus to Bellus, containing a receipt and further demand for payment of transport costs (see footnote 11 in Chapter 2 for a transcription of this text). The incisions on the surface can be seen to be complex, whilst the woodgrain, surface discoloration, warping, and cracking of the physical object demonstrate the difficulty papyrologists have in reading such texts.

Manufactured from softwood with a recessed central surface, the hollow panel of the stylus tablets was filled with coloured beeswax. Text was recorded by incising this wax with a metal stylus, and tablets could be re-used by melting the wax to form a smooth surface. Unfortunately, in nearly all surviving stylus tablets[3] the wax has perished, leaving a recessed surface showing the scratches made by the stylus as it penetrated the wax[4]. In general, the small incisions are extremely difficult to decipher[5]. Obtaining a reading is complicated further by the pronounced woodgrain

[3] It is suspected that around 2000 of such tablets exist outside Egypt (Renner 1992).

[4] Only one stylus tablet, 836, has been found so far with its wax intact. Unfortunately this deteriorated during conservation, but a photographic record of the waxed tablet remains to compare the visible text with that on the re-used tablet. A discussion regarding this tablet can be found in Bowman and Tomlin ( Forthcoming 2003).

[5] Even Roman readers sometimes had difficulty with reading stylus tablets. In a scene from Pseudolus, by the Roman playwright Plautus, the eponymous character, the slave Pseudolus, is unimpressed by the writing on a wax tablet sent to his master, Calidorus, by the young man's lover:
Pseudolus: All these letters, they seem to be playing at fathers and mothers, crawling all over each other.
Calidorus: Oh, if you're going to make a joke of it -
Pseudolus: It would take a Sibyl to read this gibberish; no one else could make head nor tail of it.

of the fir wood used to make the stylus tablets, staining and discoloration of the surface, damage over the past two thousand years, and the palimpsestic nature of the re-used tablets. A skilled reader can take several weeks to transcribe one of the more legible tablets, and several months to transcribe a portion of some of the more damaged texts, whilst many have to date defied any interpretation. Prior to the current project, the only way for the papyrologists to detect incisions in a tablet was to move the text around in a bright, low raking light. In doing this, indentations are highlighted and candidate writing strokes become apparent through the movement of shadows, although this proves to be a frustrating, time consuming, and inefficient way of reading the texts.

## 1.2 Image Processing Techniques

In 1998 the Department of Engineering Science and the Centre for the Study of Ancient Documents at the University of Oxford were jointly awarded a research grant by the Engineering and Physical Sciences Research Council (EPSRC) to develop techniques for the detection, enhancement and measurement of narrow, variable depth features inscribed on low contrast, textured surfaces (such as the Vindolanda stylus tablets). To date, the project has developed a wavelet filtering technique that enables the removal of woodgrain from images of the tablets to aid in their transcription (Bowman, Brady et al. 1997). In addition, a technique called "Shadow Stereo" or "Phase Congruency" has been developed, in which camera

---

Calidorus: Why are you so unkind to those dear little letters, written on that dear little tablet by that dear little hand?
Pseudolus: A chicken's hand was it? A chicken surely scratched these marks ... (Plautus 1965, Act 1, Scene 1).

[6] Papyrology can simply be defined as obtaining "a body of knowledge ... from the study of papyri." It is now taken to cover "as a matter of convenience ... the study of all materials carrying writing ... done by a pen" (Turner 1968, p. vi).

position and the tablet are kept fixed; but a number of images are taken where the tablet is illuminated by a strongly orientated light source. If the azimuthal direction of the light sources (that is, the direction to the light source if the light were projected directly down on to the table) is held fixed, but the light is alternated between two elevations, the shadows cast by incisions will move but stains on the surface of the tablet remain fixed. This strongly resembles the technique used by some papyrologists who use low raking light to help them read the incisions on the tablet (Molton, Pan et al. Forthcoming (2003)). Edge detection is accomplished by noting the movement of shadow to highlight transitions in two images of the same tablet, and so candidate incised strokes can be identified by finding shadows adjacent to highlights which move in the way that incised strokes would be expected (Schenk ,2001). Although this is not a standard technique in image processing, encouraging results have been achieved so far (see 4.3.1), and a mathematical model has been developed to investigate which are the best angles to position the light sources (Molton, Pan et al. Forthcoming (2003)). Work currently being undertaken is extending the performance and scope of the algorithms (Pan, Brady et al. Forthcoming (2003)), and the papyrologists are beginning to trust the results and suggestions which are being made about possible incisions on the tablets (Bowman and Tomlin Forthcoming 2003). Future work will be done in relating the parameters of analysis to the depth profile of the incisions to try and identify different overlapping writing on the more complex texts.

## 1.3 An Interface for the System

Whilst the techniques discussed above have had some success in analysing the surfaces of the tablets, there needs to be a method developed to aid the

papyrologists in utilising generated results. The algorithms developed could easily be added to readily available image manipulation software (for example using the Visual C++ plugin with PhotoShop). Although this would allow others to apply the algorithms themselves it would do little to actually provide a tool that would actively help the papyrologists in the transcription of texts; a complex process which has been described as "teasing information out of material which is all too often barely legible, fragmentary, or obscure (or all three at once)," (Bowman 1994, p.10). The focus of this thesis is to investigate the possibility of developing an intelligent system that can aid the papyrologists in their task.

## 1.4 Papyrology and Computing

The use of computing in the field of Papyrology[6] has enjoyed some notable successes. There are many established imaging projects, such as those at the CSAD[7], and the Oxyrhynchus Papyri Project[8]; excellent database projects and systems such as the Duke Bank of Documentary Papyri[9], and APIS (Advanced Papyrology Information System)[10]; and repositories of information in a user friendly format such as the Perseus Project[11]. Many standards are already in place for the digitisation and markup of ancient texts, and papyrologists are making more use of the kind of image manipulation tools provided by the likes of PhotoShop (Bagnall 1997). Simple image processing techniques have been used to aid scholars in the reading of individual texts (from both the ancient and modern period), such as the

[7] http://www.csad.ox.ac.uk
[8] http://www.csad.ox.ac.uk/POxy/
[9] http://odyssey.lib.duke.edu/papyrus/texts/DDBDP.html
[10] http://odyssey.lib.duke.edu/papyrus/texts/APISgrant.
[11] http://www.perseus.tufts.edu, or the mirror site at CSAD.

Beowulf manuscript (Kiernan 1991; Prescott 1997). Image manipulation tools are being developed which enable experts to virtually manoeuvre texts and light sources (Lundberg 2002). However, research done in the area of image processing of ancient documents often concentrates on the computational element, with little focus being given to the needs of the experts trying to read the documents (see Stark 1992; Seales, Griffioen et al. 2000; Brown and Seales 2001 for examples of poorly thought out projects involving the imaging of ancient texts). There is often very little consultation with the experts the tools are being developed for, and scant understanding of their requirements[12].

No systems currently exist to support papyrologists in the *process* of reading ancient texts[13]. Indeed, there is little research published which discusses how information is actually extracted from these texts (see 2.1.1) and there does not exist detailed cognitive and/or perceptual information processing models of the papyrology process. From a Cognitive Psychology stance, although there has been much consideration of the processes involved in reading (see 2.1.2) few conclusions have been drawn as to how a reader would approach such damaged, fragmentary, foreign language texts and construct a logical, acceptable meaning. Also, although image processing is an expanding field in the discipline of Engineering Science (see Gonzalez and Woods 1993 for an introduction) little work has been done on the role of knowledge and reasoning in the analysis and understanding of complex images. Proposals for integrating image analysis algorithms with techniques for the

[12] Newer ethnological methods of Knowledge Elicitation continue to be developed which work closely with users, see Preece, Rogers, and Sharp (2002).

[13] Levison, and, independently, Ogden considered the potential for the use of computers in the reconstruction of ancient manuscripts in the late 1960's but were hampered by lack of computing power (Levison 1965; Levison 1967; Ogden 1969; Levison 1999). Wacholder and Abegg (Wacholder and Abegg 1991) used a computer in an effort to reconstruct the unpublished Dead Sea Scrolls, but that is a much simpler task than the one that is described here.

representation and mobilisation of knowledge (the subject of the field of Artificial Intelligence) remain few.

There are some major issues that need to be tackled to develop a complete system to aid papyrologists in their task:

1. An understanding of how papyrologists operate was required.
2. Once this was achieved, Knowledge Elicitation needed to be carried out on all semantic levels of the process: from the information used regarding character form and identification (palaeography), to an understanding of the type of language that was used in the texts.
3. A way of representing the elicited knowledge was required. A system was needed to capture and mobilise information regarding letter forms. Statistics regarding the use of language at Vindolanda had to be obtained.
4. It was necessary to find a way of linking this data to the existing image processing techniques that have been developed to aid in the reading of the stylus tablets, so that stroke data from the feature detection algorithms could be compared with the information captured regarding character forms.
5. A way of comparing graphical and linguistic information had to be found, in order to generate probable interpretations of the texts.
6. A graphical user interface should be developed to deliver the system as a stand alone application to the experts to aid them in their day-to-day task.

This thesis extends significantly the understanding of how papyrologists carry out their task. It also makes explicit the type of information used by the experts regarding letter forms, and provides a way of representing and utilising this

information. Statistical knowledge about the language contained within the Vindolanda texts was generated. The elicited information regarding letter forms is linked to the work done on image processing through the use of a stochastic intelligent system. This system also incorporates linguistic knowledge, to generate possible interpretations of image data. Finally, some suggestions are made as to how this system could be delivered to the papyrologists as a desktop application.

## 1.5 Thesis Approach

### 1.5.1 Knowledge Elicitation

In order to identify the tools that could be built to aid the papyrologists in their transcription of the Vindolanda tablets, it was first necessary to try and gain an understanding of what the papyrology process actually entails. A program of "Knowledge Elicitation" was undertaken (see 2.2), and from this, a model of how papyrologists approach and start to understand ancient texts was developed (see 2.8.3). This was then related to current cognitive psychology theories regarding the resolution of ambiguity in texts during the process of reading. There are many papers regarding this phenomenally complex process, but only a few computer programs have been implemented to test the models postulated. Parallels between one of these major theories, the Interaction Activation and Competition Model of visual word recognition, which aims to explain the Word Superiority Effect, developed by McClelland and Rumelhart (McClelland and Rumelhart 1986, see 2.8.2), and the findings of the knowledge elicitation experiments were obvious. The model of how the papyrologists operate was taken as the basis for developing the architecture of the system described in Chapter 4 of this thesis.

### 1.5.2 Handwriting Recognition

The field of handwriting recognition is expansive and complex (see Impevedo (1993) for an introduction), but although has been a great deal of work on pattern recognition algorithms aimed at recognising text, the vast majority is irrelevant to this research. Most techniques are aimed at recognising printed text, making them incompatible with the hand-written, cursive text found on the Vindolanda tablets. Work done on hand written non-cursive text primarily approaches the problem with pattern class or neural network learning: where the system learns individual characters from thousands of examples. There are simply insufficient examples to be able to train such a neural net, even from the corpus of Vindolanda ink texts. Other work, largely from the late 1960s and early 1970s, emphasised "syntactic pattern classification": the idea that a character is composed of strokes that have a certain relationships to each other (see Connell and Brady's approach to shape representation (1987), and Fu and Swain's introduction to Syntactic Pattern Recognition (1969)). The attempts to teach a machine to "read" text in this manner were hampered by the problem of stroke detection: image processing techniques were not developed enough to provide the necessary data. Due to the subsequent advances made regarding feature detection (see 1.2), a similar, syntactic way of modelling character information was adopted in this thesis (see 3.3), as a means of capturing a set of data with which to compare information extracted from the images of the text.

There have been previous attempts to use connectionist inspired models of human reading as the basis on which to build systems to "read" cursive handwriting (Dodel and Shinghal 1995; Parisse 1996; Côté, Lecolinet et al. 1998). These systems have

had limited success: they are dependent on over simplification of word shape and contour, the models rely on strong contextual constraints, and are only successful with small lexicons of 25-35 words. Appropriating these systems in an attempt to build a tool to aid the stylus tablets would be unsuccessful for the same reason that existing image processing techniques could not be used to analyse the surface data: the data is too fragmentary, too noisy, and too complex. It has been suggested that using Minimum Description Length may provide a way to successfully model systems when only sparse data exists (Robertson, 2001).

### 1.5.3 Minimum Description Length

A major puzzle in perceptual psychology has been how the brain reconciles different sorts of information, for example colour, motion, boundaries of regions, or textures, to yield a percept (Eysenck and Keane 1997). For example, in image segmentation, by changing the texture model it is possible to increase or decrease the number of regions that are identified. Similarly, the number of boundary shapes that are found can be increased or decreased by changing the search parameters. The reading of ancient documents is just one (complex) example of an interpretation problem, where the individual is faced with visual ambiguity and competing information, which has to be reconciled and resolved with other types of information (in this case linguistic data) to generate a plausible solution. There has been substantial consideration, by psychologists and image processing experts, of how these different processes are combined. One suggestion is that there is a common value that can be used to calculate the "least cost" solution when comparing different types of information. This has been adopted by the field of Artificial Intelligence, in the concept of Minimum Description Length: (MDL).

First introduced in the late 1960s, and developed in the 1970s, (Wallace and Boulton 1968; Rissanen 1978), MDL applies the intuition that the simplest theory which explains the data is the best one[14]. MDL can be used as a means of comparing data in coding theory, in which the goal is to communicate a given message through a given communication channel in the least time or with the least power. MDL is a very powerful and general approach which can be applied to any inductive learning task, and can be used as a criterion for comparing competing theories about, or inductive inferences from, a given body of data. It is well suited to interpretation problems: where solutions can be generated by comparing unknown data to a series of models, when the most likely fit can generate plausible solutions to the problem.

MDL has been applied to numerous image processing problems, to provide a means to compare information and choose between competing solutions. Leclerc used MDL in an attempt to solve the image partitioning problem, to delineate regions in an image that correspond to semantic entities in a scene (1989). Zhu and Yuille utilised MDL to develop a novel statistical and variational approach to image segmentation through the use of region competition (1996). Other recent research which uses MDL as a means of comparing and contrasting possible solutions in the field of image processing includes Lanterman (1998), Rissanen (1999), and Gao and Ming (2000).

[14] The basic concept behind MDL is an operational form of Occam's razor, "Pluralitas non est ponenda sine neccesitate" or "plurality should not be posited without necessity." Formulated by the medieval English philosopher and Franciscan monk William of Ockham (ca. 1285-1349), the phrase has been adopted by Communication Studies to suggest that one should not increase, beyond what is necessary, the number of entities required to explain anything (Forster 1999).

In his doctoral thesis "A Self Adaptive Architecture for Image Understanding" Paul Robertson expands the remit of the type of information exchanged within an MDL based stochastic systems architecture (2001). Although previous (known) research had limited the use of MDL to image understanding, Robertson incorporated linguistic information into his system to build an application that could effectively "read" a hand-written phrase (see 4.1). Robertson's GRAVA system uses a model of reading similar to the Interaction Activation and Competition Model for Word Perception (McClelland and Rumelhart 1986) combined with MDL (see 4.1.1.1) and Monte Carlo Methods (4.1.1.2) to propagate possible interpretations of a written text. Robertson's example is limited to the reading of a small phrase from a nursery rhyme. It was suggested in his thesis that this architecture could be suitable for the implementation of a system that deals with much more complex data. There was much work to be done to adapt Robertson's system to the needs of this project.

### 1.5.4 System Construction

In the research presented within this thesis, the GRAVA system constructed by Robertson is adopted as the means by which to construct a system that can effectively "read" and "reason" about the texts found at Vindolanda. The original GRAVA system needed to be significantly adapted to carry out this more complex task. Also, much intricate data needed to be collected in such a way that it could be mobilised efficiently, to provide the information on which to base the system, in order to make the adoption of GRAVA feasible.

Firstly, an investigation was undertaken into the type of letter forms found at Vindolanda, which enabled an encoding scheme to be developed. Images of ink

and stylus texts were then annotated, using this encoding scheme, to build up a corpus of letter forms which could be used to train the system (see Chapter 3). Statistics regarding the language used at Vindolanda were generated from the ink tablet corpus and prepared for integration with the system. The GRAVA system was then adapted, trained on the image corpus, and used to generate interpretations of the Vindolanda texts (see Chapter 4). The resulting system was also tested with data that was generated from the automatic feature extraction algorithms by other members of the team (see 4.8). It is demonstrated that such a stochastic MDL architecture provides a means by which to interpret images of the Vindolanda documents, and as such provides the basis for a stand alone application with which to aid the papyrologists in their reading of the stylus texts.

It should be stressed here that the construction of this system is not an attempt to build an "expert system" that will automatically "read" and provide a fixed transcription of the texts, thus negating the input of the papyrologists. Rather it is an attempt to build a system with which papyrologists will be able to mobilise disparate knowledge structures, such as linguistic and visual clues, and use these in the prediction process to aid in the resolution of the ambiguity of the texts. The system as it stands propagates possible solutions to the input data. Integration of this into an application which allows the papyrologist to adjust parameters of the system would enable them to maintain an explicit record of the alternative hypotheses developed as they attempt to read such a text, whilst suggesting possible solutions to aid them in their task. This would allow them to switch effortlessly between initially competing hypotheses, allowing them to see the development of their reading of the texts and trace any conclusions back to their initial thought processes: something which is difficult to do at present.

The overall aim in this research was to aid in the transcription of the stylus tablets, but it is hoped that the system may be eventually be used by papyrologists working on other texts. Also, the techniques used would be easily adaptable to other fields: there are many applications for a computer system that can effectively reason about data contained within images. Additionally, this research has also provided a great deal of information regarding how papyrologists work. Most importantly, it demonstrates that MDL can be used effectively as a unifying method to compare and resolve different forms of complex data, and that MDL can provide the basis for a cognitive visual system.

## 1.6 Thesis Synopsis

This thesis comprises of six chapters and three appendices, which cover the following topics:

- **Chapter 1** contains background and contextual information regarding the research, plus an overview of the direction of the thesis.
- **Chapter 2** asks the question: how do experts read ancient texts? After considering all available research on this topic, this chapter details an investigation, using Knowledge Elicitation techniques, into how the experts read both the Vindolanda ink and stylus texts. The findings are rationalised into a connectionist model on which to base the development of the computer system detailed in Chapter 4.
- **Chapter 3** concerns the letter forms used within the texts. It is shown that the stylus tablets should contain similar letter forms to the ink texts. An encoding scheme is developed to capture information about the letter forms by monitoring the information discussed by the experts as they identify individual characters.

A corpus of annotated images is built up to provide a data set on which to train the system developed in Chapter 4.

- **Chapter 4** discusses the adoption of the MDL based GRAVA architecture in order to build a system to effectively "read" the stylus texts, and generate possible interpretations of the text contained within the images.
- **Chapter 5** details future work, suggesting how the developed system could be expanded, and eventually delivered to the experts as a stand-alone application to aid them in their task of reading the stylus tablets.
- **Chapter 6** summarises the findings of this research, and evaluates the success of the project.
- **Appendix A** details the encoding scheme used to annotate the images and provides an introduction to the corpus of annotated images.
- **Appendix B** contains a graphical representation of the stroke data of all the individual characters contained in the annotated corpus.
- **Appendix C** details the contents of the **CDROM**, which contains all the data sets developed, used, and discussed during this research. The CDROM also contains a means of viewing the annotated corpus of ink and stylus text images.

# CHAPTER 2

## How Do Papyrologists Read Ancient Texts?

### Knowledge Elicitation and the Papyrologist (1)

"It seems to me that translation from one language into another ... is like looking at Flemish tapestries on the wrong side; for though the figures are visible, they are full of threads that make them indistinct, and they do not show with the smoothness and brightness of the right side ... But I do not mean by this to draw the inference that no credit is to be allowed for the work of translating, for a man may employ himself in ways worse and less profitable to himself."
Cervantes, Don Quixote. (1922, p.731)

Before designing and building any tools to aid papyrologists in the reading of texts, it is a necessary requirement firstly to ask: just what does a papyrologist do when trying to read and understand an ancient text? A review was undertaken of all relevant research in this area, before an investigation was carried out to elucidate this process. Techniques borrowed from the field of Knowledge Elicitation were used to gather quantitative and qualitative information about how papyrologists work, resulting in an in-depth understanding of the ways different experts approach and reason about damaged and abraded texts. The process is resolved into defined units, with characteristics about each being documented. General procedural information is also presented. Particular issues regarding problems in reading the Vindolanda stylus texts are highlighted, indicating areas in which computational tools may be able to aid the papyrologists in reading such texts. This results in a proposed model of how experts read ancient documents which is used in subsequent chapters as a basis for the development of such a computer system.

## 2.1 Papyrology Discussed

The readings generated from ancient documents provide one of the major primary information sources for classicists, linguists, archaeologists, historians, palaeographers, and scholars from associated disciplines. Although there has been some discussion of the history of papyrology (Hunt 1930; Turner 1968; Pattie and Turner 1974; van Minnen 1995), and the contribution the transcription of such texts has made to both literary and non-literary classical studies (Hall 1913; Kenyon 1918; Winter 1936; Turner 1968; Ullman 1969; Turner 1973; Reynolds and Wilson 1991), the process entailed in transcribing a text remains opaque. Papyrology is, in essence, a "self consuming labor which leaves little or no trace of itself," (Youtie 1963, p.11) and the expertise of papyrologists, as with the expertise of any professional, is a valuable but surprisingly elusive resource. There are only a few discussions by papyrologists themselves that query the nature of papyrology. Additionally, very little research has been done in the field of Cognitive Psychology regarding how experts may read damaged and deteriorated texts, or even texts in languages other than (clearly printed) English.

### 2.1.1 Papyrologists on Papyrology

There are a handful of papers written by those working on ancient documents which attempt to analyse and relay the nature of the processes used to generate satisfactory readings; two papers by the papyrologist H. C. Youtie, a guide to the reading of cuneiform texts by the assyriologist E. Reiner, a pamphlet regarding decipherment by P. Aalto, a Victorian guide to the reading of "ancient" documents, and a forthcoming paper regarding the problems encountered in reading the ink and stylus texts by Bowman and Tomlin.

Youtie's papers, "The Papyrologist, Artificer of Fact" and "Text and Context in Transcribing Papyri" (1963, 1966) both attempt to describe the processes that the papyrologist goes through when transcribing a text; "What do papyrologists do?" (1963, p.9). Youtie demonstrates that most accounts of papyri and the reading of papyri consider the studies to which they make a contribution and not the act of transcription itself, and talks about the mechanics of papyrology – the publication, annotation, and editing of transcriptions of ancient texts. Youtie then goes on to attempt to describe what happens when he is reading such an ancient text:

> A Theban receipt is not written to be deciphered letter by letter; it presupposes a reader who has sufficient information to read intuitively (1963, p.14).

Youtie's attempts at describing this activity (although at times hyperbolic) demonstrate the complex, recursive nature of the task:

> He tries to take account of the text as a communication, as a message, as a linguistic pattern of meaning. He forms a concept of the writer's intention and uses this to aid him in transcription. As his decipherment progresses, the amount of text that he has available increases, and as this increases he may be forced to revise his idea of the meaning or direction of the entire text, and as the meaning changes for him, he may revise his readings of portions of the text which he previously thought to be well read. And so he constantly oscillates between the written text and his mental picture of its meaning, altering his view of one or of both as his expanding knowledge of them seems to make necessary. Only when they at last cover each other is he able to feel that he has solved his problem. The tension between the script and its content is then relaxed: the two have become one (Youtie 1966, p.253).

He also elucidates the difficulties involved in the process of reading such damaged and deteriorated texts:

> ... the memory of all the effort that it has cost, the doubts, the hesitations, the numerous false starts and new beginnings, the guesses sometimes confirmed, sometimes rejected by the script, the continual recourse to books for information of every sort – lexical,

> grammatical, palaeographic, historical, legal; the every threatening awareness of ... visual and intellectual inadequacy; the interludes of exhaustion and depression ... (Youtie 1963, p.17).

Ultimately, it is demonstrated that it is difficult to explain in the final annotated version of the text how the papyrologist got from papyri to transcription and translation. The complex process of transcribing a text is lost in its documentation.

Aalto's pamphlet "Notes On Methods of Decipherment of Unknown Writings and Languages" (Aalto 1945) aligns the reading of ancient texts with the work of cryptographers, demonstrating how techniques such as transposition and substitution are used in both cracking military codes, and reading ancient texts[1]. He demonstrates how different techniques can be used to understand texts where the script, language, or both are unknown, and suggests the importance of statistical information, the generation of probabilities, and the comparison of grammatical and etymological elements when trying to decipher texts. Like Youtie, he stresses the recursive nature of the task: "Every interpretation thus implies a long series of trials (and errors too!) before it can possibly be regarded as verified" (p.18).

Reiner's paper, "How We Read Cuneiform Texts" (1973), interestingly adopts "the fiction of programming a machine to read cuneiform" (p.16) to sketch a very precise and detailed model of the process of identifying and understanding the written Assyro-Babylonian language. She shows that

> the reading of a Cuneiform text is based on information which includes all three components: syllabary, grammar, and dictionary ... and also that the three components are interrelated, or in Saussure's words "tout se tient" in the system that we call language (p.15).

---

[1] Aalto discusses how experts in ancient languages from the British Museum were successfully employed in WWI as code breakers for the British Army.

The reading process is stratified into four sections;

1. The basic value look up table
2. Finding the ultimate value
3. Segmentation
4. Morphosyntactic analysis (p.7)

and Renier talks of the "generalized cross reference" between these sections (p. 22) needed to reach a conclusion regarding the text. However, although these steps may be extrapolated to cover the reading of other languages, the remainder of the paper focuses exclusively on issues regarding particular aspects of cuneiform, such as "the features of voice and emphasis in stops and sibilants" (p.25). Whilst providing an interesting introduction into how to approach such a complex language system, this paper is mostly a comparative analysis of contemporary research on Akkadian phonetics, phonemics, and morphology.

The one book which purports to explain "How to Decipher and Study Old Documents, Being a Guide to the Reading of Ancient Manuscripts" (Thoyts 1893) details how to gain access and place into historical context "ancient" texts, such as Parish registers, Deeds, and Monastic Charters. However, Thoyts shows little insight into how one may actually *read* such texts:

> Written in a language I knew not, relating to customs no longer existing, all was strange and unfamiliar. I toiled on, by degrees light dawned and the difficulties melted away. (p.xi).

Thoyts' conclusion indicates the paucity of advice she gives regarding how to approach and reason about documents;

> 'Persevere and Practice' is the best motto I can give to those interested in the matter, for proficiency comes quickly to those who seek it, and, as in all subjects, 'Nothing succeeds like success' (p. 143).

Although an intriguing text regarding the growth of antiquarianism in the late 19th Century, this guide offers little in the understanding of how experts read, transcribe, and understand ancient, often damaged, texts, never stretching beyond the obvious: "a transcriber's work properly consists chiefly in correctly putting into modern handwriting the deeds which are only illegible to the uninitiated" (p.9).

In contrast, Bowman and Tomlin's forthcoming paper (2003) details the experiences they have had over the past quarter century or more in reading damaged and abraded texts, particularly the wooden stylus tablets from Vindolanda, providing the most comprehensive illustration of the exercise to date. They provide some introspection and analysis of their own working processes:

> First we identify the shapes of letter forms, which fall into a range of types and different hands, then we read individual letters and combinations of letters to the point where we can construct words, phrases, sentences and finally whole texts. But of course we do not normally transcribe letter by letter in a completely neutral and automaton-like fashion, and then realise that we have transcribed a word or sentence; there is a point, very quickly or perhaps immediately reached, at which we bring into play our corpus of acquired linguistic, palaeographical and historical information which in effect predisposes or even forces us to predict how we will identify, restore or articulate letters or groups of letters. And this is a recursive process. We do the easy bits, then make hypotheses about the problematic bits and test them in the context which we think we have established by what we can read (p. 3).

An account of how they read and formulated a meaning of Vindolanda Tablet 974 is presented, giving an example of this process. They show how the conclusions they draw rely on parallels with other texts, and how the accumulation of textual knowledge from other similar documents can improve their reading of each (the expanding knowledge base regarding Vindolanda aiding the reading of each

subsequent text encountered). Reading an ancient document is then "not a letter-by-letter transcription like sleep walkers, but a wakeful testing of possibilities in the light of other knowledge" (p.7).

The only other description of the nature of the process of papyrology is to be found in a text by Turner (1968), and whilst he, again, describes the mechanics of the papyrology in describing the Leiden system[2] and publishing formats, his attempts at describing the actual act of transcribing and translating such ancient texts descends into romanticism[3]:

> What does a papyrologist do? He is engaged in eliciting new knowledge and doing his best to guarantee the reliability of that knowledge. His proper and professional field of competence is in relation to the text themselves and what they say. Curiosity and excitement will have started him on his quest; a passion for truth will bring him to its conclusion (Turner 1968).

## 2.1.2 Psychology and Papyrology

How papyrologists read ancient texts, as well as being seldom addressed by the experts themselves, has not been the focus of any psychological study. The field of Psycholinguistics is dominated by more mainstream questions regarding language (for example the acquisition of language, the relationship of linguistic knowledge to language usage, and the comprehension and production of speech (Aitchison 1998)) rather than the mechanics of such a complex task. Although there is a growing

---

[2] A series of symbols which denote various characteristics of the original text, see 2.5.3.3.

[3] For the act of reading and translating ancient documents as wistful subject matter for the arts see Tom Stoppard's play "The Invention of Love" (1997), or Lesley Saunders' poem "The Uses of Greek" (1999).

interest in translation studies[4] in the humanities, there has been little work done on the psychology of translation[5] (papyrology differing from translation on many levels anyway[6]). The Psycho-linguistic studies which do exist regarding the use of a second language focus on the psychological, social and educational issues surrounding bilingualism in individuals, or the learning of a second language. Studies on how expert readers access the separate lexicon of a second language is in its infancy (Hornby 1977; de Groot and Barry 1994). Most research done on linguistics concerns the acquisition and use of the English Language, and the research that does exist "suggests that differences may exist in the reading processes between languages with different writing systems, orthography, morphology, and syntax" (Asher 1994, p.3461). This would mean that the large structural differences between, say, English and Latin[7], would render the reading of a Latin document by a native English speaker a very different process than the reading of an English text

---

[4] The emergent field of translation studies (Venuti 2000) addresses questions regarding the production of "a target language version of the source language text" (Danks and Griffin 1997, p.164).

[5] Cognitive psychologists have not explored straight translation thoroughly never mind how experts deal with the reading of degraded and difficult ancient texts. This may be partly because they are not really aware of the emergent discipline of Translation Studies, or that those in the field who have recognised its relevance may have given priority to the study of simpler tasks regarding how humans read, of which there is scant understanding. Solutions to those may provide the answer to further questions regarding the cognitive processes involved in translation (Groot 1997, p.29).

[6] All the surveyed research in translation studies assumes that there is no problem in the actual *reading* of the text whilst comparatively studying texts and their translations (Carr, Durand et al. 1994). Although some of the techniques used in Translation Studies can be adopted in analysing the procedures papyrologists use, the focus of the two disciplines is different, as papyrologists primarily aim to deliver a transcription of the text in its native language. As Youtie stated, "we do of course translate the texts that we derive from papyri but ... we must first obtain the texts" (Youtie 1963, p.9). Undoubtedly there will be some shared cognitive processes between the two fields, but Translation Studies has little to offer at present in understanding how papyrologists read ancient texts.

[7] Although English derives many words from Latin, the structure of the languages differ widely (Ellegard 1963). English is much more dependant on word order than Latin, which utilises word endings more as indication of case, Latin words reflecting their purpose by the altering of their structure. (Morwood and Warman 1990, p 46) The Latin concept of gender in language is entirely different that of English with nouns being assigned gender specific traits. Latin makes more use of noun and adjective stems and less use of adverbs than English. Latin uses an entirely different case structure than English. (Conway 1923, 84-106) Put succinctly: "You cannot apply the rules of Latin grammar to English." (Morwood and Warman 1990, p 48)

by the same reader. In turn, most current research on reading would therefore be inappropriate to this thesis.

Research which specifically addresses the resolution of ambiguity in texts, whether that is lexical (Hogaboam and Perfetti 1975; Swinney and Hakes 1976), or structural (Hirst 1987) concurs on one thing: that there is no agreement on the principles which readers use for disambiguation when reading texts. There has, however, been a great deal of research done on how readers carry out simple reading tasks, and this is covered in 2.8.1.

### 2.1.3 Papyrology Undiscovered

Thus, although ancient manuscripts provide a primary source for historians, linguists, and others, very little attention until recently has been given to the process entailed in deciphering such texts. When it has been discussed, no attempt has been made to make explicit the reading process, the accounts being discursive rather than quantitative. There has been no comparison made of the differences and similarities between the work of individual papyrologists. Ultimately, when the readings of ancient documents are published, there is a focus on the texts themselves rather than the processes undertaken to read them:

> The general accounts, the surveys, the reports, whatever they are called, tell us nothing about the work done by the papyrologist. They all take up where he leaves off. They talk about papyri as they are after the papyrologist has finished with them, when he has already completed his transcriptions, added his philological and sometimes historical commentaries, and made them available in learned journals or volumes of papyri to any specialists who have a use for them (Youtie 1963, p.11).

Youtie terms this documentation "public papyrology". To be able to discover and describe the process of "private" papyrology ("what a papyrologist does privately, in the solitary confines of the library, in order to make public papyrology possible" (p. 11)), a detailed study of how individual papyrologists work was undertaken. "Knowledge Elicitation" techniques, frequently used in the fields of Computing and Engineering Science to enable engineers to capture expertise and encapsulate it into computer systems, were used to elucidate this process.

## 2.2 Knowledge Elicitation: A Brief Guide

The problem with trying to discover the process that papyrologists go through whilst reading an ancient text is that experts are notoriously bad at describing what they are expert at (McGraw and Harbison-Briggs 1989). Experts utilise and develop many skills which become automated and so they are increasingly unable to explain their behaviour, resulting in the troublesome "knowledge engineering paradox": the more competent domain experts become, the less able they are to describe the knowledge they use to solve problems (Waterman 1986). Added to this problem is the fact that, although knowledge acquisition and elicitation[8] are becoming increasingly necessary for the development of computer systems, there is no consensus within the field as to the best way to proceed in undertaking such a study.

Discussions regarding how best to elicit knowledge for the basis of an expert system first appeared in the late 1970s (Feigenbaum 1977). Early attempts at eliciting,

[8] Knowledge acquisition is conventionally defined as the gathering of information from any source. Knowledge elicitation is the subtask of gathering knowledge from a domain expert. (Shadbolt and Burton 1990)

formalising, and refining expert knowledge were so unfruitful that Knowledge Elicitation was labelled the "bottleneck" to building knowledge-based systems (Feigenbaum 1977). Throughout the 1980s and early 1990s, protocols (often referred to as the "traditional" or "transfer" approach to Knowledge Elicitation) were developed regarding how a knowledge engineer should interact with a domain expert to organise and formalise extracted knowledge so that it is suitable for processing by a knowledge based system (Diaper 1989; McGraw and Harbison-Briggs 1989; Boose and Gaines 1990; McGraw and Westphal 1990; Morik, Wrobel et al. 1993). These discussions centre on the suitability of different techniques (often derived from clinical psychology and qualitative research methods used in the social sciences) for the capture of knowledge, such as:

- Unstructured, semi-structured, and focussed interviews with the expert(s).
- Think Aloud Protocols (TAPs), where an expert is set a task and asked to describe their actions and thought processes, stage by stage (see 2.3.2).
- Sorting, where the expert is asked to express the relationship between a pre-selected set of concepts in the domain.
- Laddering, where the expert is asked to explain the hierarchical nature of concepts within the domain.
- The construction of Repertory Grids, where concepts are defined by the way they are alike or different from other related concepts in the domain by comparing and contrasting values (see 3.2.3).

A discussion contrasting these with other knowledge acquisition techniques, highlighting their suitability regarding the elicitation of certain types of knowledge, can be found in Cordingley (1989).

Since the early 1990s the field has focussed on Automated Knowledge Elicitation, incorporating the same psychological techniques into computer programs, to make the interactions more productive, assisting, and in some cases replacing, the knowledge engineer (White 2000). Such tools were, at first, implemented in a stand-alone, domain independent way, focussing on the collection of particular types of data. For example, the ETS and the AQUINAS systems (Boose 1990) are computerised representations of the Repertory Grid method and have been used to derive "hundreds" of small knowledge-based systems. Gradually, programs appeared offering implementations of various techniques bundled together as a Knowledge Elicitation "workbench" such as the research prototype ProtoKEW (Reichgelt and Shadbolt 1992) which was later repackaged and marketed as the commercial PC-Pack system. Researchers have now started to utilise the internet, developing distributed knowledge acquisition tools such as RepGrid[9] and WebGrid[10], which can be used remotely to build up knowledge bases and data sets. However, these computer tools produce the best results when applied to very small domains to build knowledge based systems which carry out well defined tasks, and are not successful at providing overviews of complex systems, or when used to describe domains about which little is known from the outset (Marcus 1988; White 2000).

Although much research has therefore gone into the different individual techniques and tools which can be used to elicit knowledge from an expert, "no single knowledge acquisition standard has emerged" as McGraw and Harbison-Briggs are keen to point out (1989, p.52). There are actually few useful guides on how to

[9] http://www.csd.abdn.ac.uk/~swhite/repgrid/repgrid.html
[10] http://tiger.cpsc.ucalgary.ca/WebGrid/WebGrid.html

undertake such an exercise from conception to completion, and how to choose the best selection of tools suited to the domain being examined (Dermott 1988). The knowledge engineer is left to try and ascertain how best to identify, collect, and rationalise any sources of knowledge, use any computational tools, interact with the experts, and develop and check any conclusions reached regarding the domain. External factors, such as amount of time available, amount of relevant external source material, and the availability of the experts, all affect the process. The scope of the study also affects the tools and techniques chosen: is the knowledge engineer trying to gain an understanding of an overall process, or an intricate understanding of a small task? Knowledge Elicitation remains a complex, time-consuming process, and the choice of method and technique is specific to each domain investigated.

## 2.3 Knowledge Elicitation and Vindolanda

The primary questions to be asked in this study were: is there a general process that experts use when reading ancient texts? Can this procedure be elucidated? Additionally, what are the differences and similarities between individual experts' approaches to the problem? Also, although it has been noted that "cognitive and knowledge-based problems ... are in general common to ink texts and inscribed texts" (Bowman and Tomlin Forthcoming 2003, p.3), how does the format and medium of a document affect the reading process? Finally, how does the provided documentation relate to the reading of an ancient text, and can it indicate anything about the process?

A number of steps and observations were undertaken to gain an understanding of the general process the experts utilise when approaching an ancient text, and specifically, the Vindolanda ink and stylus tablets. Firstly, as with all knowledge acquisition tasks, the domain literature was researched, as reviewed above in 2.1. Secondly, any other associated literature was collated. Although not a direct comment on the act of reading and transcribing, the two published volumes regarding the Vindolanda ink tablets contain detailed apparatus of the individual texts (Bowman and Thomas 1983; Bowman and Thomas 1994). The standard publication format (the transcribed text, marked up using the Leiden system, followed by the critical apparatus: variant readings, misspellings, line by line comments and explanations, and the translation of the text)[11], is all that is presented of the process which was undertaken in the reading and understanding of the documents. This apparatus aims to cover comprehensively the difficulties, reasoning, and alternative hypotheses regarding the final transcription. As such, it

[11] As an example of this format, here is the published commentary of Stylus Tablet 836, by Bowman and Tomlin (Forthcoming, 2003, p. 6). In the text presented below, letters printed in **boldface** are those which can be read with confidence; letters printed in ordinary type are read with some measure of conjecture; underlinings indicate traces of letters which cannot be identified with confidence:

**banus bello suo sal**utem
(traces only)
acc**__erunt in in uecturas**
**de_arios oc**to **reliquos** solues
**rios nouem qua__r_r___**
**sam dari debeb__**
(interlinear addition?)
**em libris**
**dus uale**

'Albanus to his Bellus greetings ... they have received for transport costs 8 denarii. You will pay the remaining 9 denarii ... ought to be given (?) ... nine pounds (?) ... Farewell.'

Notes:
1. There is a trace between the first and second **l** in **bello** which might or might not be a letter. The scratches on the wood show that this overlies an earlier text.
2. The correct reading is almost certainly **acceperunt**.
3. The word at the end of the line presents particular difficulty. Of the first three letters of **solues** only the **o** is certain. There is a clear high horizontal which has to be ignored if the first letter is read as **s**. The third letter might be **p**, and there is another apparent high horizontal which is

is the best representation of the different types of knowledge used in the reading of texts available without carrying out further investigation (although the sequential order of the different stages in their reading are lost due to the reporting format). These texts were obtained in digital format to enable in-depth study.

Three experts were then identified who were working on the ink and stylus texts, and who were willing to take part in this investigation. (They shall be referred to as Expert A, Expert B, and Expert C, so as to spare their blushes, and not to imply any criticism of their work: this investigation is concerned with asking how an expert operates, rather than making judgements on those operations.) Expert A works equally on the ink and stylus texts from Vindolanda. Expert B works mostly on the stylus tablets, and other similar incised texts from the period. Expert C's studies mainly focus on the ink tablets, and other similar texts from the period. All three experts are English male academics who have been working on such texts for over twenty-five years. They graciously gave their time, and permission, for this study.

A series of investigations were carried out, utilising knowledge elicitation techniques, particularly Think Aloud Protocols. The data captured provides explicit and quantitative representation of the way the papyrologists approach damaged and abraded texts. General procedural information was also collated.

## 2.3.1 First Stages in Knowledge Elicitation

The first stages of knowledge elicitation were fairly informal. The experts were observed whilst going about their tasks, and unstructured interviews were

---

discounted. The attraction of reading the word **solues** (from the verb **soluere** 'to pay) is obvious

undertaken, where the experts described their domain, and the individual processes and techniques that they preferred. More structured interviews were then undertaken, when the experts were asked to describe particular facets of their task, such as the identification of letter forms (which is the focus of Chapter 3 in this thesis), and the role of grammar, word lists, and external historical and archaeological resources in the reading of the documents. Joint sessions, where Expert A and Expert B discussed their readings of the stylus texts, were also attended. These sessions were documented, and a broad understanding of papyrology was formed, whilst building up a good working relationship with the experts.

### 2.3.2 Think Aloud Protocols

However, a more formal approach was needed to build up some quantitative data on the reading of such texts. It was decided to undertake a series of Think Aloud Protocols (TAPs) a technique adopted from experimental psychology, where the expert is urged to utter every thought that comes to mind whilst undertaking a specified task. These sessions are recorded, transcribed, and analysed to allow the observer to identify the different steps characterising conceptual processes more precisely. Although "an expensive and meticulous research method that has had its share of growing pains" (Smagorinsky 1989, p.475), the collection of verbal data in this manner has been a procedure used in the social sciences for three quarters of a century (Duncker 1926). Protocol analysis has been shown to be

> ... a very useful addition to the repertoire of research tools ... The data from most other tools yield little about the internal structures of cognitive processes, particularly when the

---

if the word '**denarios**' occurs twice in lines 4-5.

> tasks are complex. Think-aloud protocols, in contrast, can yield significant information about the structure of processes (Smagorinsky 1989, p.465).

Ericsson and Simon (1993) show that verbalisation does not interfere with the cognitive processes discussed, and that there is little difference between results whether the information is collected concurrently (whilst the expert undertakes a task) or retrospectively (when the expert details what they did whilst undertaking a task). Most cognitive studies of translation and interpreting use TAPs as the tool of choice (Danks, Shreve et al. 1997, p.xv), for example Kiraly (1997) effectively used TAPs to investigate how translators work.

The three experts were set various tasks to gain an insight into how they approach and reason about the ink and stylus texts. It was explained to each the nature of the Think Aloud Protocols, and how and why the experiment was going to be carried out[12]. For the most part, these TAPs took place in the workplace of the experts, where they would usually work on such texts. Various data was collected:

- All three experts were given a pack containing various images of tablet 1543[13], and asked to come up with the best reading they could at their own leisure. In the session they were asked to talk through how they arrived at their reading, and to describe the processes they undertook whilst coming up with their final text.

[12] Available on the CD-ROM in Chapter 2/Think Aloud Protocols/Instructions/.

**Figure 2.1:** Ink Tablet 1543. A larger image is available on the accompanying CD-ROM.

➢ All three experts were presented, on the day, with an image of an ink text (1491). They had not been asked to prepare this text, and so discussed how they would go about tackling a text from the outset. (This text was identified as being one of the ink texts the experts would be least familiar with. In actuality, Expert A and Expert C had some prior knowledge of this text. Expert B had never seen this text before). This provided data to compare how experts reason about texts they have had the time to prepare, and those they have had little experience with.

[13] Larger images of the four texts used in the Think Aloud Protocols are available on the accompanying CD-ROM in Chapter 2/Think Aloud Protocols/Images/.

**Figure 2.2:** Ink Tablet 1491. A larger image is available on the accompanying CD-ROM.

- Expert B was asked to talk through his reading of 1491 again at a later date, to describe how he had read the document. He had not looked again at the ink text in the interim. This was done to investigate how retrospectively talking about the reading of texts differs from the concurrent explanation of the process.
- The experts who deal with stylus texts (Expert A and B) were given various images of two stylus texts, 1593 and 797, from which they tried to obtain readings. They were asked to discuss these in full, to provide data on how reading the stylus tablets differs from reading the ink tablets.

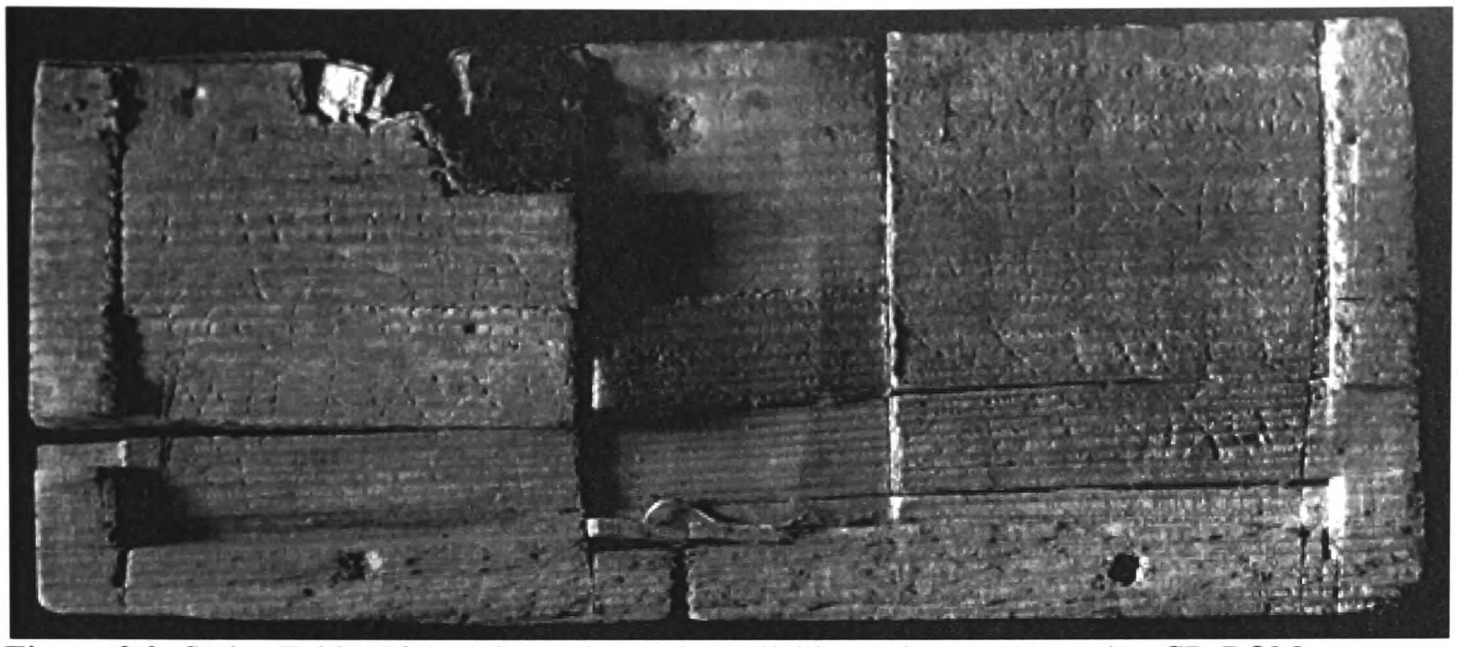


**Figure 2.3:** Stylus Tablet 797. A larger image is available on the accompanying CD-ROM.

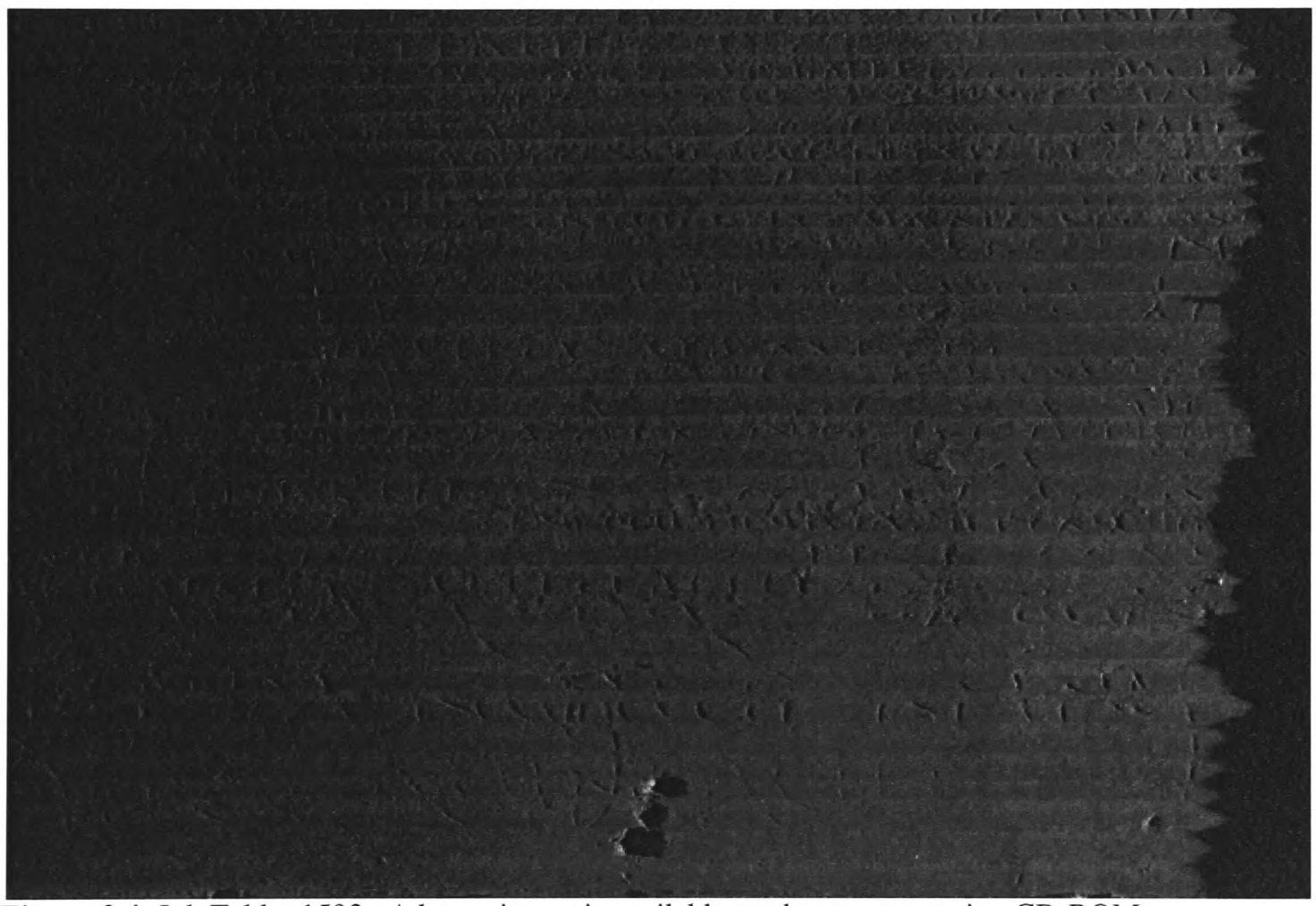


**Figure 2.4:** Ink Tablet 1593. A larger image is available on the accompanying CD-ROM.

These sessions were taped, and transcribed verbatim, giving a large set of data (23,000 words of discussion; 16,000 words pertaining to the reading of the ink tablets, 7,000 regarding the reading of the stylus tablets)[14].

[14] These transcripts are available on the accompanying CD-ROM in Chapter2/think aloud protocols/.

## 2.4 Associated Techniques: Content and Textual Analysis

The majority of the data collected during this investigation was textual, due to the standard way of documenting Think Aloud Protocols, and the format of the Vindolanda text apparatus. Various techniques have been developed in the social sciences and the humanities to analyse such data, the two most important being Content and Textual Analysis.

### 2.4.1 Content Analysis and Vindolanda

Content Analysis, a "method of studying and analyzing written (or oral) communications in a systematic, objective, and quantitative manner" (Aiken 1971, p.433) is an unobtrusive, context sensitive, empirical process in which texts are reduced and condensed into a numerical format in order to estimate some phenomenon in the context of data (Holsti 1969; Krippendorff 1980; Stemler 2001). It has been used extensively since the 1930's in the analysis of large volumes of data, such as textbooks, comic strips, speeches, advertising, and the psychological analysis of personal communication, and is a large field in Communication Studies. Although Content Analysis has its problems with reliability and statistical validity (Franzosi 1990), has been accused of being subjective (Anderson 1974), and can be time consuming and laborious (Aiken 1971, p.279), it remains the most thorough and useful way to gain an empirical insight into the structure and content of complex textual sources, and is the standard technique used in encoding and analysing the data from Think Aloud Protocols

(Ericsson and Simon 1993). Texts are divided into defined units[15] and labelled, and these units can be used as a basis for statistical analysis.

In the case of the data from Vindolanda, the *subject* of the linguistic unit was identified as the defining feature, and the text split into sections where the subject of the phrase, sentence, or sometimes paragraph, changed. A number of pilot studies were undertaken before settling on an inventive encoding scheme which comfortably encompassed the data from the apparatus, transcripts, and preliminary knowledge elicitation exercises, resolving the different types of knowledge presented into an overall framework. The final, novel encoding scheme is presented below.

| Reading Level | Thematic Subject |
|---|---|
| 8 | Meaning or sense of document as a whole |
| 7 | Meaning or sense of a group or phrase or words |
| 6 | Meaning or sense of a word |
| 5 | Discussion of grammar |
| 4 | Identification of possible word or morphemic unit |
| 3 | Identification of sequence of characters |
| 2 | Identification of possible character |
| 1 | Discussion of features of character |
| 0 | Discussion of physical attributes of the document |
| -1 | Archaeological or historical context[16] |

**Table 2.1:** The encoding scheme resolved from an analysis of the Vindolanda textual data, an original scheme devised for this project.

[15] There are many different ways in which to break down a text in Content Analysis, such as word, sentence, paragraph, item, or theme (Holsti 1969, p.180).

[16] This is presented as "-1" to mark the fact that the experts are explicitly referring to other sources, and not only this document.

The Think Aloud Protocols were encoded in this manner, the time spent (in seconds) discussing each different subject also being recorded as well as the amount of words spoken.

A selection of critical apparatus from *Tabulae Vindolandenses II* (Bowman and Thomas 1994) were also analysed. It was important that the critical apparatus which were analysed contained enough information (some of the published texts are very fragmentary) to be useful, and contained the type of text that would be relevant; there was no point in comparing official army documentation with, say, literary texts. For this reason the texts were not selected randomly[17] from the publication, but the authors were consulted to see which texts would be most significant in relation to the task. Knowledge elicitation techniques have shown that asking experts to recall the five or ten most important instances of their work can yield the best examples of the process (Morik, Wrobel et al. 1993), and so one of the experts was asked to point out the most relevant and important texts from their publication. Seven of the critical discussions were analysed.

Because of the nature of the project there was only one person encoding the texts, leading to problems with validity. To increase the reliability of the analysis each text was encoded in such a way twice, and the two resulting spreadsheets compared to highlight any problematic areas, which were then re-analysed to make sure that the system was consistent and therefore reliable. All texts were encoded before any analysis of the data took place. Such techniques are the best way to ensure the

[17] It can be argued that the selection of texts that are published are quite statistically random due to the nature of their survival.

validity of results when relying on one researcher for the Content Analysis of texts (Stemler 2001).

## 2.4.2 Textual Analysis and Vindolanda

The textual data from both the Think Aloud Protocols and the published commentaries were also analysed using a common Corpus Linguistics program for lexical analysis, WordSmith[18], in order to look for patterns in the language used whilst discussing the documents. Such tools have been used in Corpus Linguistics since the 1970s (Sinclair 1991; Lawler and Dry 1998), and can be used for both quantitative and qualitative linguistic investigations (Popping 2000). The generation of word lists, frequencies, collocates, concordances, and key words can "unpack the political, social, and cultural implications of texts" (Hoey 2001, p.3). For example, textual analysis has been used to investigate the occurrence of sexist language in children's literature, and how language spoken in the court room can affect the outcome of a case (Stubbs 1996). In respect to Vindolanda, these tools were used to look for linguistic differences between the experts, and to analyse the order in which they undertook discussions. The differences in language usage regarding the different tasks set was investigated, to see if this could aid in the investigation of how the experts work. These tools were also used to highlight information regarding different individual letter forms, as discussed in 3.2.1.

No Automated Knowledge Elicitation tools were used in trying to gain an understanding of how the papyrologists worked. This was because the task investigated was too amorphous to be suited to any of the tools currently available,

[18] http://www1.oup.co.uk/elt/catalogue/Multimedia/WordSmithTools3.0/

which require well defined parameters regarding small domains to be successful (White 2000). However, a Repertory Grid tool was used to collect and analyse data regarding the letter forms of Vindolanda, as discussed in 3.2.3.

## 2.5 Results

### 2.5.1 The Technique of Individual Experts Compared

It is perhaps unsurprising that, superficially, the three experts used in this study seem to read documents differently. The individual tools and techniques they prefer differ greatly; Expert A makes most use of digital images and PhotoShop to examine the texts, Expert B favours drawing his own representation of the text as he reads, Expert C relies mostly on photographs. The three experts spent various amounts of time discussing the texts, and the word counts of these discussions also varied greatly. The speed which they talked differed a great deal, with Expert B and C more prone to periods of silence whilst they considered the documents.

| Expert | Average Word Count per Document | Average Time per Document (seconds) | Average Words per Second |
|---|---|---|---|
| A | 2187 | 417 | 5.24 |
| B | 3205 | 1109 | 2.89 |
| C | 1622 | 947 | 1.71 |

**Table 2.2:** Comparison of different individuals from transcripts of discussions. These results vary considerably, and suggest that the experts work differently. However, it is subsequently shown that there are underlying similarities to their techniques.

Each individual also had their own linguistic style in discussing the texts, which can be seen in the differences in language that they use[19], and illuminated by providing examples of their reading of the word *adfectum* in tablet 1491.

- **Expert A** described readings that he was very certain of, and described the final conclusion of the texts firmly. The words which he used most often which differentiated him from the other experts were CLEARLY, PROBABLY, CLEAR, LOOKING, ABLE, HAVING, REFERENCE, FOLLOWING, and REMEMBER, showing a very concrete approach to demonstrating what he could actually see in the document. For example:

  There is a reference to ADFECTUM. And then ANIMUS MEUS, and clearly it's talking about emotional matters. The last two lines read HUNC ENIM ADFECTUM ANIMUS and then something (Expert A, 1491).

- **Expert B** was comfortable talking about the reasoning process he went through, and could detail very closely the problems he encountered. He was much more likely to raise different possibilities and conclusions from the smallest change in letter forms, quickly changing hypothesis (akin to the "rapid prototyping" model of software development). The words he used most often were SORT, SEEMS, SUPPOSE, SEQUENCE, HYPOTHESIS and ASSUMING, indicating that he was concerned with the generation of different conclusions as he read a text:

  A letter to be sure either a B or a D followed by a FEC. So it is pretty certain A D F E C, ADFEC. Which is a verbal part of ADFECERE, to affect. Then one would either see if it continues into the next space, and the next line is pretty clearly TUM. So we have ADFECTUM (Expert B, 1491).

- **Expert C** was much more cautious in making assumptions, and details his hypothesis by stressing the uncertainty of his readings at all times. The words

[19] The complete word lists are available on the accompanying CD-ROM in Chapter2/comparison of individuals/.

he most often uses which differentiates him from the others are IMMEDIATELY, SUGGESTS, WILL, VERB, INSTRUCTION. Again, this indicates a fairly certain attitude to the discussion of what he can and cannot see, and also the fact that he discusses grammar much more than the other two experts:

> So then we go straight away and read the third line. You can straight away read ADFECTUM. And that's alright, masculine singular accusative so that's easy enough. What ADFECTUS means ... it's a pretty vague word, and it refers to various sorts of moods and possibly it may refer to physical illness or mental illness or something like that. Depression. Anyway. That's not a matter of reading, the problem is a matter of reasoning, ADFECTUM is a problem word. The reading is clear, HUNC ENIM ADFECTUM (Expert C, 1491).

The order in which experts discussed key features of the documents and identified constituent words also varied greatly. For example, whilst discussing tablet 1491, the identification of the words in the document happened in very different orders. This can be illustrated by plotting where three words from the text, *adfectum*, *fecit*, and *qua*, were used (and therefore identified) in the course of the discussion.

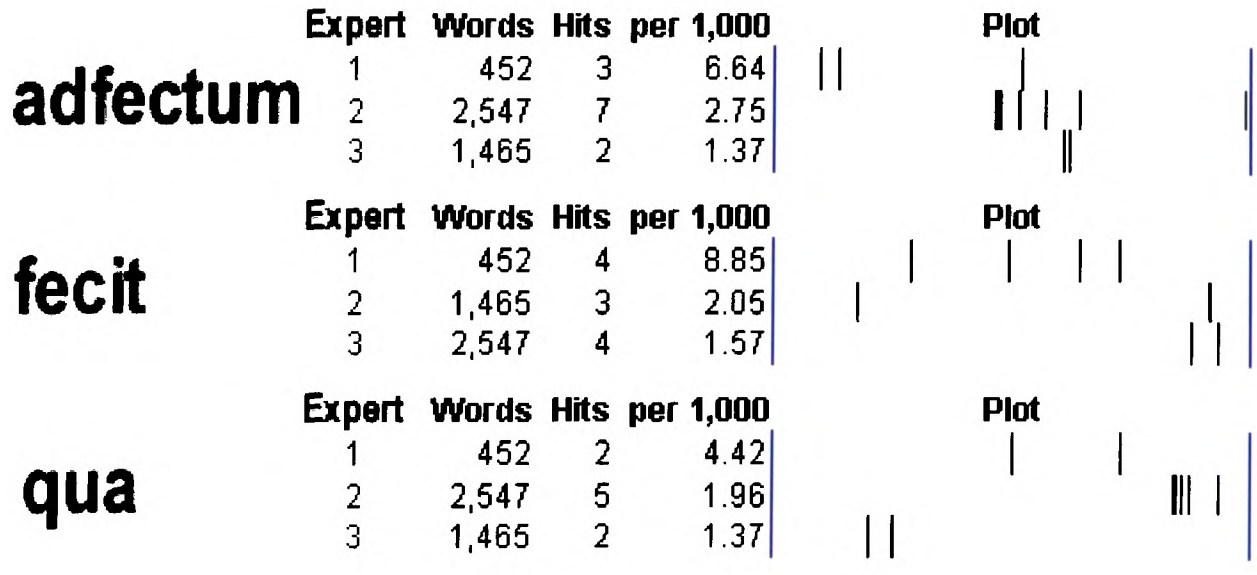



**Figure 2.5:** Occurrence of key words of document in discussion of ink tablet 1491 by three experts.

It can clearly be seen that the experts identify the words in different orders (Expert A: *adfectum* → *fecit* → *qua*, Expert B: *fecit* → *adfectum* → *qua*, Expert C: *qua* → *adfectum* → *fecit*) and at different times in the discussion.

However, this difference in order is hardly surprising when it becomes clear that when an individual expert is asked to discuss a text on two separate occasions, the order in which he discusses the content varies considerably (although there does not seem to be any order to this: he does not present them in a linear fashion as they appear on the document on the second reading):

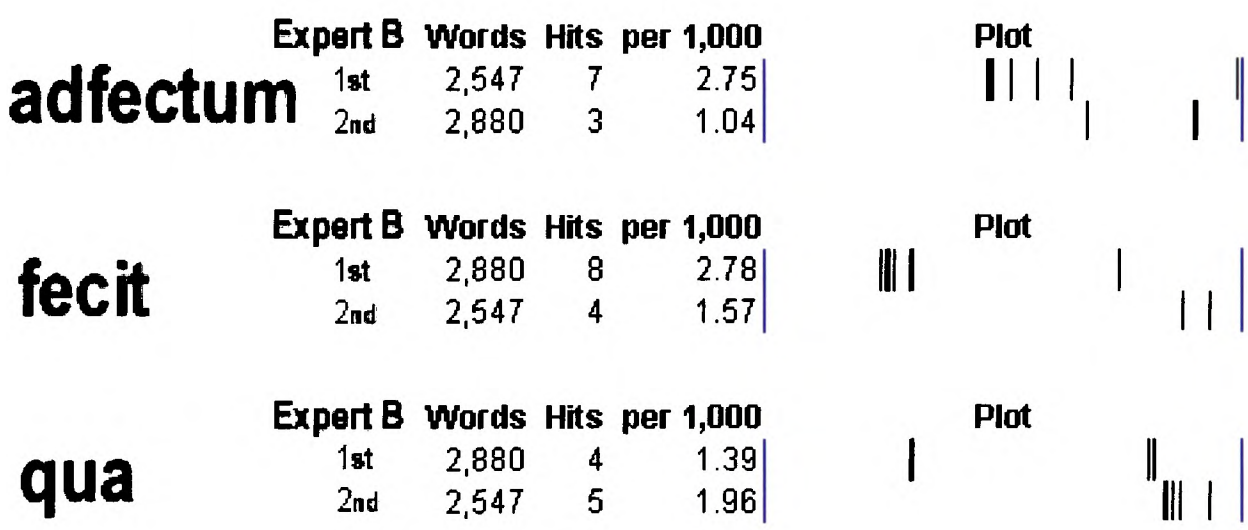



**Figure 2.6:** Occurrence of key words of document in discussion of ink tablet 1491 by the same expert on two separate occasions.

Although on the surface, the three individual papyrologists seem to discuss these texts in different manners, using different techniques, and in different orders, the readings which emerged from these different sessions were, for the most part, converging towards the same conclusions, and highlighting the same areas of difficulty. For example, compare the three experts' readings of 1491:

- And we have got something dot MAE… FECIT QUA ME something CUNDE CONSOLARIS SIC UT MATER ET HUNC ENIM ADFECTUM. (Expert A)

- FECIT QUA ME ... and then in the next line what I was reading as FACUNDE but actually may be ECUNDE... And then this CONSOLA word, And then another letter, And then what looks like UT MATER ... Then what I think is this sequence NUNC ENIM ADFECTUM ANIMUS ME. (Expert B)
- FECIT ... QUA ME ... before that we have got MA possibly MINA... So you then pick up the individual letters C U N D E... that could be CONSOLARIS ... SICUT... And then MATER looks good, Then this could be something from the verb FACEO... This ...reading is clear, HUNC ENIM ADFECTUM ... ANIMUS. And then you have M E and what looks like part of a U. (Expert C).

All three experts show the same confusion surrounding the reading of the word(s) prior to FECIT, the word which may be CUNDE, and the letters around SICUT. All clearly read FECIT, had some difficulties with the sequence QUA ME, and wondered about the possible meaning of ADFECTUM whilst clearly reading the words around it. Further analysis of the Think Aloud Protocols, utilising Content Analysis techniques, revealed hidden similarities between the experts' processes which illustrate the similarities in the way they reason about the texts.

## 2.5.2 The Cyclic Reasoning Process

A theme develops throughout the literature regarding papyrology: Youtie's "oscillation" between different stages, Aalto's series of trials and errors, Reiner's "interrelated" components and "generalized cross referencing", Bowman and Tomlin's "recursive process", and even Thoyts' "toiling on by degrees", indicating a cyclical process in transcribing and understanding the text under scrutiny. This can be shown where the experts discuss the overall process they use when transcribing a text. It is not a process of transcribing letter by letter (as Youtie,

Aalto, and Bowman and Tomlin noted), but rather of proposing hypotheses and reasoning about these as more information comes to light:

- Some people when faced with something like this will start by saying well we can identify 1,2,3,4, lines of writing here and start at the beginning of the first line and work their way through 'til the end of the last line and I suppose the idea is that you get some sort of objective view of the writing or build up of the letter, letter by letter, but I don't work that way, I never have and I don't believe that it really works very well. So what I have actually done, as I do with all these things is really go the point at which I think I can identify some of the letters in some of the words to start with and that begins to give me some sort of clue (Expert A, 1543).

- I suppose what I would do would be try and do a letter by letter transcription and start seeing if anything was making sense ... this is going to be a series of interlocking hypotheses which don't necessarily resolve themselves. In a way one is doing what one often does do, which is to at this point you are sort of on the hypothesis, but you can't really be sure that it is so until you find something that kind of makes such obvious sense, that it must be right, whereas the first two lines or so do seem to work and are self contained. I don't know if you ever have tried life drawing, but its often that you draw part of the figure and it all fits beautifully. Then you find maddeningly that the foot or something is too close, and you kind of doing what kind of requires a strength of character which is to redraw the good bit, to make it fit in with the other bit, otherwise you lose the proportion, the relationship of the whole, so that it is always the problem in reading a text; how long you hold onto something that you are certain of, if it just won't fit in with anything else (Expert B, 1491).
- I could make out individual letters at first. What we're trying to do, or what I'm trying to do, is get words to make sense ... not individual letters ... (Expert C, 1543).

The identification of the core subjects covered in these discussions (Table 2.1) shows a novel scheme that can be used to illustrate the fact that discussions regarding texts vacillate between the identification of features, letters, and words, and the production of meaning regarding these components. When plotted over

time it becomes obvious that the reasoning process is far from linear, and depends on a complex cycle of interlocking elements. This can be seen in Expert A's discussion of ink tablet 1491. Similar graphs result when any of the discussions are plotted in this manner.

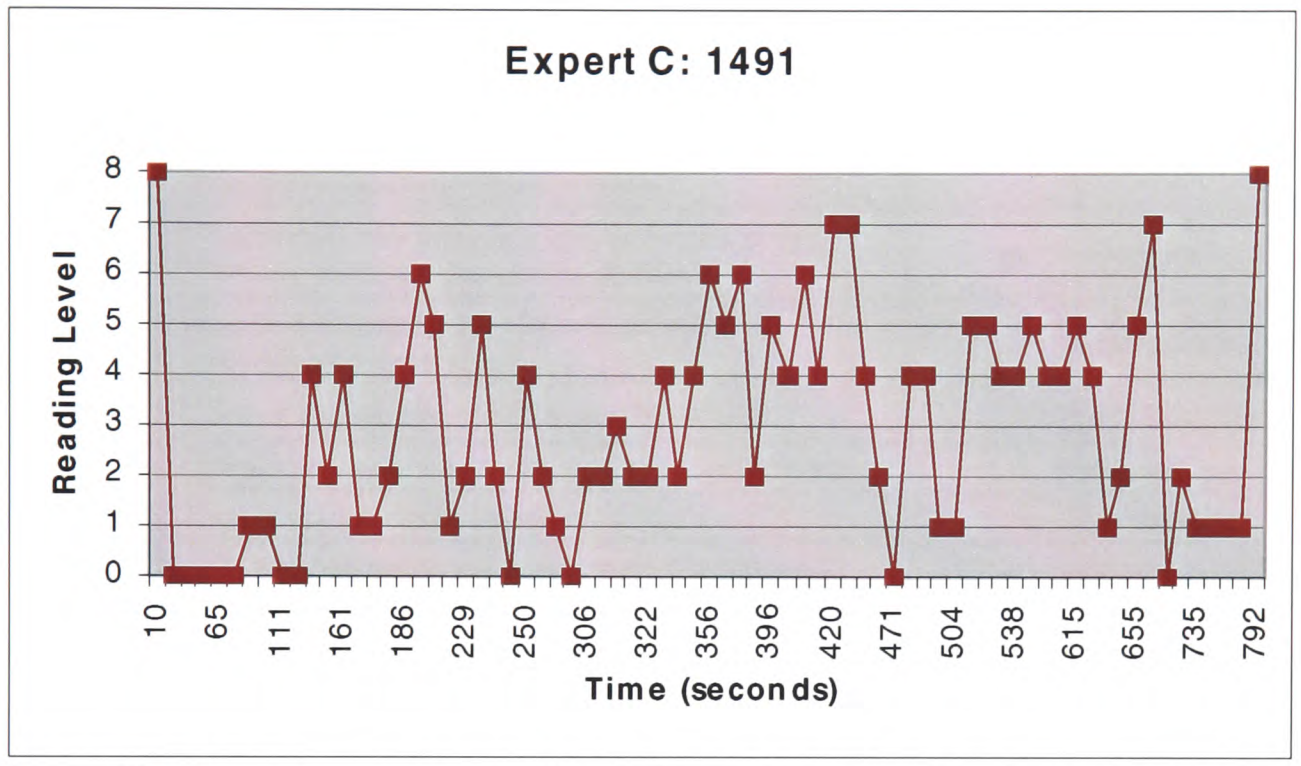



**Figure 2.7:** Discussion of tablet 1491 by Expert C, plotted by subject matter over period of discussion.

Expert C begins by drawing some conclusions about the meaning of the document (level 8) before looking its the physical attributes (level 0). He then discusses what could be possible features of the text (level 1), before noting more physical attributes of the document (level 0). He then produces a word (level 4), looks at the characters within this word (level 2), and revises his initial word. Checking of the features (level 1) leads to identification of a character (level 2), the noting of a possible word (level 4) and a discussion of meaning of that word (level 6). In this manner the expert vacillates between the different levels in reading a document, until a resolution is reached regarding the sense of the document (level 8), or until he has exhausted all possibilities regarding the text. An extract from this

discussion, illustrating how the Reading Levels are appropriated, is presented below.

| Timing (secs) | Transcribed text | Reading Level | |
|---|---|---|---|
| 131 | Then, when we start reading it, well this looks like a very good Latin word FECIT - F E C I T. | 4 | Identification of Word |
| 139 | And then immediately at the end what looks like Q U A N E. | 2 | Identification of Characters |
| 161 | Which is a bit of a problem... if that is the end of the line is has to be QUA ME instead of QUAN, You wouldn't really expect to see it there. | 4 | Identification of Word |
| 167 | But immediately when one gets as far as this you can see that this stoke coming down is E. | 1 | Identification of Feature |
| 174 | So it's a descender from the line above, so we haven't got the top of N. | 1 | Identification of Feature |
| 182 | Before that we have got M A possibly M I N A. | 2 | Identification of Characters |
| 186 | Which obviously works with the Latin word DOMINA. | 4 | Identification of Word |
| 190 | A woman has done something. | 6 | Discussion of Word Meaning |
| 202 | QUA is then better, the woman is then in the perfect. | 5 | Discussion of Grammar |

**Table: 2.3**. Excerpt from Expert B's discussion of 1491 illustrating the appropriation of Reading Levels to the text, and demonstrating the flow of the discussion between different levels.

All the experts' discussions regarding the texts followed a similar pattern, with various hypotheses concerning the identification of features, characters, and words, being checked against other information from the document, until some resolution of ambiguity was reached. Modelling this process computationally, by implementing each different level as an agent, and using MDL as a means to pass information between levels (see 1.5.3), should provide a way to construct a tool to aid the papyrologists in their reading, as it would enable this cyclic process to be replicated.

### 2.5.2.1 The Order of the Reasoning Process

It is possible to work out the latencies between different stages in the process (data from Think Aloud Protocols often being used to predict the sequence of cognitive

processes executed in a given task (Ericsson and Simon 1993)). This illustrates the likelihood of particular levels following on from each other, showing the order in which these processes occur. An attempt was made to see if there were any patterns occurring, resulting in some complex data[20]. This is best represented in a general format, as shown below where the likelihood of one topic following another in the discussions is plotted on a grid. A "high" probability means it occurs in over 5% of the total discussions; likely, between 3% and 5% of the discussion; often, between 1 and 3%; low, between 0.05% and 1%; unlikely, beneath 0.05% of the total discussions[21].

| First Topic of Discussion | Second Topic of Discussion: -1. Archaeo Info | 0. Physical | 1. Feature | 2. Character | 3. Char. Sets | 4. Word | 5. Grammar | 6. Word Meaning | 7. Phrase Meaning | 8. Meaning of Doc |
|---|---|---|---|---|---|---|---|---|---|---|
| -1. Archaeol Info | | | | | | | | | | |
| 0. Physical | | Often | High (M) | High | Low | Low | Low | Low | Low | Low |
| 1. Feature | | Often | High | High (M) | Often | Likely | Low | Low | Low | Low |
| 2. Character | | Often | High | High (M) | Often | High | Low | Low | Low | Low |
| 3. Char. Sets | | Low | Low | Often | Low | Often (M) | Low | Low | Low | Low |
| 4. Word | | Often | Likely | Often | Low | High (M) | Often | Often | Likely | Often |
| 5. Grammar | | Low | Often (M) | Low | Low | Low | Low | Low | Low | Low |
| 6. Word Meaning | | Low | Low | Low | Low | Low | Low (M) | Low | Low | Low |
| 7. Phrase Meaning | | Often (M) | Low | Often | | Low | Low | Low | Low | Low |
| 8. Meaning | | Often (M) | Often | Often | Low | Low | Low | | Low | Low |

**Table: 2.4**. Likelihood of one reading level following another in the discussions. (M) denotes most frequent occurrence on that level.

The interesting factor in this table is that only the Word level, Level 4, seems to have common connections with all of the different levels (save the archaeological and historical context level, which didn't occur during any of the TAPs). The other commonly spoken about subjects, features (Level 1) and characters (Level 2) (see Figure 2.4) seem to relate closely to the nearest related subjects. For example, discussions concerning features of characters are most often directly followed by discussions regarding other features and characters. The word level is the only one

[20] The data sets from which this data was derived are available on the accompanying CD-ROM in Chapter2/data analysis/comparison.xls

[21] It is worth noting that just because discussions directly follow on from each other they may not be directly related in some cases. However, for the most part, as the experts were talking exclusively about their reading of the texts, the order of the discussions should indicate something about the order of their thought processes.

where the next probable discussion is spread fairly equally across the range of topics available.

This is interesting when compared to an initial "subject-object" analysis of the critical apparatus from *Tabulae Vindolandenses II*[22]. Although these apparatus do not relate the order in which individual sections of the texts were read, it is possible to categorise the data that is presented both into the encoding scheme described, then into a more complex inner scheme which looks for any connections between the subjects by noting any secondary reference to other subjects. For example, the Expert may be noting a difficulty about the identification of a word, which depends on a problem with the characters within the word, giving a main subject of word, and secondary subject of character. Similar subject/object models, or "semantic triplet" (subject-action-object) models, have been recently developed and adopted by others in the field of Content Analysis, such as Franzoni's analysis of narrative structure (Franzosi 1997). By using such a method it is able to illustrate the transient relationships between different types of knowledge, seeing what type of knowledge the different instances depend on to reach their conclusions. The patterns which emerge from within the critical apparatus of the Vindolanda Texts also indicate that the Word level is the most linked of all the topics discussed.

[22] The data sets from which this data was derived are available on the accompanying CD-ROM in Chapter2/apparatus analysis/

| Main Topic of Discussion | Secondary Information: 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | -1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | High (m) | Low | Low | Low | Low | Low | Low | [illegible] | [illegible] | High |
| 7 | Likely | High (m) | Likely | Low | Low | Likely | Low | [illegible] | [illegible] | Likely |
| 6 | Low | Low | High (m) | High | Low | Low | [illegible] | [illegible] | Low | Likely |
| 5 | Low | Low | High | High (m) | Low | Low | [illegible] | Low | [illegible] | Low |
| 4 | Low | Likely | High (m) | High | High | Likely | Likely | Low | Low | Low |
| 3 | Low | [illegible] | Likely | Likely | Likely | High (m) | Low | Low | [illegible] | [illegible] |
| 2 | Low | Low | Low | [illegible] | Likely | Likely | High (m) | Likely | Likely | [illegible] |
| 1 | [illegible] | [illegible] | Low | [illegible] | [illegible] | Low | Likely | High (m) | Likely | [illegible] |
| 0 | [illegible] | [illegible] | [illegible] | Low | [illegible] | [illegible] | Low | Likely | High (m) | Low |
| -1 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | High (m) |

**Table 2.5.** Relationship of main and subsidiary topics within the Vindolanda Apparatus. (M) denotes most frequent occurrence on that level.

While the lower level processes, such as the identification of features and characters, mostly relate to each other, and the upper levels, such as discussion of word meaning and meaning of document, mostly relate to each other, it is only the word level which shows a relationship with all the different types of information discussed.

The identification of words, then, seems to be the topic which is most likely to relate to all the different types of information used within the process of reading an ancient text. Although not the most commonly discussed topic (see Figure 2.4) it appears to be the most flexible, deriving identification of word units from various sources. Perhaps this suggests that it is the most basic and central of the whole process, but this needs to be the focus of further research. The relationship between different reading levels in the process of reading an ancient text may serve as a useful starting point for further statistical analysis in the future. A general summary of the discussions most likely to precede and follow each level can be found in table 2.7, below.

### 2.5.3 Something to Talk About

By categorising the discussions in the manner discussed above, it is easy to illustrate which different topics the experts talk about most when describing their reading of a text. There seems to be a general trend in all the discussions.

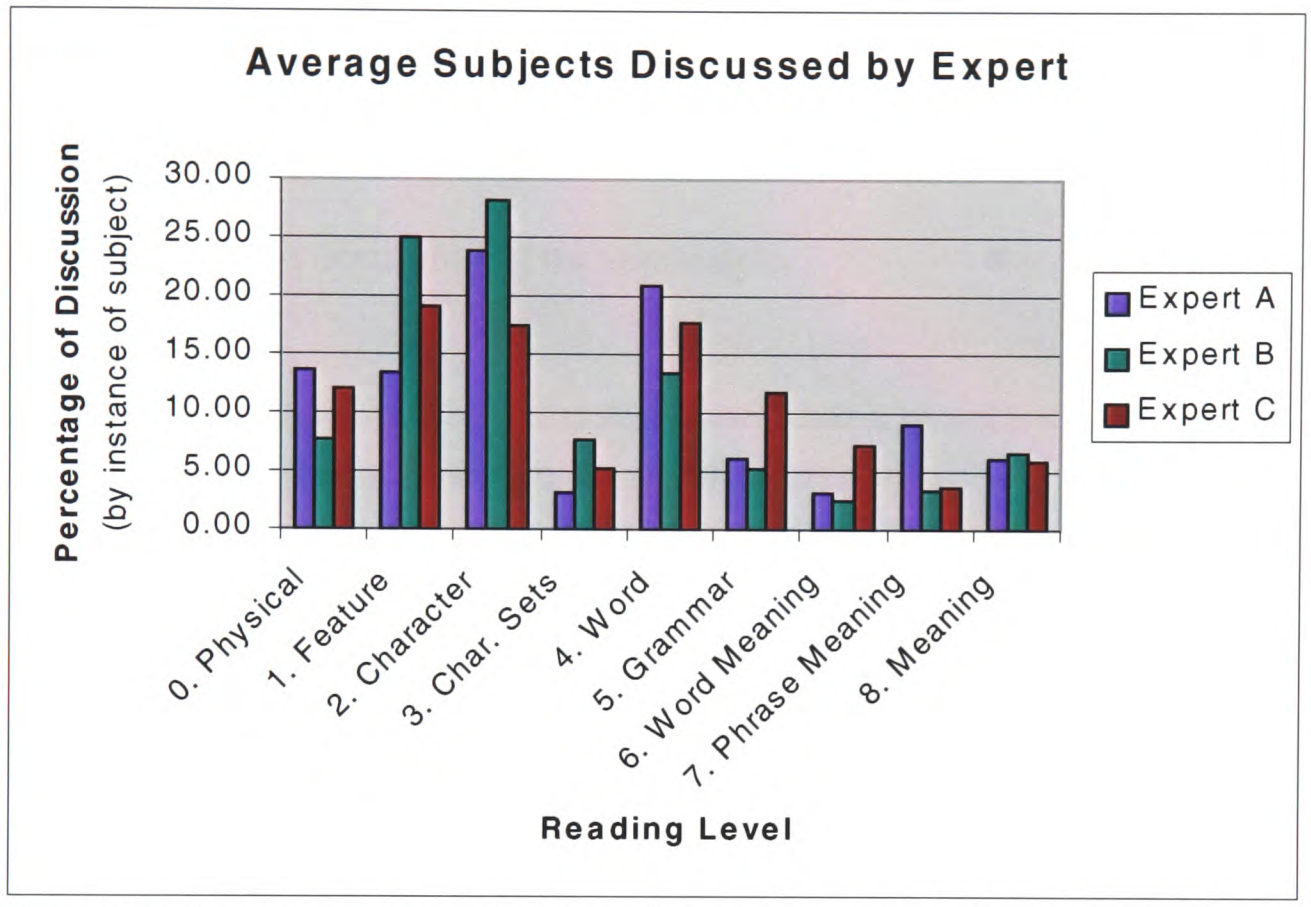



**Figure 2.8:** The average amount of subjects discussed by each expert throughout the discussions, by percentage of the discussion by instance[23].

It can be seen from this illustration that all three experts mostly talk about the features of the characters, the identification of characters, and the identification of individual words throughout the discussions. In comparison, the discussion of the meaning of the document as a whole actually takes up a small part of the process, as do discussions regarding word and phrase meaning, grammar, and the identification of character sets. Although there are some differences between the distribution of the data between experts, this figure shows the predominance of discussions

[23] Again, the data sets from which this data was derived are available on the accompanying CD-ROM in Chapter2/data analysis/comparison.xls

regarding features, characters, and words in the reading of ancient texts above that of grammar, phrase meaning, etc.

Conversely, the discussions regarding the identification of characters and words are most often the shortest amongst those occurring in the transcripts, being mostly declarative ("And then E and an N and an E" (Expert B, 1491) "And then FECIT" (Expert A, 1491)). This can be seen by comparing the average time spent discussing the different reading levels with the amount of separate instances of these discussions which occur during the transcripts.

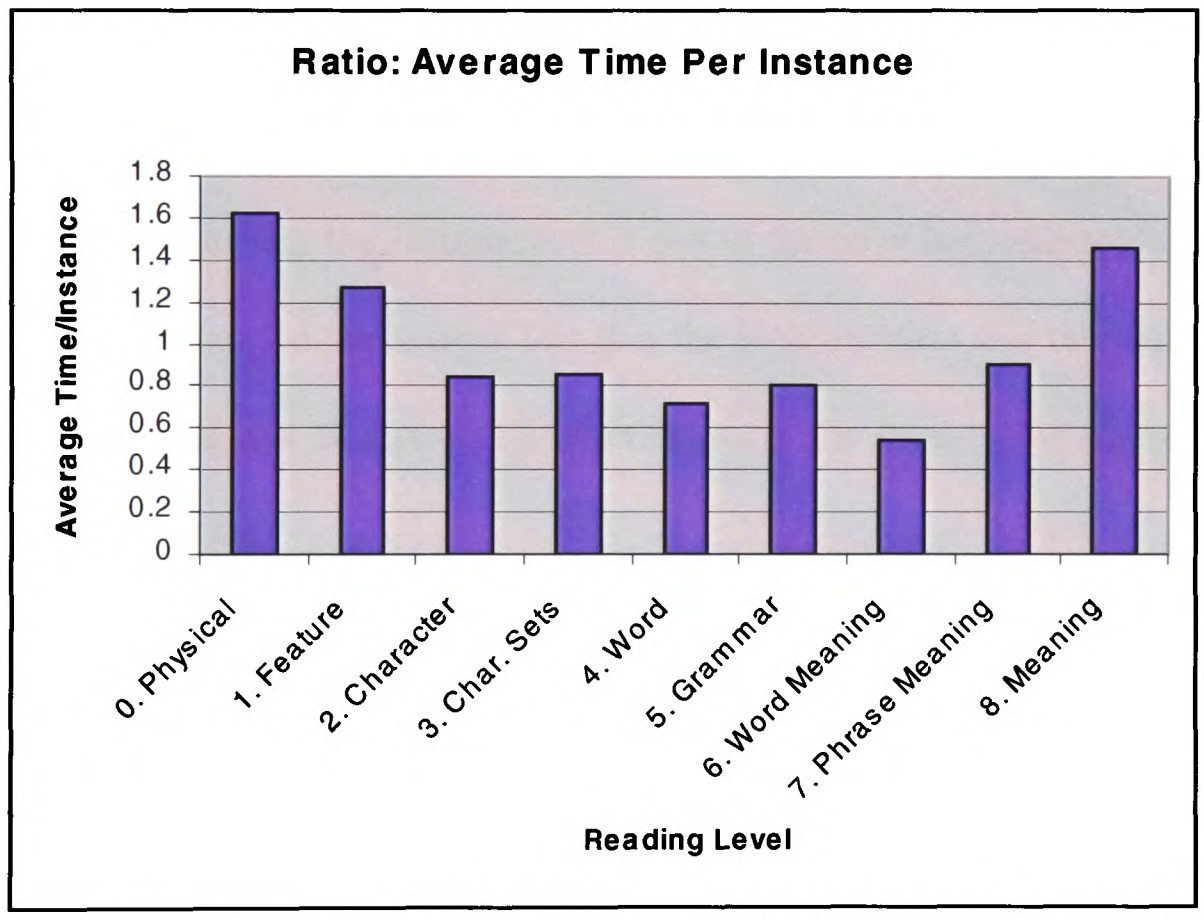



**Figure 2.9:** The average time spent discussing each subject compared to the amount of separate occurrences of those discussions within the transcripts. A ratio of 1 indicates the same length of time as the instances of discussion. Over 1 indicates that discussions are longer than average per instance, less than 1 indicates discussions are shorter than average.

The subjects which are discussed for the longest length of time per instance are the physical characteristics of the document, and the overall meaning of the text. The information regarding these levels is much more complex than the identification of

characters and words, and tends to be discursive. For example, when discussing the meaning of ink tablet 1543 Expert C explains:

> This all seems to make good sense. An instruction ... will all of you diligently take care that if any friends come they are well received, or something like that. He seems to have got a clear sense, starting here, which would fit very well with the beginning of a letter or instruction, which suggests that we probably have got the top of the tablet, and that this is a letter of instruction, possibly to the freed men, possibly even to the slaves as well ...

Discussions regarding the features which constitute characters, however, tend to be both frequent, and lengthy:

> And then, I'm afraid, more of these wretched things, this unit is part of so many different letters, its basically part of an A but could be part of an M, can even be part of a rather slanting N, I've got sort of three of these things, here. It's really rather difficult to tell them apart (Expert B, 797).

This indicates that the feature level is one of the most key, and most complex, in the process. It can be taken from this that the identification and rationalisation of what may and may not be a part of a letter is one of the most trying stages in the reading of ancient texts.

A summary of the frequencies of topic, and the length and type of discussions regarding each topic can be found in Table 2.7, below.

### 2.5.4 Reading Ink Texts Versus Reading Stylus Tablets

An analysis of the Think Aloud Protocols also indicates the differences between reading the ink texts (carbon ink on wood) and the stylus texts (incised texts) (see 1.1). Although from the same period and using similar letter forms (see 3.1.1), the stylus tablets are much harder to read due to their physical characteristics. Exactly how this affects the discussions regarding the texts can be seen below.

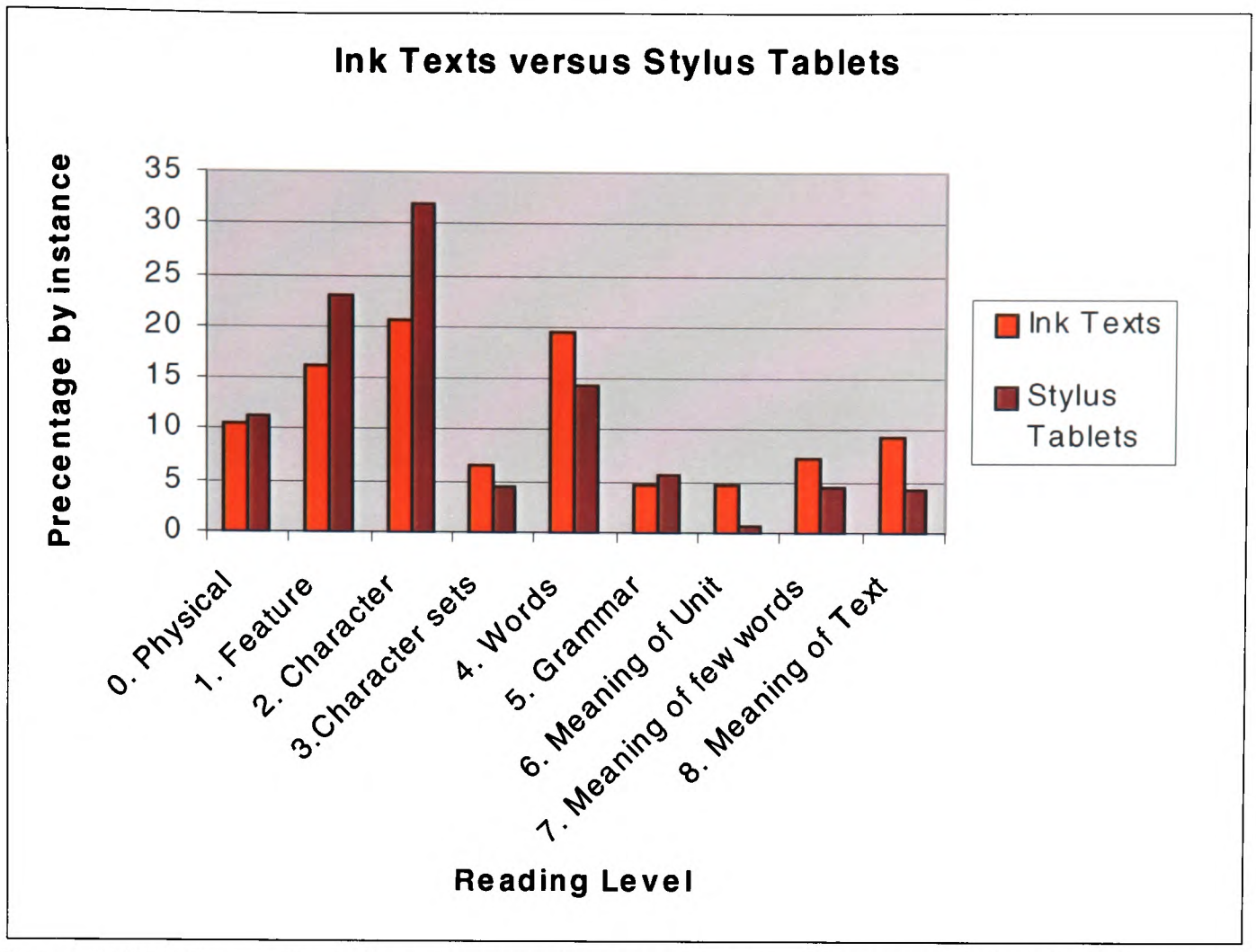



**Figure 2.10:** Comparison of percentages of subject matter in discussions of the ink and stylus texts. Two ink texts and two stylus texts discussed by Expert A and Expert B were used to generate the data[24].

It can be clearly seen from this figure that there are more instances of the feature and character level in discussions of the stylus tablets rather than the ink texts. There are more instances of the word, and meaning levels in the discussions regarding the ink texts, indicating how hard it is to generate meaning from the writing on the stylus tablets. This is backed up when comparing the length of time different subjects are discussed in the reading of the ink and stylus texts.

[24] The spreadsheet from which this data was derived is available on the accompanying CD-ROM in Chapter2/ink v stylus/

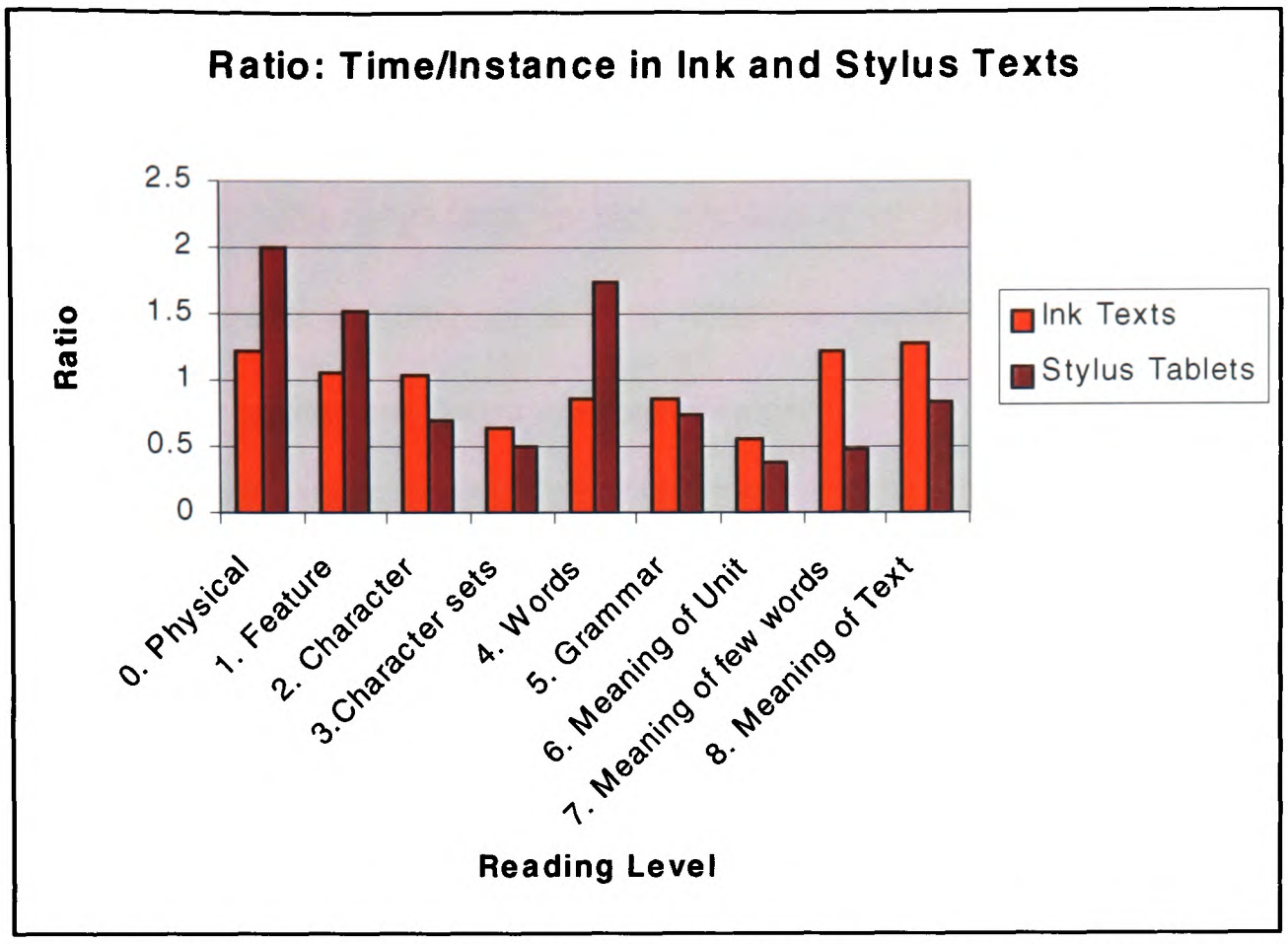



**Figure 2.11:** The average time spent discussing each subject compared to the amount of separate occurrences of those discussions within the transcripts of the ink and stylus texts. A ratio of 1 indicates the same length of time as the instances of discussion. Over 1 indicates that discussions are longer than average per instance, less than 1 indicates discussions are shorter than average.

Discussions regarding features of characters in the stylus texts are lengthy as well as frequent: as above in 2.5.2, and are much more lengthy than those of the ink texts. The discussion of features is therefore crucial to the reading of stylus texts[25]. Interestingly, the discussions regarding the physical attributes of stylus texts are lengthy and complex, indicating that much more attention is paid to the format of the text when reading stylus texts. Although discussions regarding words are not as frequent in the stylus texts as in the ink texts (Figure 2.6), discussions identifying words in the stylus texts are much lengthier, indicating how hard it is to identify words in these type of documents. A table summarising the different characteristics of these writing levels is presented below (Table 2.7).

[25] It has been noted that "the more difficulty the reader has with reading [a text], the more he relies on the visual information" (Smith 1971, p.221).

The difficulty in identifying features, characters, and words in the stylus texts can also be illustrated by comparing the word lists derived from the ink and stylus text discussions[26]. The keywords in the discussion of the stylus texts (which are unusually frequent in comparison with what one would expect on the basis of the discussions regarding ink texts) are shown below:

| N | WORD | FREQ. | STYLUS TXT % | FREQ. | INK TXT % | KEYNESS |
|---|---|---|---|---|---|---|
| 1 | HERE | 100 | 1.14 | 30 | 0.21 | 83.8 |
| 2 | THERE | 224 | 2.56 | 182 | 1.26 | 52.1 |
| 3 | THIS | 195 | 2.23 | 170 | 1.17 | 38 |
| 4 | SORT | 83 | 0.95 | 53 | 0.37 | 30.6 |
| 5 | YOU | 173 | 1.98 | 163 | 1.12 | 26.9 |

**Table 2.6**. Key words in the discussions regarding stylus tablets. Keyness is calculated using the classic chi-square test of significance with Yates correction; any word with a keyness of 25 or over is taken to be significant (Sinclair 1991).

Although these words are distinctly unglamorous, they show that the experts need to continually point out elements of the text (here, there, this), try to evaluate their judgements more (sort, whether this is used to "sort" between hypothesis, or explain how they "sort of" function), and explain to the knowledge engineer the features that they see, and how they undertake the task (you).

A further comparison of the word lists derived from the two sets of discussions[27] illustrates the difference in approach towards the two types of texts further: the experts use the words AFRAID, ASSUME, CONFUSING, CONVINCE, DECIDING, SURPRISED, and TRIED in relation to the stylus texts but not the ink texts. Words occurring in the ink text discussions but not the stylus text discussions tend to be less about the decision making process and the frame of mind with which the experts approach the texts, and more about features of the texts themselves, such as HORIZONTAL, BOLD, FORMAT, and DISCOLORATION.

---

[26] Available on the accompanying CD-ROM in Chapter2/ink v stylus/

Reading the stylus texts is more laborious than reading the ink texts, with much more attention to the features of characters necessary, and much more ambiguity surrounding the hypothesis generated regarding word identification. Less attention is given to discussing the meaning of the document as a whole, mainly because they do not seem to get that far in these two examples: the experts have problems enough in making certain assumptions about the characters and words within the texts. Further investigation into this area could look to see if there was any difference in the order that experts discussed different types of knowledge between the stylus and ink texts, the data sets here being rather small to allow valid conclusions to be drawn regarding this.

## 2.5.5 Recounting the Process

There are two further questions which shall be addressed regarding the process of "private" papyrology. Firstly, how does the initial reading of a text relate to an explanation of that reading at a later date (and so how valid is the evidence presented by papyrologists when they explain how they reached a reading). Secondly, how does the subject and textual content of the expert's discussions regarding the texts relate when compared to those of the published volume of apparatus containing the ink texts, *Tabulae Vindolandenses II.*

[27] Available on the accompanying CD-ROM in Chapter2/ink v stylus/

### 2.5.5.1 Retelling The Story

It has already been shown that when an expert was asked to recount how he came to a reading of a text after an initial reading, he discussed the prime elements in a different order (Figure 2.2). The primary tenet of Knowledge Elicitation is that experts find it difficult to relay the processes they went through whilst undertaking a task. However, analysis of concurrent and retrospective Think Aloud Protocols by Ericsson and Simon (1993) show that there is very little difference between the two. Expert B's two discussions regarding 1491 are very similar[28]. The first discussion lasts 1347 seconds, the second 1246 seconds, and the frequency of subject matter discussed correlates closely, as shown below.

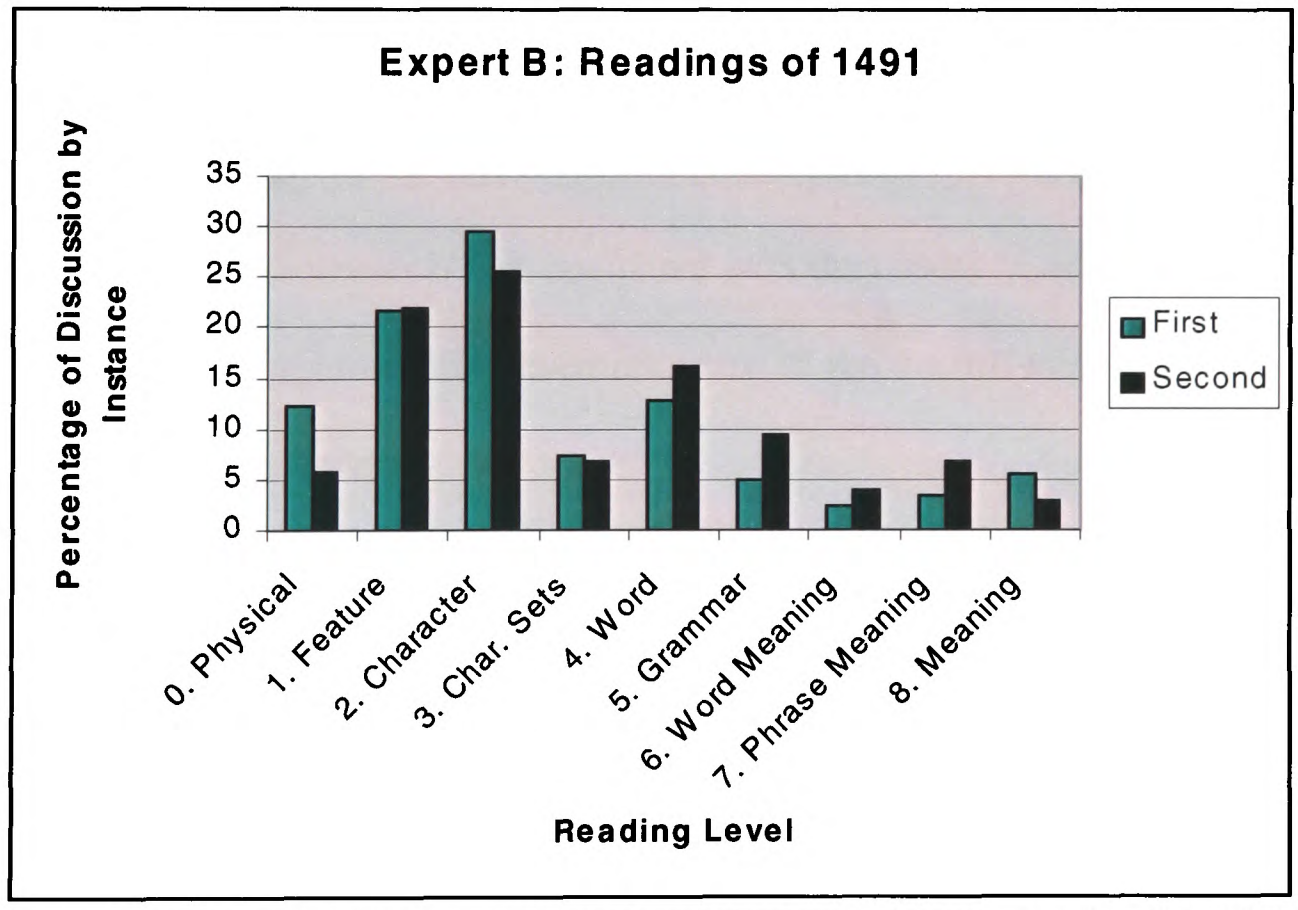



**Figure 2.12:** Comparison of Expert B's reading and retrospective account of reading 1491.

When an error rate of 4.1% is considered, it becomes obvious that there is very little difference in the distribution of subject matter between the two readings. The same

is also true for the length of time talking about the subject matter in each case. There is also little difference between the words used to discuss the texts, and the frequencies or collocates of these words[29]. It would seem that the retrospective discussions of how the experts approach a text are closely linked to the actual processes they got through (or the ones they think that they go through and can verbalise) when trying to read a text. The problem for the knowledge engineer, then, is not when the experts discuss how they carried out a process, but if there is any difference between their account and what is actually taking place, a gulf there is no means to cross at present.

### 2.5.5.2 Published Commentaries Versus Discussions

How do the published commentaries regarding the ink texts differ from the discussions regarding the reading of the ink texts, and what can they show about the process of reading an ink text? A content analysis of seven critical apparatus from *Tabulae Vindolandenses II* was compared with the results from a similar analysis of the discussions regarding the seven ink texts, to see the difference in subject matter covered between the two.

---

[28] See the accompanying CD-ROM in Chapter2/Expert B 1491/
[29] Again, see the accompanying CD-ROM in Chapter2/Expert B 1491/

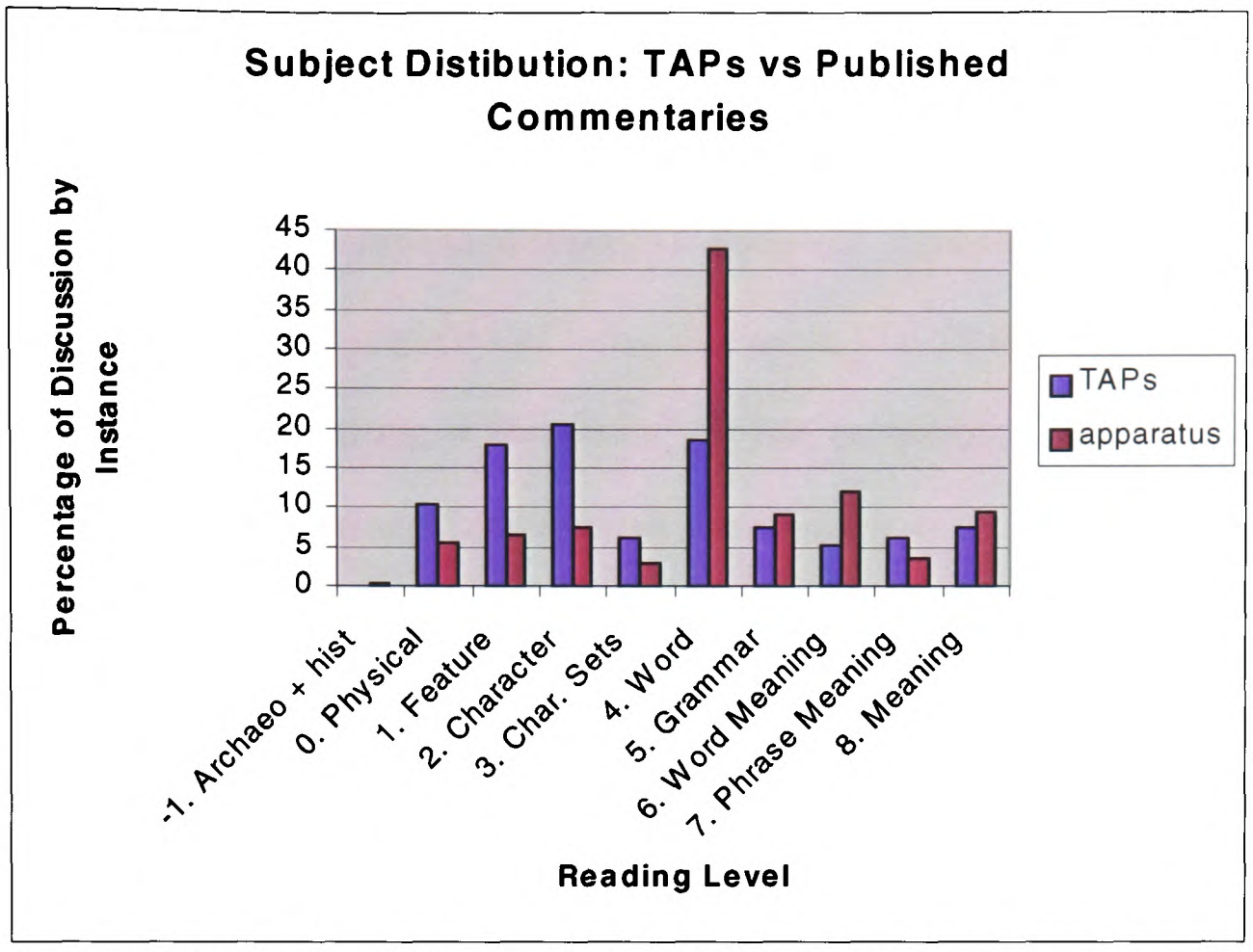



**Figure 2.13:** Distribution of subject matter in the Think Aloud Protocols compared with those of the critical apparatus of the ink texts.

It is obvious that the published commentaries focus mostly on the identification of words, whilst the Think Aloud Protocols concentrate equally on the identification of features, characters, and words, and place attention on the physical nature of the document. This is partly due to the publishing format of the apparatus, but also shows that the published commentaries are slightly removed from describing how texts are read, not describing the basic processes of reading a text. However, the Leiden system does preserve some information regarding the certainties of the readings of the texts, and an analysis of this provides some interesting information about the role of certainty in the reading of ancient documents.

### 2.5.5.3 Textual Analysis of the Published Vindolanda Ink Texts

The Leiden system of transcription, a series of symbols that denote various characteristics of the original text[30], was devised in 1931 (Turner 1968) and remains the standard in notating a document[31]. The published Vindolanda ink texts annotated in this manner from *Tabulae Vindolandenses II* (Bowman and Thomas 1994) and the ink texts prepared for publication (so far) for *Tabulae Vindolandenses III (Bowman and Thomas Forthcoming 2003)* were collated. This gave a corpus of 286 ink texts in total[32], comprising 2301 lines and 27364 characters of Vulgar Latin, grouped into 6532 words, or groupings of characters[33].

Of these 27364 characters, 25915 characters were identified as actually being present within the documents, the remaining 1449 being probable text which is missing from the document due to breakage, damage, etc. These were grouped in 761 separate instances, often ending or beginning words which were left incomplete due to damage. This means that 5.3% of the characters, and 11.7% of the words in the readings of the Vindolanda texts were actually not physically present, and predicted by the experts because of their prior knowledge of the language and

---

[30] The Leiden system is as follows:
[ ] probable text which is missing from the document,
( ) an abbreviation which has been expanded,
{ } an editorial deletion,
[[ ]] an original scribal correction or deletion,
‘ ’an insertion above the line,
. (below a letter) a letter which cannot be read with certainty.

[31] Although the application of the Leiden system can be shown to vary between different readings of a text by both the same and separate individuals (Youtie 1966).

[32] Although over a thousand tablets in total have been found at Vindolanda, many of them contain little or no writing, or are in such a fragmentary state that nothing of value can be read from them (Thomas 1992), and so only the published texts were included in this corpus.

[33] Available on the accompanying CD-ROM in Chapter2/Vindolanda ink texts corpus/

similar texts, being restoration rather than reading. There were 772 (2.8%) instances of characters taken to be abbreviations within the documents.

2566 characters from the total 25915 characters were underscored, indicating an uncertainty about the reading of the letter. Although there is no way to measure the extent of that uncertainty (i.e. whether they are a little uncertain of the reading, or very unsure that it is the correct reading) this shows that 9.9% of the characters read in the Vindolanda ink texts were marked as being uncertain: a fairly high proportion.

Thus ambiguity is a main feature of the readings reached of the documents, even in their published version. This is not a critique of the work of the papyrologists; published versions are open to correction, and merely detail the extent to which the author has resolved the reading of the text at that moment in time[34]. However, it does show that there still remains a great deal of uncertainty about the reading of the ink texts, and that this is very seldom exhausted. The process revolves around the resolution of ambiguity through prediction, prior knowledge, reasoning about the characteristics of the document, and the addressing of uncertainties. This ambiguity would be difficult to implement in a Artificial Intelligence system, as it would be hard to provide enough real world knowledge to reflect the amount of contextual information necessary to undertake such a task.

---

[34] Bowman and Tomlin (Forthcoming 2003) provide examples of readings which have changed dramatically between different versions of published texts. Tomlin (1994) demonstrates how the correct reading of a lead tablet was achieved by rotating the tablet through 180 degrees and re-reading the text.

## 2.6 General Observations

There was substantial procedural information gathered from the literature review, the elementary stages of knowledge elicitation, and the data from the Think Aloud Protocols. These general observations regarding the reading of the ink and stylus tablets illustrate further the complexity of the reading process. Some observations peculiar to the reading of stylus tablets indicate the difficulties faced by the experts when trying to read such damaged and abraded texts.

The papyrologists continually refer to other material, such as other texts, word lists, grammars, and archaeological evidence. Reference to such information is not always made explicit. Only when all parallels to other texts are exhausted are they comfortable with making any new assertions regarding the language or subject of a document: there is an element of appropriateness to the resolution of ambiguity regarding the texts.

The understanding of convention is paramount in being able to read a text. There are both physical conventions, regarding format, and textual conventions, regarding language. The format of the document can indicate whether it is a letter, an account, or official documentation, and these standard formats can allow the experts to immediately focus expectations regarding the possible subject matter of the texts. Linguistically, there exist conventions of letter writing, phrases commonly used to express certain situations, and conventional ways of conveying certain information, for example literary phrases, and high class spellings. The papyrologist is always on the look out for any conventions at work (and alternatively, unconventional usage) in order to gain an understanding of the text. As the body of texts read from

Vindolanda increases, the conventions regarding format and language are becoming better understood, and so aid in the reading of subsequent documents.

The starting point in reading a document is always the clearest text on the document, not the first characters at the beginning of the document. Uninscribed or clear areas (whether or not deliberate word separation has been used) can give vital first clues to where words begin and end, and the space between these can indicate the number of characters needed to complete the lexical unit. The papyrologists often note the relationship of the writing with the grain of the wood and how it flows along the line, which can give some indication of the letter forms. There is also consideration of the physical act of writing, asking what could be the possible intention of the writer, and if there was a problem in making individual characters due to the format, for example squeezing in characters so that a word fits into the end of a line. Differentiation in the thickness of strokes in the ink tablets can help identify characters.

Once a letter form has been identified in the document, this is often used as a template to compare more ambiguous letters, to give some measure of the likelihood that a group of features can be identified as a letter. The expert's familiarity with the different handwriting already identified in the corpus can give some indication of the subject matter of the text; if they can identify the writer they can have a good idea of who the document was addressed to, and the sort of information it will probably contain.

Probable sequences of letters give a clue to what words might be there, and it is often easier to work backwards from the end of word than the beginning to identify it. Possible words are continually checked with external sources. Deciding whether the words wrap around to the next line of text can prove problematic. The identification of words does not happen in isolation: context is everything[35]. It is perfectly acceptable to leave readings as conjecture when they still remain unclear.

As far as the stylus texts are concerned, the first task is deciding which way up the document goes, which is often tricky[36]. The format is paid more attention to, as it can often give a clearer indication as to the purpose of the document than the format of the ink texts, which is more uniform. It can be difficult to decide whether the writing on the stylus text continues all the way across the document, or if it is written in columns. Possible lines of text are identified, to give an indication of where characters may actually be. Uninscribed areas are particularly important in the reading of the stylus texts, as they provide some definite information regarding the flow of text.

The inscribed nature of the writing with the Stylus point means a signature mark is left when it "bumps" over woodgrain, and this can differentiate actual letter strokes from background noise. Care needs to be taken to accommodate the effect of changes of pressure in the making of strokes in order to identify them properly, as tails of letters can often lift off and become faint. Strong definite strokes are identified first, and signature strokes, such as descenders and ascenders, can give

[35] This may present problems when moving to a more fully automated system, in that trying to model the contextual information is almost impossible due to the vast nature of associated information that papyrologists use when trying to read a text.

clues as the use of key letters, for example the letter S (see 3.7). Care must be taken to avoid reading persistent strokes that remain from text(s) written on the stylus tablet previously. Some consideration must be given to differences in letter forms (from those used in the ink texts) caused by writing on a different, and more difficult, medium (see 3.1.2).

## 2.7 Preliminary Conclusions

Although each expert has his own individual style in reading an ancient text, it has been shown that there are some unifying procedures in the process of reading an ancient document. The first success of the knowledge elicitation exercise was making explicit the different topics discussed regarding the ink and stylus tablets, providing a representation of the different types of information the experts use whilst reasoning about such texts. The study also enabled the identification of characteristics regarding each level, and the frequency and order of occurrence of each topic. Moreover, this gave specific information regarding how reading the stylus tablets differs from reading the ink texts, indicating the features of reading that become more important when encountering more damaged and abraded texts. This information has been summarised, below.

| Reading Level | Thematic Subject | Characteristics |
|---|---|---|
| 8 | Meaning or sense of document as a whole | This is discussed surprisingly rarely, but when it is, the reasoning tends to be lengthy, complex, and discursive. It is usually preceded by discussions regarding the identification of words, and followed by discussions regarding the physical characteristics of the document, features of characters, and the identification of characters. There were more often, and lengthier, discussions |

[36] "I decided it was probably the other way up I think, did I or did I not? No I think it is this way up" (Expert B, 1593).

| | | |
|---|---|---|
| | | regarding this level in relation to the ink texts than the stylus tablets. |
| 7 | Meaning or sense of a group or phrase or words | This is rarely discussed, and tends to be regarding linguistic convention. It is usually preceded by identification of words, and followed by discussions regarding physical characteristics and the identification of character sets. Discussions tend to be short, although they are marginally longer with regard to the ink texts. This level occurs marginally more often with regard to the ink texts. |
| 6 | Meaning or sense of a word | Surprisingly, this is seldom discussed. It is the shortest of all discussions, tending to be very declarative. It is usually preceded by the identification of a word, and followed by discussions regarding grammar. There are marginally more occurrences of this level with regard to the ink texts. |
| 5 | Discussion of grammar | This is seldom discussed (although Expert C refers to grammar often.) It is usually preceded by the identification of a word, and followed by discussions regarding the feature level. Discussions are short, and are distributed equally over the transcripts regarding the ink and stylus texts. |
| 4 | Identification of possible word or morphemic unit | One of the most commonly discussed subjects by the papyrologists, although discussions tend to be short and declarative. It tends to be preceded by the feature, character, character set, and word levels, and is often followed by every other level in the representation: the only one to have this characteristic. Discussions regarding words are more frequent in relation to the ink texts, but are considerably longer in relation to the stylus texts. |
| 3 | Identification of sequence of characters | This is fairly seldom discussed, and discussions tend to be short. It is usually preceded by discussions regarding the feature and character levels, and followed by discussions regarding the identification of characters and words. It is equally distributed in both sets of transcripts. |
| 2 | Identification of possible character | This is the most discussed topic, on average. Discussions tend to be declarative and short. It can be preceded by discussions regarding physical characteristics, features, characters, and less often discussions regarding character sets, phrase meaning, and the meaning of documents. It is most often followed by discussions regarding features, characters, character sets, words, and sometimes the physical characteristics of the document. It is much more frequent in the transcripts regarding the stylus tablets, |

| | | |
|---|---|---|
| | | but discussions regarding this in relation to the stylus tablets tend to be shorter. |
| 1 | Discussion of features of character | This is one of the most commonly talked about topics in the transcripts, but discussions tend to be complex, lengthy, and discursive. It is usually preceded by discussions regarding the physical characteristics of the document, the feature of characters, the identification of characters, and sometimes identification of words and character sets. It is usually followed by discussions regarding features, characters, character sets, and identification of words. It is more frequent in relation to the stylus texts, and discussions of this level also tend to be lengthier in relation to the stylus texts. |
| 0 | Discussion of physical attributes of the document | This is discussed fairly regularly, and discussions tend to be lengthy, complex, and discursive. It is often followed by discussions regarding the physical attributes, features, and identification of characters, and often preceded by discussions regarding the physical characteristics, features, and characters. It is important in both sets of transcripts, but discussions of this level in relation to the stylus tablets tend to be lengthier. |
| -1 | Archaeological or historical context | Direct reference to external resources was surprisingly rare. However, the experts continually relate the hypotheses they generate to their own extensive knowledge regarding similar texts. This level was mostly present in the published commentaries, providing examples to aid in the discussions of other levels. |

**Table 2.7:** Features of the reading levels identified in the discussions regarding the ink and stylus texts.

The investigation also revealed the cyclic reading process; the process of reading a text is not linear, building up one character at a time, but depends on the propagation of hypotheses, and the testing of these regarding all available information concerning a text. Reading a document is a process of resolution of ambiguity, and depends on the interaction of all the different facets of knowledge available to the expert. Specific details of how the experts proceed in reading ancient documents were also made explicit.

These conclusions can be drawn together to propose an elementary model of how experts read ancient texts. (However, firstly it will be necessary to review other models of reading that have been developed in the field of psychology). The conclusions can also be used to indicate which part of the process the experts need assistance with to aid in the reading of the remainder of the stylus texts. The lower levels of the process (discussions regarding physical attributes, identification of features, identification of characters, and words) are much more a focus in discussions regarding the stylus tablets, and any subsequent computational tools should concentrate on aiding the papyrologists in these areas. As such, this is the focus of the system discussed in 5.3.

## 2.8 Models of Reading and Papyrology

Although the act of reading an ancient document has never been the focus of a psychological investigation, and there exists very little psychological research that directly relates to the reading of an ancient text (see 2.1.2), there exists a large body of research covering general aspects of reading. Various attempts have been made to construct some kind of model of the reading process, but it has so far proved impossible to condense all the different elements involved into one precise, albeit high level, theory. However, successful models exist which relate to specific tasks in reading. Using the results from the investigation, above, it is possible to propose a model of how experts approach and read ancient texts.

## 2.8.1 Psychology and Models of Reading

Reading, as a focus of cognitive study, covers a large, problematic, critically diverse area. Constructing models of the reading process has proved problematic, but so has defining what "to read" actually means; "Reading is extracting information from text." (Gibson and Levin 1976, 5);

> Readers do not so much extract meaning from print, but rather engage in an active construction of meaning based on the signs provided by the print (Crain and Steedman 1985).

Although the physical process of reading is well documented, for example the physiology of the eye and the muscle movements made whilst reading a text (Gibson and Levin 1976; Oakhill and Garnham 1988; Asher 1994; Gregory 1994; Manguel 1997), the cognitive process that results in assigning a text meaning remains obscure.

Many attempts have been made to illustrate the reading process using systems modelling techniques borrowed from the field of computer science. Two types of these models of reading exist; sequential and componential. Componential models have proven problematic as it is impossible to form a definitive list of universally applicable components of the reading process (Asher 1994). Sequential models have been slightly more successful in their aim to chart the time course of reading from the start of the process, and can be classified into three sections. "Bottom up" models such as Gough's "One Second of Reading" (Gough 1977) utilise perceptual information and skills, treating each occurrence of reading as a new cognitive puzzle to be solved individually. "Top Down" models, such as Goodman's "psycholinguistic guessing game" (Goodman 1967) define reading as a process of

prediction, confirmation and correction and rely on previously stored linguistic information. Interactive, or connectionist, models, such as McClelland and Rumelhart's "Interaction Activation and Competition Model of Word Recognition" (McClelland and Rumelhart 1986) cycle between the Top Down and Bottom Up processes, and are the most favoured of all models by current theorists (Ellis and Humphreys 1999).

The extent to which these models actually explain the reading process can be illustrated by sections of Gough's model: "Merlin" is the name given to the mechanism that magically applies syntactic and semantic rules to viewed text, and the central processing plant is named TPWSGWTAU, "The Place Where Sentences Go When They Are Understood". In recent years attempts to develop a definitive model of the reading process has slowed as it has become obvious that reading is a complex system which encompasses various diverse processes:

> all that can be said with a fair degree of certainty is that the skills readers use for the extraction of meaning from print are extremely complex, automated to a fairly high degree, and highly flexible, comprising a number of different styles (Asher 1994, p.3457).

The flexibility of the reading process has led researchers to identify a number of reading styles, dependant on both reading purpose and text type, for example skim reading, detailed reading and puzzle solving (Tunmer and Hoover 1992; Asher 1994; Gibson and Levin 1976). As "there is no single reading process, there can be no single model for reading" (Gibson and Levin 1976, p. 438). However, they can give insight into individual processes, such as McClelland and Rumelhart's "Interaction Activation and Competition Model of Word Recognition" (McClelland and Rumelhart 1986) which aims to explain the mechanism of the word superiority effect.

## 2.8.2 The Interaction Model of Reading

The finding that a word can be identified more accurately than a single letter has been known since the earliest considerations of the reading process (Cattel 1886) The Word Superiority Effect, the fact that "Letter recognition will be better for letters in words than for letters in non-words or even for single letters" (McClelland and Rumelhart 1986) is accepted by most in the field.

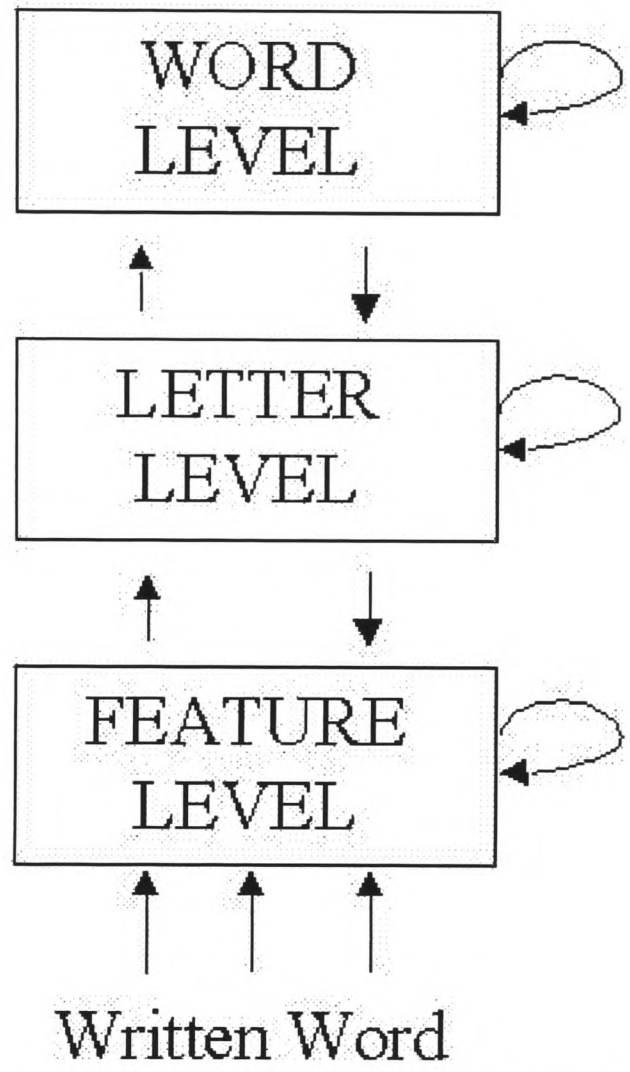



**Figure 2.14**: McClelland and Rumelhart's "Interaction Activation and Competition Model of Word Recognition" (McClelland and Rumelhart 1986).

McClelland and Rumelhart's model demonstrates the resolution of ambiguity in the reading process by the use of a recursive mechanism, which favours the identification of letters in known words. Features are identified at the basic level and then checked to see if they are likely characters. If not then the features are re-identified. Once a letter appears to be likely, it is related to the surrounding letters

to see if they are likely to combine and become a word. Words which are already known are therefore most likely to be the result of this process, and words take precedent over identifying random characters independently: a character is most likely to be identified if it is part of a known word. This is a very simplistic, recursive, model, shown to be very effective in practice (Ellis and Humphreys 1999). An implementation of a model similar to this was used by Robertson (2001) to demonstrate the effectiveness of the architecture of a system based on Minimum Description Length (see Chapter 5). Robertson's system, based on this recursive model, successfully "read" a hand written text.

### 2.8.3 Proposed Model of the Papyrology Process

The papyrologists seem to operate in a similar way as the model shown above, as it has been demonstrated that they use a recursive reading mechanism which oscillates between different levels, or modules, of reading. To this extent, it is possible to propose an overall model of how the experts work. The model shown below is stacked hierarchically, with the "Overall Meaning of Text" agent nominally being taken to be the highest level in the sequence of events before a transcription of the text can be prepared for publication. As suggested above, it is possible that the Word level is actually the most important of the process, but the levels are presented in this order to suggest a basic order of interaction whilst trying to develop a reading of the text.

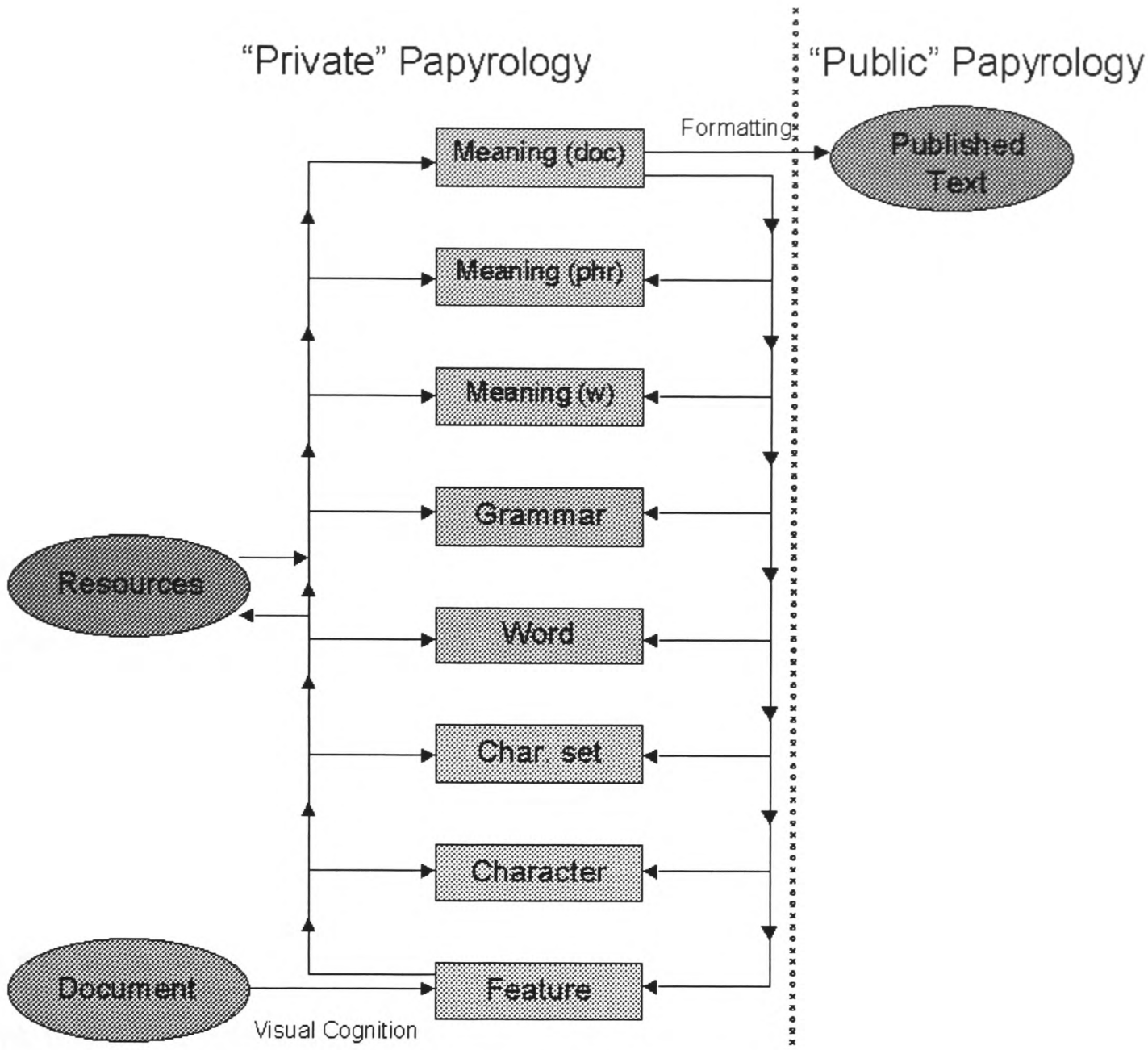



**Figure 2.15:** Proposed model of the procedure used to read an ancient text.

An expert reads an ancient document by identifying visual features, and then incrementally building up knowledge about the document's characters, words, grammar, phrases, and meaning, continually proposing hypotheses, and checking those against other information, until s/he finds that this process is exhausted. At this point a representation of the text is prepared in the standard publication format. At each level, external resources may be consulted, or be unconsciously compared to the characteristics of the document.

Although this is a simple representation of the process carried out, it shows the overall scope of the process of reading an ancient text. This model was used as the basis of the computer system developed to aid in the reading of the stylus texts, as discussed in Chapter 4.

## 2.9 Conclusion

This investigation has been the first attempt to make explicit the process of reading an ancient text. Although it has been shown that individual experts have different approaches to reading a text, it has been demonstrated that there are overall similarities in their working methods. The process has been stratified into defined modules, investigating the characteristics of each, and how they relate to each other. The problems in reading ink and stylus texts have been illuminated, with specific conclusions drawn regarding how reading the stylus texts differs from reading the ink texts. The role of ambiguity in the production of readings has been highlighted. The general process has been condensed into an overall model in order to provide a basis for the development of computational tools to aid the experts read the stylus tablets.

There is no doubt that the organisation, documentation, and analysis of the Think Aloud Protocols and published commentaries was a time consuming and tedious task. Some of the conclusions reached regarding the papyrology process may seem obvious (the more difficult and deteriorated a document, the more the experts pay attention to features of the document: who would have thought?) However, it has provided data which has successfully given insight into a hitherto undocumented area, and has shown how utilising such techniques can provide rich quantitative and qualitative evidence on how experts carry out their tasks. There remains a great

amount of research which could be done to investigate the cognitive processes the experts go through when reading an ancient text. For example, more graded tasks could be given to the experts. The role of eye movements in reading such texts could be investigated; little attention has been paid to the role of visual cognition in reading ancient texts. The techniques of experts who work with other language systems could be studied. Possible future work regarding Knowledge Elicitation is detailed in 5.1.

The conclusions reached in this chapter also pertain to the process of reading texts of a particular period, location, and medium, and it is more than possible that an investigation into the reading of different types of ancient document and inscriptions may reveal different findings. However, this research has provided a concrete starting point for the development of computational tools to aid papyrologists in reading the Vindolanda texts.

# CHAPTER 3

# The Palaeography of Vindolanda

## Knowledge Elicitation and the Papyrologist (2)

"A language is a dialect with an army" (Weinreich 1945, p.13).

To enable the construction of a system which could be used to read the stylus tablets, it was imperative to gain an understanding of the letter forms that might be contained within the texts. The only significant body of contemporaneous documents to the stylus tablets are the Vindolanda ink texts, and it can be demonstrated that the stylus tablets should contain similar character forms to those found on the ink tablets, even though the two types of texts are written on substantially different media. The construction of a corpus of annotated images, based on the letter forms found in the ink tablets (and the few stylus tablets that have been read), provided data with which to train a system to read unknown characters in the stylus tablets (the focus of Chapter 4 in this thesis).

To create such a data set, it was firstly necessary to undertake a review of what is known about the letter forms contained in the Vindolanda texts. Both sets of texts contain handwriting in the form that is known as Old Roman Cursive (ORC). Although many aspects of Latin palaeography have been studied in depth, ORC is the focus of academic debate, due to the paucity of documents available from this period. A series of knowledge elicitation exercises were undertaken with the three

experts (see Chapter 2) to gain an understanding of the types of information they use to describe and identify ORC character forms. From this a schema was constructed, detailing the relationships between the different types of information identified. An encoding scheme was developed from this schema, enabling the annotation of images of the Vindolanda texts. Nine documents were annotated, using this encoding scheme and an annotation program which was adapted from a tool used to capture data regarding aerial images. This resulted in a corpus of annotated images which comprises of data regarding 1700 individual letters from the Vindolanda corpus. As such, it constitutes a unique information source that could be used in the future by palaeographers and papyrologists, and is the source of information used to train the system described in Chapter 4. The corpus can also be used to demonstrate that the letter forms found within the ink and stylus tablets are indeed similar.

## 3.1 Palaeography

Palaeography, the study of ancient handwriting,

> in the strictest sense deals only with the old styles of writing, whereas palaeography in the wider sense embraces everything related to written texts of the past: the technique of writing, its material support, the mode of its transcription, and also the ways texts were diffused and circulated (Boyle 2001, p.xi).

Palaeography incorporates elements of codicology, epigraphy, philology, diplomacy, and papyrology. Its traditional task was to date manuscripts, and also to investigate the authenticity of documents (Bately, Brown et al. 1993). Its methodology is fairly transparent, as the palaeographer deals with the forms of

letters on an individual basis, providing in-depth documentation[1] to chart the development and use of styles of handwriting. The handwriting of Latin manuscripts has been the focus of systematic study since the late 17th Century (see Bischoff (1990) for a comprehensive introduction).

### 3.1.1 The Palaeography of the Vindolanda Ink Tablets

The letter forms in the Vindolanda ink tablets are "Old Roman Cursive" (ORC)[2] (Bowman and Thomas 1983), the everyday Roman script during the first three centuries AD (as opposed to the formal bookhand used for literary works). Although ORC was the commonly used hand, and has been studied since the early 1800s, extant sources are rare[3], and are mostly from the South and Eastern reaches of the empire. Scholars' understanding of the forms and conventions utilised is still incomplete, despite much academic interest in the area[4].

[1] For example, a comprehensive chart of letter forms from the Bath Curse Tablets, and discussions regarding these characters, can be found in Tomlin (1988).
[2] This form of writing has also been called "Scrittura usuale" (Cencetti 1950), "l' écriture commune classique" (Mallon 1952), "Ancient Latin Cursive" (Thomas 1976), and "Ancient Roman Cursive" (Tjader 1986). It will be called ORC here for the sake of consistency.
[3] The primary sources for ORC, aside from the Vindolanda Corpus, are 200 tablets from Pompeii and Herculaneum (pre A.D. 79), tablets from Dacia (A.D 131-67), 45 tablets from North Africa (5th Century A.D.), a variety of tablets from Egypt (1st- 4th Centuries A.D.), 400 tablets from Vindonissa in Switzerland (mid 1st Century A.D.), and a handful of tablets from Britain. These sources are discussed in Bowman and Thomas (1983, p. 33-7). A complete list of known examples was compiled by Marichal (1950), which he later updated (1955).
[4] For a comprehensive bibliography see Bowman and Thomas (1994), and Tjader (1977).

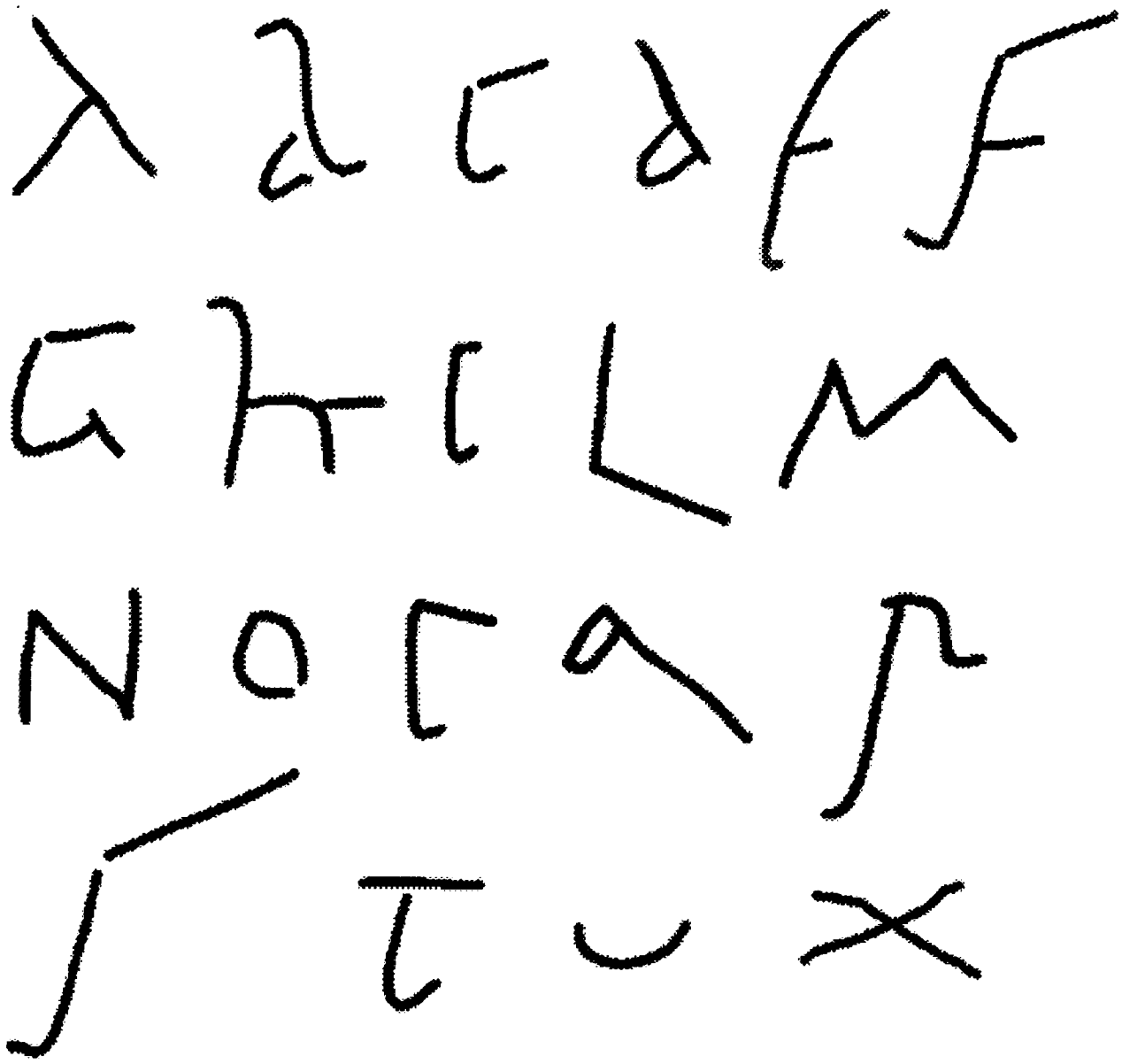


**Figure 3.1:** Characteristic letter forms in ORC, based on Bowman and Thomas (1983, p.54). The alphabet consists of 20 characters (not having the letters j, k, v, w, y, and z present in modern day English script).

From around 300 AD ORC was replaced by New Roman Cursive (NRC), often referred to as "miniscule cursive"; the transition between the two being the focus of much academic debate (see Tjader 1979). (NRC led indirectly to Carolingian miniscule in about 800 AD, and so to the script we employ today). Although the documents from Vindolanda are from a much earlier period than that of NRC, the ink and stylus tablets provide an excellent palaeographical source for studying the development of ORC Latin script, and the use of ORC in Roman Britain. Bowman and Thomas explain:

> only a few of the tablets will actually have been written at Vindolanda, but the remainder are hardly likely ... to have travelled far, and we may fairly take the find as exemplifying the type of writing in use in Britain at this period. We thus have now for the first time a not

> inconsiderable body of written material from a part of the Empire from which hitherto virtually no such material had come to light (1983, p.52).

Bowman and Thomas provide a comprehensive discussion of the individual letter forms and other general palaeographic detail regarding the ink tablets (1983; 1994).

### 3.1.2 The Palaeography of the Stylus Tablets

In order to aid in the reading of the stylus tablets, it is important to know what type of letter forms they will contain. However, due to the fact that only a handful of stylus tablets have been read so far, it is impossible to provide a traditional palaeographic review of all the letter forms used within them. Those that have been read utilise ORC script, having the same letter forms as the ink tablets. This is fairly unsurprising as they are contemporaneous documents from the same source, albeit more official in nature. There has been some discussion of different forms of letters being employed in official documents (Bischoff 1990), but this tended to apply to comparisons involving papyri and stylus tablets, and there is little evidence that this was the case in Vindolanda. Cencetti notes that around this period the script in legal documents as well as in military and civilian administration took on a very uniform character (1950). The many individual hands (over three hundred) present in the ink tablets, which are both official and civilian correspondence, utilise the same script, showing "that there was a single, standard type of script in common use, and that this script could be written both quickly and slowly" (Thomas 1976, p.41). There is no reason to think that the letter forms in the Vindolanda stylus tablets will differ wildly from those in the ink tablets, aside for individual cases, such as the letter E, which is written in a different format on wax in most cases.

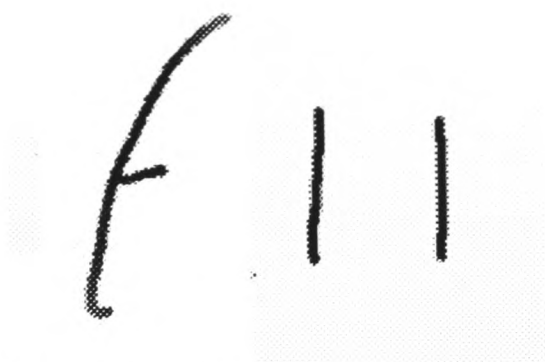


**Figure 3.2:** Letter E. On the left is the form commonly found on the ink tablets. On the right is the form commonly found on stylus tablets. The ink form is occasionally found on the stylus tablets. There has been one instance, so far, of the stylus form being found in the Vindolanda ink texts.

However, it cannot be ignored that the writing-surfaces of the two types of documents differ greatly (see 1.1). Earlier comparisons between writing on papyri with ink, and the inscribed writing of a stylus on wax indicate that there are differences in the process of writing between the two: "In writing with a stilus upon wax there is a natural tendency to make short disjointed strokes" (Van Hoesen 1915, p.3). Thompson talks of the "smooth but clinging surface of wax scratched with the point of the stilus, or the less impeding ... wood ... inscribed in ink with the reed" (1912, p.315), giving examples of the different letter forms found on documents at Pompeii:

> we have writing on two kinds of material, and different accordingly; that of the deeds themselves, incised on the waxed pages with the stylus in decidedly cursive characters; and that of the endorsements and lists of witnesses, written in ink upon the bare wood of the pages which were not coated with wax, in a generally more restrained style and employing other forms of certain letters ... the natural tendency, in writing on a resisting or clinging surface such as wax, is to turn the point of the writing implement inwards and hence to slope the letters to the left. The letters employed by preference, where a choice is possible, would usually be those which are more easily written in disconnected strokes, such as the two-stroke E and the four-stroke M ... On the other hand, we find here the ordinary capital N ..." (Thompson 1912, p.319).

### 3.1.2.1 Forensic Palaeography and the Stylus Tablets

A few papers exist which investigate the effect different writing implements have on handwriting, a topic of interest in forensic science and criminology (Hilton 1959; Hilton 1984; Masson 1985). The most useful study is detailed by J. Mathyer in his paper "Influence of Writing Instruments on Handwriting and Signatures" (1969), in which he systematically compares 98 standards prepared by 14 persons using 7 different writing implements. Mathyer aimed to investigate an hypothesis suggested by Edmond Solange-Pellat (1927):

> Le geste graphique est sous l'influence immediate du cerveau. Sa forme n'est pas modifée par l'organe scripteur si celui-ci fonctionne normalement et se trouve adapté à sa fonction[5].

Mathyer considers the role of the

> writing organ, ie the fingers and hand (right or left) for a person writing on a sheet of paper, the hand, forearm, the arm and the shoulder for a person writing at a black board (p. 106),

demonstrating that a change in the physiology of the writing process has little effect on the writing produced, indicating that it is a higher level cognitive process. He also graphically shows how a change in instrument does not effect a writer's handwriting, adapting the earlier hypothesis to his findings:

> *The form and line quality of the handwriting of a person is not modified by the writing instrument if this one works normally.*
>
> When the writing instrument does not work well... it can of course contract characteristics which indicate that the writer has tried with more or less success to obtain an acceptable result with a bad instrument... The influence of the instrument is rarely very important; this influence is frequently non-existent (p.106).

It can be presumed that, for the most part, the writers of the ink and stylus tablets try to use the same letter forms on each. Mathyer's paper indicates that differences in

---

[5] Mathyer translates this as "The graphic movement is under the immediate influence of the brain. Its form is not modified by the writing organ if this one works normally and is sufficiently adapted to its function" (p. 105).

letter forms between writing on wood with ink and incising the letters in wax should be minimal. Although there may be some variance caused by the different mediums, the scribes would have had the *intention* to write the same type of letter forms, and for the most part, these should be similar (aside from the letter E, as shown above. This, indeed, is demonstrated in section 3.10.) These small differences, however, would prove enough to throw off any existing image processing techniques that could be used to scale, find, and match letter forms found in the ink tablets to those on the stylus tablets. Each individual usage of the characters is not identical in the ink tablets: it is human handwriting that is being dealt with here, not machine printed text. Instead, it was necessary to find a way of modelling the existing letter forms, which captured the types of knowledge the papyrologists/palaeographers employ when identifying and discussing individual letter forms. In doing so, it would provide a way of representing the letter forms contained within the stylus tablets, and so aid in their identification.

## 3.2 Knowledge Elicitation and Palaeography

The aim of this knowledge elicitation exercise was not to collect and summarise information regarding each individual letter form, which has more than adequately been covered in Bowman and Thomas (1983, 1994, Forthcoming (2003)), but rather to enquire: how do the papyrologists identify individual letters? What are the characteristics and relationships of strokes that are noted by the palaeographers, which build up to make a character? Is there a way of modelling this information, so that a formal way of documenting characters can be developed, which would enable the information to be eventually processed by a computer (modelling being the first step in documenting data structures so that they can be processed

automatically (de Carteret and Vidgen 1995)). Developing an encoding scheme would also enable a training corpus to be constructed (the use of corpora in building trainable intelligent systems being a growing trend in the field of Artificial Intelligence, particularly in natural language processing (Charniak 1993; Lawler and Dry 1998), and computer vision (Robertson 1999; Robertson and Laddaga Forthcoming (2002))). A program of knowledge acquisition and elicitation (see 2.2) was undertaken to gain an understanding of the types of information the experts use when discussing and reasoning about letter forms.

### 3.2.1 Textual Sources

The major textual sources which deal with the letter forms from Vindolanda are, of course, Tab. Vind. I and Tab. Vind. II (Bowman and Thomas 1983, 1994). These were read in detail, and formed the basis for the information collected about the letter forms. There were additional associated articles that proved helpful, by Cencetti (1950) and Casamassima and Staraz (1977), which deal with other sources of ORC, but detail the letter forms closely. Older sources, such as Thompson (1912) and Van Hoesen (1915), although somewhat outdated, still provided useful information regarding the reading of letter forms.

A secondary source of information was generated from the text of Tab. Vind. II. All instances of discussions of individual letter forms from within the published commentaries of the Vindolanda ink texts (see 2.3) were collated, utilising WordSmith (see 2.4.2). This provided instances of discussion regarding the identification of individual letters, particularly where this identification was in doubt. Likewise, each instance of discussions of individual letter forms from within

the Think Aloud Protocols (see 2.3.2) was collated, providing information regarding the letter forms on both the ink and the stylus texts[6].

### 3.2.2 First Stages in Knowledge Elicitation

To gain an understanding of the field, each expert was approached and asked to discuss at length the letter forms that are present within the Vindolanda texts. At first these were general discussions to gain an insight into the field. The experts were then interviewed using a semi-structured interview technique and asked to focus on specific issues, such as:

- individual characters
- standard forms of characters
- deviations from standard forms
- the physical relationship of characters to one another (ligatures, serifs, etc)
- characters which are often confused with another
- specific features, such as ascenders and descenders
- the identification of similar hands in different documents.

This resulted in a large volume of data regarding the different characteristics of individual letters, and it was necessary to resolve this into a simpler form. This meant an encoding scheme could be developed that would enable a corpus of letter forms contained in the Vindolanda texts to be constructed. This, in turn, could be used to aid in the reading of the stylus texts (see Chapter 4).

[6] These are available on the accompanying CD-ROM in Chapter 3/letter forms/.

### 3.2.3 Use of Repertory Grid

To aid in aggregating the data regarding characteristics of letter forms, a Repertory Grid was employed. Repertory Grids are based on Personal Construct Theory (Kelly 1955), where individual concepts, or "elements", are defined by the way they are alike or similar to other concepts. The identifying topics are the constructs in the relationship[7]. By using a Repertory Grid to accumulate data regarding the Vindolanda letter forms, the main constructs become apparent, highlighting the most important characteristics of letter forms mentioned in all the available data. (It should be noted that the Repertory Grid was not used with the experts, which would be another possible means of capturing their knowledge regarding letter forms, as there were limitations on time and availability. In future, this may be another tool that can be used. Here, it was used by the Knowledge Engineer as a means of accumulating information from different varied resources.)

The program used was WebGrid[8], developed by Gaines and Shaw (1997) at the Knowledge Science Institute, University of Calgary. WebGrid is a web-based knowledge acquisition and inference server that uses an extended repertory grid system for knowledge acquisition, inductive inference for knowledge modelling, and an integrated knowledge-based system shell for inference. In this case, it was used primarily for knowledge acquisition and modelling, to build up a detailed understanding of the constructs utilised when discussing the letters used in the documents at Vindolanda.

---

[7] For example, Border Collies, Dachshunds, and Xoloitzcuintles are all breeds of dogs (and therefore individual "concepts" or "elements"). They have different characteristics such as height, weight, length of coat, intelligence, athleticism, rarity, etc. These are "constructs", in that they can be compared to each other on a numerical scale to build up a profile of each type, or concept, of dog. In this way, an individual concept can be defined by comparing it to other concepts.

[8] http://tiger.cpsc.ucalgary.ca/Webgrid/webgrid.html

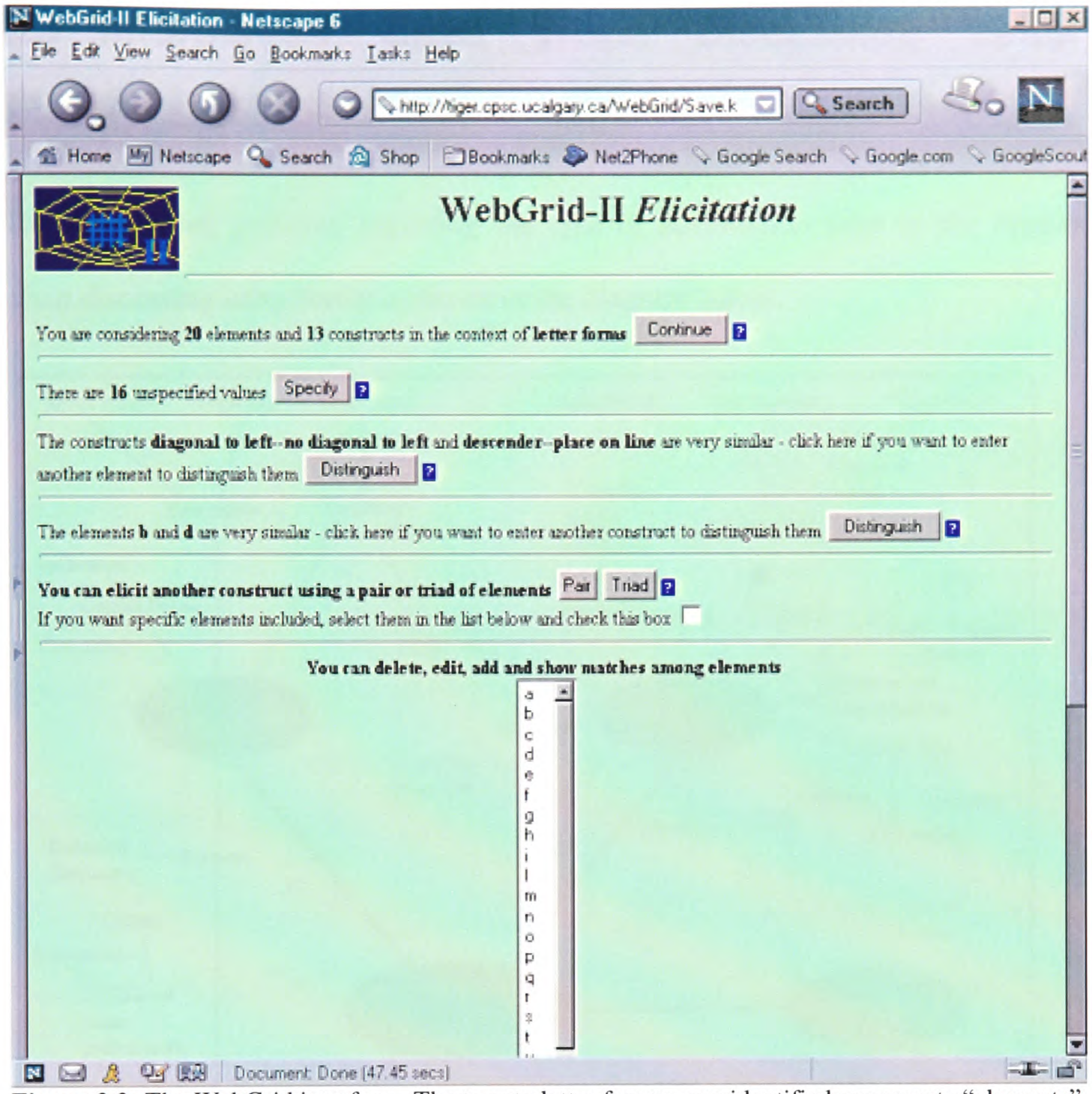



**Figure 3.3:** The WebGrid interface. The twenty letter forms were identified as separate "elements", and the characteristics identified as "constructs". The program identifies areas which need qualification; in this case it recognises that the characters B and D are very similar, and asks the user to add another construct to distinguish them. In this way a list of constructs was built up which encompasses all the information included in the various sources.

Once the data was captured in this manner, it was resolved into an overall schema which describes the types of information experts refer to when discussing and identifying individual letter forms. This is shown below.

## 3.3 Information Used When Discussing Letter Forms

The information gathered regarding the type of information used by the experts when discussing letter forms is shown in the diagram below.

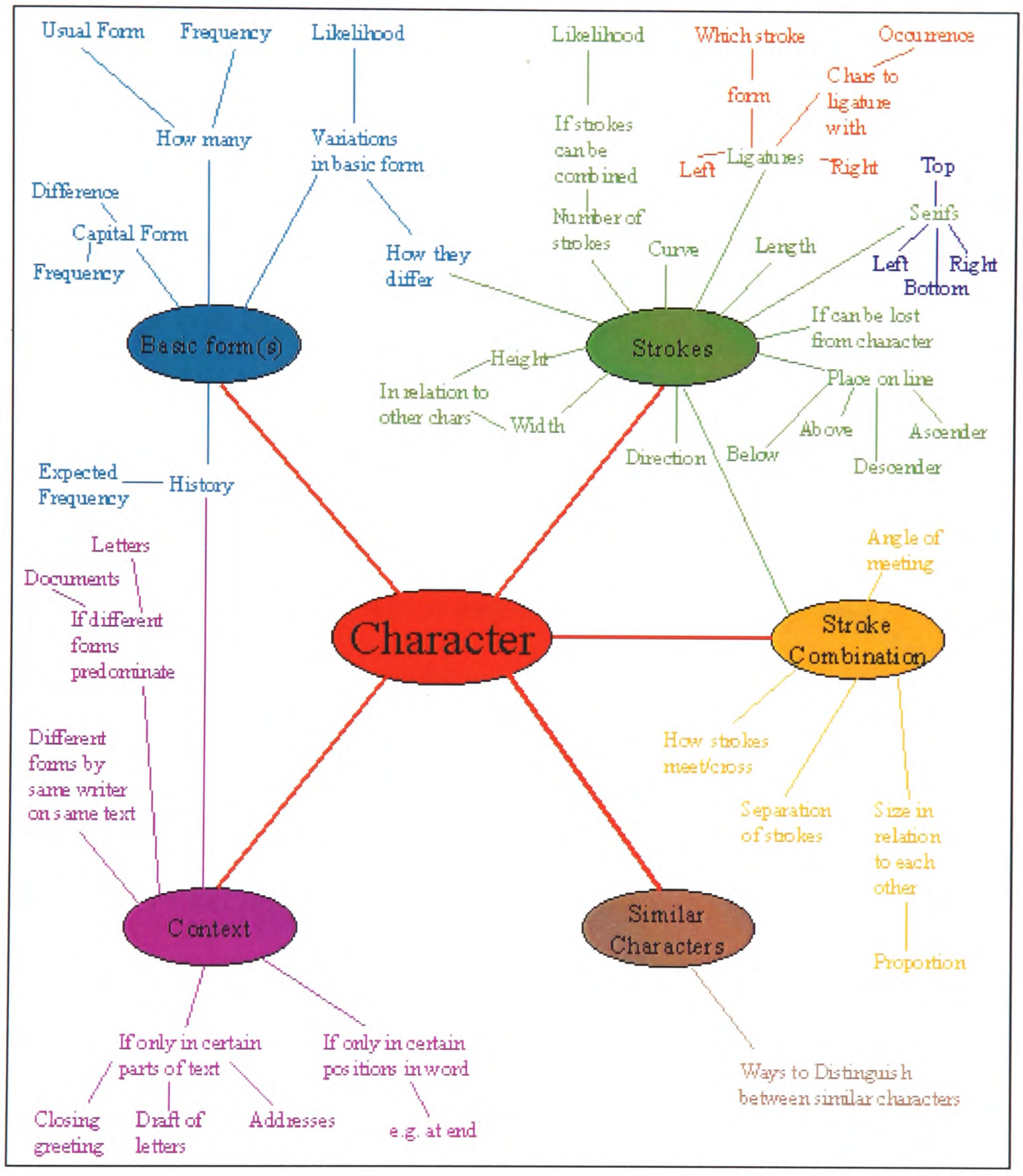



**Figure 3.4:** Information used when discussing letter forms, presented as a semantic network[9].

[9] A similar approach to the collection and representation of information in this manner can be found in Connell and Brady (1987).

Each character has one or more basic forms. These consist of one or more strokes, which combine to make a character. Character forms may differ depending on the context they are placed in, and may also be hard to distinguish from similar character forms.

There may be more than one basic form, one of which may be more usual or common than the others. There may also be a different capital form which is used in the texts. These may differ substantially from the lower case letter, or may be very similar. The upper case character may be more or less frequent than the lower case. There may be variations on the basic form(s), which differ slightly on the stroke level. Some variations will be more or less common than others.

Each character consists of one or more strokes. Character forms will have a usual number of strokes, some of which may be combined in the variations on the basic form. The strokes will be made in a certain order (although it is difficult to definitely prove the order in which strokes were made). Sometimes, it is possible to see where characters are made in a different way or order from the standard. Some strokes can be lost from the character in some variations in letter form: the fact that this stroke is missing not hindering the identification of the letter.

Each stroke has a height, width, and length. These can also be compared with other strokes, possibly in other instances of the same character or even in other different characters within the same word, line, or document. Each stroke also has a relative direction and curve. Strokes (and characters) have a place on the base writing line: ascenders and descenders can give important clues as to character identification.

Strokes can also be (or have) ligatures and serifs. Ligatures join one character to another in a fluid writing motion. They may have a particular form from a certain stroke in a letter, and may be to the right, or from the left. Each character form has common letters that it will ligature with, and letters that it will rarely ligature with. A serif is a short decorative foot at the end of a stroke. They may be to the right or left, at the top or bottom of a stroke, and occur in some characters more often than others. Some character forms can be written with or without serifs or ligatures. Some characters never have ligatures or serifs. Stroke endings with no ligatures or serifs can either be blunt, or slightly hooked.

Strokes combine together to make character forms. Strokes which constitute a character will have a relationship to each other. How the strokes meet or cross is important: strokes may meet end to end, or cross over each other. Alternatively, there may be a gap, or separation, remaining between strokes. This meeting, or junction point (whether closed or open) will have an angle. The size of strokes, in relation to each other, can also be important, as the proportion of one stroke when compared to another can impart information which leads to a possible identification.

Each character form may vary depending on the context it is placed in. In some rare cases, it has been observed that letter forms vary depending on the position the characters have in the word: the final letter may take a different form than the same letter earlier in the word. Also, it has been suggested that where the letters are found in the text can have an effect on their form: there are types of letters which have only been found in closing greetings, drafts, or addresses. There may be some predomination of certain forms of letters in official documentation, which are rarely

used in private letters (although this area requires much further research[10]). The development of letter forms can be traced throughout different texts, and the context of the letter can give some indication as to the form which would most likely be expected, as some forms are historically used in certain circumstances. Different forms of letters are commonly found on the same text written by the same scribe, with no apparent reason for the difference in usage (a common feature of handwriting no matter which language is being used).

Some characters may often be confused with others. For example, the letters B and D are often confused because of the similarity of their form[11]. When the use of specific features allowed these confused characters to be differentiated, they were noted.

Some additional information was collected regarding formatting which is not represented in the diagram above. For example, interpuncts are sometimes used as word separators. Indentation is also used in the texts to indicate where a section begins. Word spacing is also used (although not in all documents).

All of the above information has been discussed in relation to the Vindolanda letter forms, and can be used when trying to identify an unknown character. To be able to capture this data regarding individual letters, an encoding scheme was constructed,

[10] It is certainly true of the so-called "Chancery Hands" found in documents of the later Roman period.

[11] It is this similarity that would throw a character-by-character pattern recognition system, as it would plump for what appears to be the most likely individual solution. Since the characters are so similar "noise would dominate signal" – that is, small changes would sway the interpretation. In the approach adopted in this thesis, the disambiguation is left to a higher level knowledge, increasing the accuracy of the interpretation, as discussed and demonstrated in Chapter 4.

based on the above schema, which would allow images of the Vindolanda texts to be annotated in a manner that would encapsulate important character information.

## 3.4 Derived Encoding Scheme

Before being able to annotate any images, it was necessary to strip down the general model of the type of information discussed in regard to character forms, to provide a more linear structure consisting of broad headings that could easily be applied to images and their constituent parts. The complex relationship of information in the above diagram was resolved as follows:

Each area of space in the image is either a

- Character box (the area surrounding a collection of strokes which make up a character)
- Space Character (indicating the space between words)
- Paragraph Character (indicating areas of indentation in the text)
- Interpunct (indicating the mark sometimes used to differentiate words)

Each letter is comprised of strokes

- Traced and identified individually (one, two, three, etc. Although it is arguable whether you can identify stroke order from a letter form, when this was obvious the strokes were annotated in this manner. When not obvious, they were annotated from left to right.)

Each stroke has end points, either

- blunt

- hook (with direction specified, i.e. up left, up right etc)
- ligature (with direction specified)
- serif (with direction specified)

Each character may also have stroke meetings, either

- end to end
  - exact meet
  - close meet
  - or crossing,
- or middle to end
  - exact
  - close
  - crossing
- or crossing (middle to middle).

These were the most important set of characteristics, and were incorporated into image annotation software (see below 3.6) to be able to graphically encode this data. However, there was much more information available regarding letter forms. For example, one of the most important characteristics which help identify letters is whether they have strokes ascending above or descending below the writing line; the letter S is often easily recognisable because of its long descenders. A further collation of all features and types of strokes mentioned by the papyrologists as they discussed the texts was undertaken, and resolved into a textual schema. This was used to provide additional tags to the annotated regions to help in labelling them further. These tags were added manually, in tandem with the annotations made with the graphical interface.

Each Character Box had a letter identification (*) and an overall size height (SH) and width (SW). Letters that had not been confidently identified by the papyrologists were marked with a question mark (*?). Letters which were expected by the papyrologists but missing due to damage of the texts were marked with a character box (with no strokes included) but annotated with an exclamation mark, for example "*s!".

Each stroke was assigned additional textual tags, having:

- A direction (D), giving the stroke orientation and type, which included
  - Straight Strokes (DS)
  - Simple Curved Strokes (DC)
  - Complex curved Strokes (DCW)
  - Loops (DL).
    - Each of these tags included further orientation tags to indicate left, right, etc (orientation being dictated by taking a centre point of the Character Box and deriving from that up, down, left, right, etc). For example, DSdl represents a straight stroke down to the left.
- A Length (L), being either comparatively short, average, or long.
- A Width (W) being either comparatively thin, average, or wide.
- A Place (P) on the writing line, being either within the average, Descending, or Ascending.

Each stroke meeting, or junction, had

- an Angle (A), with note taken of the orientation of the meeting and whether the angle was acute, right, obtuse, or parallel.

If the strokes were broken (due to damage of the document) this was noted in the stroke ending field, where they were labelled as being blunt with the additional "broken" textual tag added.

A full textual representation of this encoding scheme is available in Appendix A.

## 3.5 Building the Data Set

Seven ink tablets were identified[12], with the help of one of the experts, which would provide good, clear images of text to annotate, and have enough textual content to provide a suitable set of test data. Of these seven, 225b and 225f have been identified as being by the same scribe (being two sides of the same document), and 248 and 291 were also identified as by being by the same hand (although this is different to the scribe who wrote 225). The three other ink texts are in different hands. Additionally, closing valedictory comments are present in some of the documents written by the hand of the author rather than the scribe, providing further different instances of letter forms. This sampling provides enough data to ensure that different letters forms and styles are covered in the training set.

[12] 225b, 225f, 248, 255, 291, 309, 311 from Bowman and Thomas (1994). Larger images of these are available on the accompanying CDROM in Chapter 3/images/

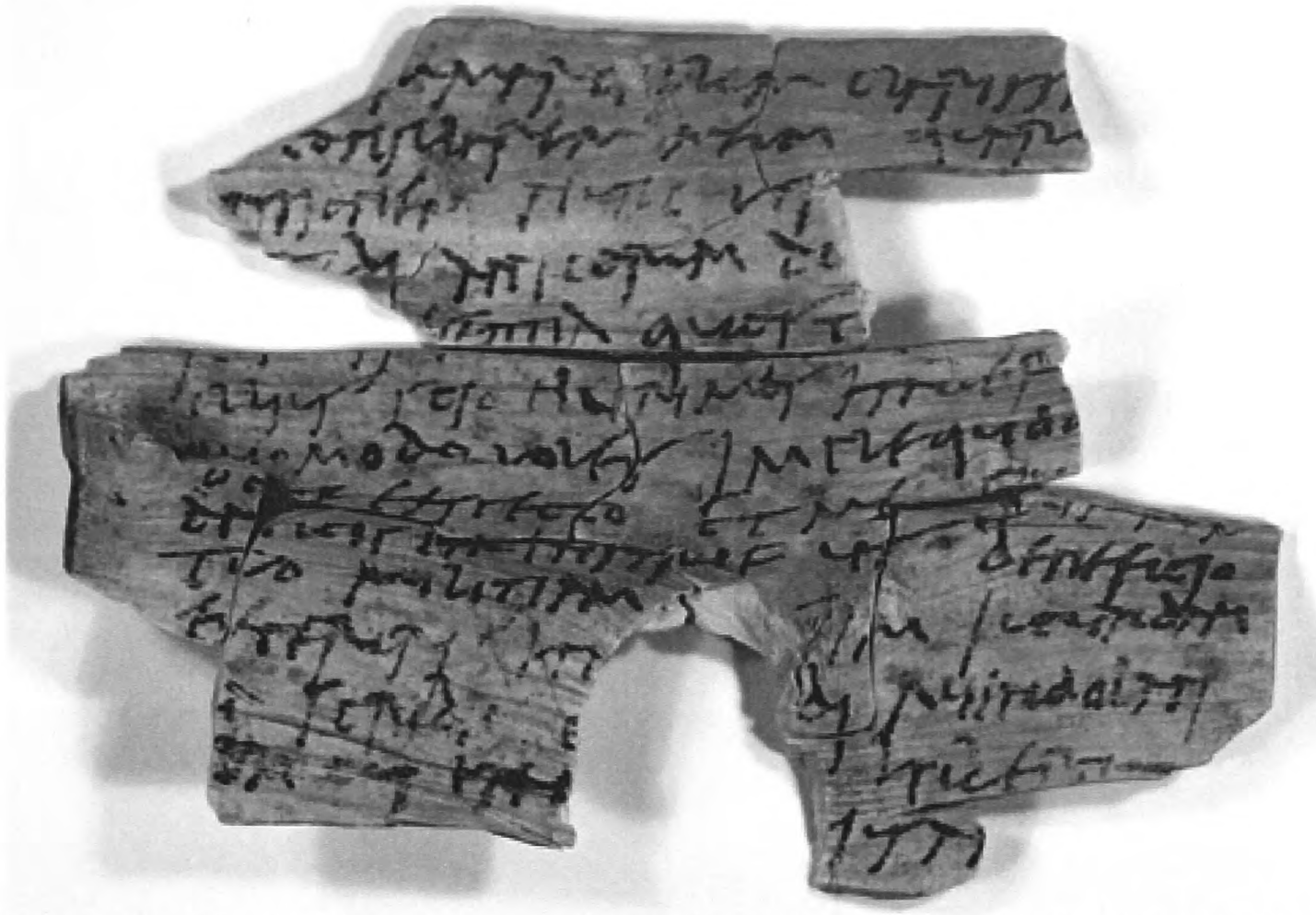


**Figure 3.5:** Ink text 225b, used as a source of ink tablet characters. Although displaying considerable damage, this tablet contains variation in letter forms and is a good source for the training set. The hand is confident and inelegant.

**Figure 3.6:** Ink text 225f. The other side of 225b. Again, this document displays a diverse selection of type and form of characters, in the same hand as above.

**Figure 3.7:** Ink text 248. A more clearly written letter on a complete leaf in two columns, with different character forms from the tablets above. Note the long ascenders and descenders.

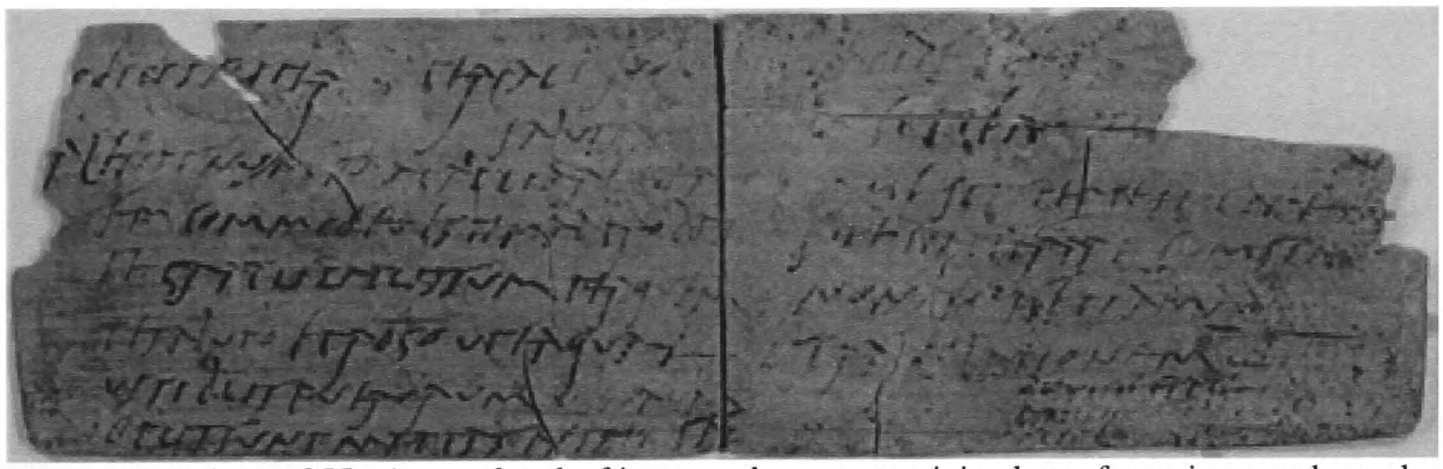


**Figure 3.8:** Ink text 255. A complete leaf in two columns, containing letter forms in a good, regular cursive hand.

**Figure 3.9:** Ink text 291: A diptych written in a slim, elegant script.

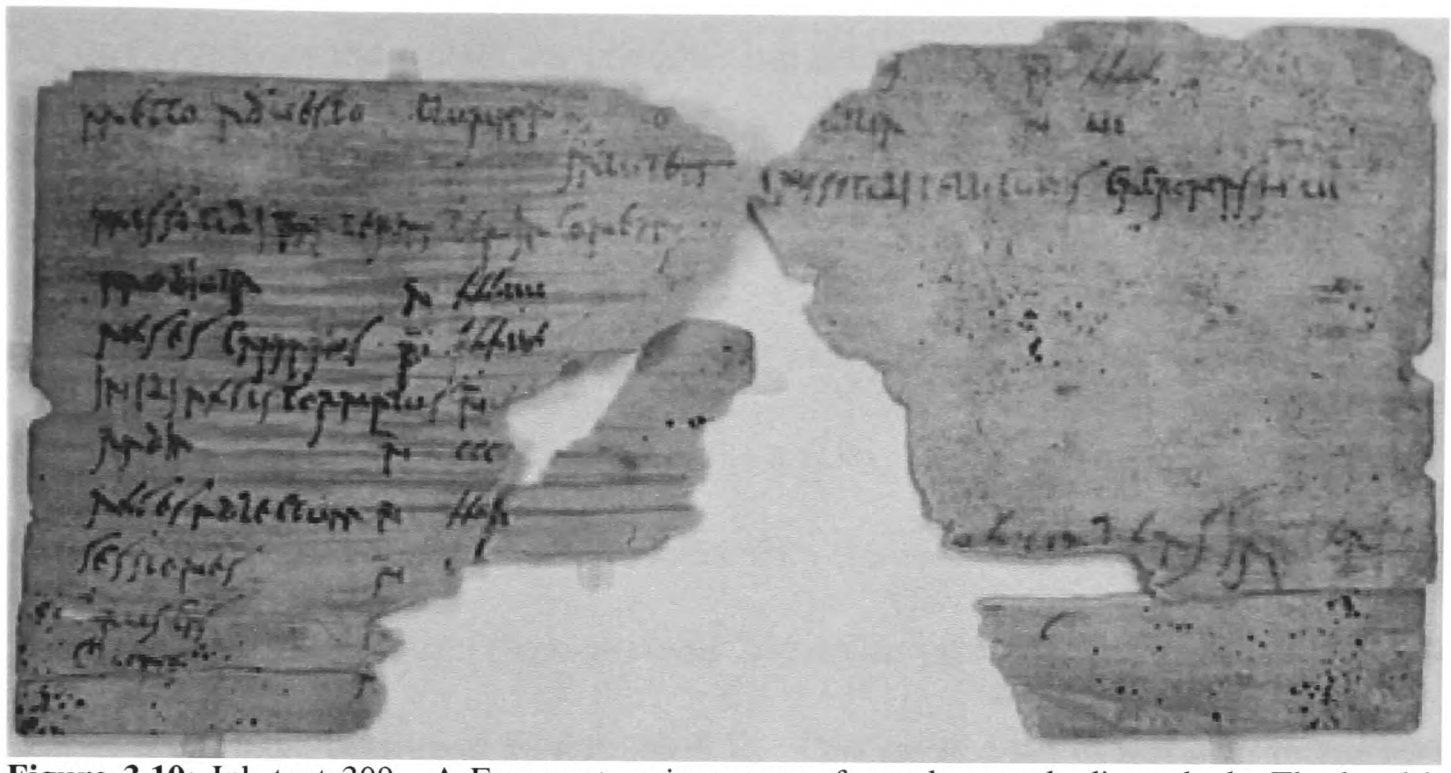


**Figure 3.10:** Ink text 309. A Fragmentary inventory of wooden goods dispatched. The hand is rather inelegant.

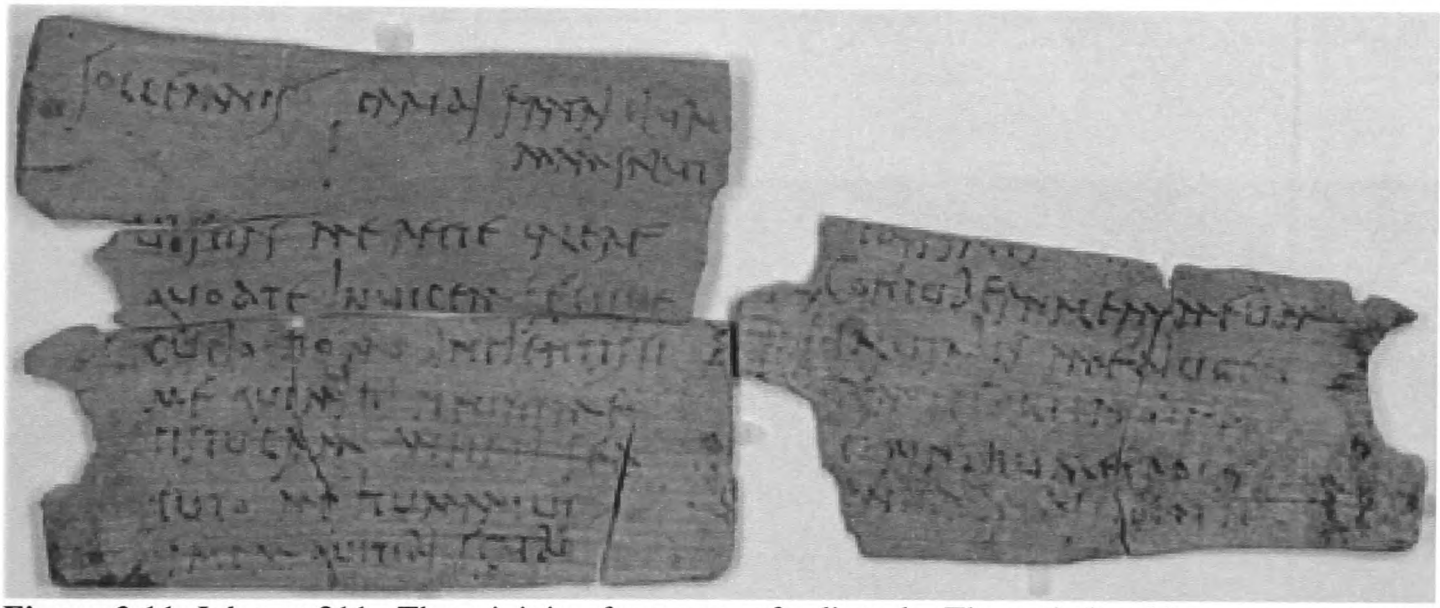


**Figure 3.11:** Ink text 311. Three joining fragments of a diptych. The main hand is a very competent and interesting squarish cursive, with carefully spaced words, and little use of ligatures.

These seven texts are mostly personal correspondence (225b, 225f, 248, and 255 being taken from the correspondence of Flavius Cerialis, the prefect of the Ninth Cohort of Batavians, and 291 from the correspondence of his wife, Lepidina. 311 is a letter from Sollemnis to Paris, who was possibly a slave of the commander). The only example that is not a letter is 309, being the account of supplies, (mostly manufactured wooden items) which were being delivered to Vindolanda. Although it is expected that the stylus tablets will contain slightly more formal subject matter,

the difference in subject matter between the ink and stylus tablets should not matter for this corpus as the reason for constructing this training set was to provide a test set to compare letter forms, not words.

Two stylus tablets were also annotated, 974 and 797. 974 possibly concerns some transactions involving the manumission of a slave, 797 is a fragment of a personal letter. These were chosen simply because they are two tablets that the experts have managed to read (although there are over 200 stylus tablets from Vindolanda only a few of them have been read so far, see 1.1). This gives a set of test data to see how effectively data regarding the character forms from the ink tablets can be used to aid in the reading of character forms from the stylus tablets (see 4.8).

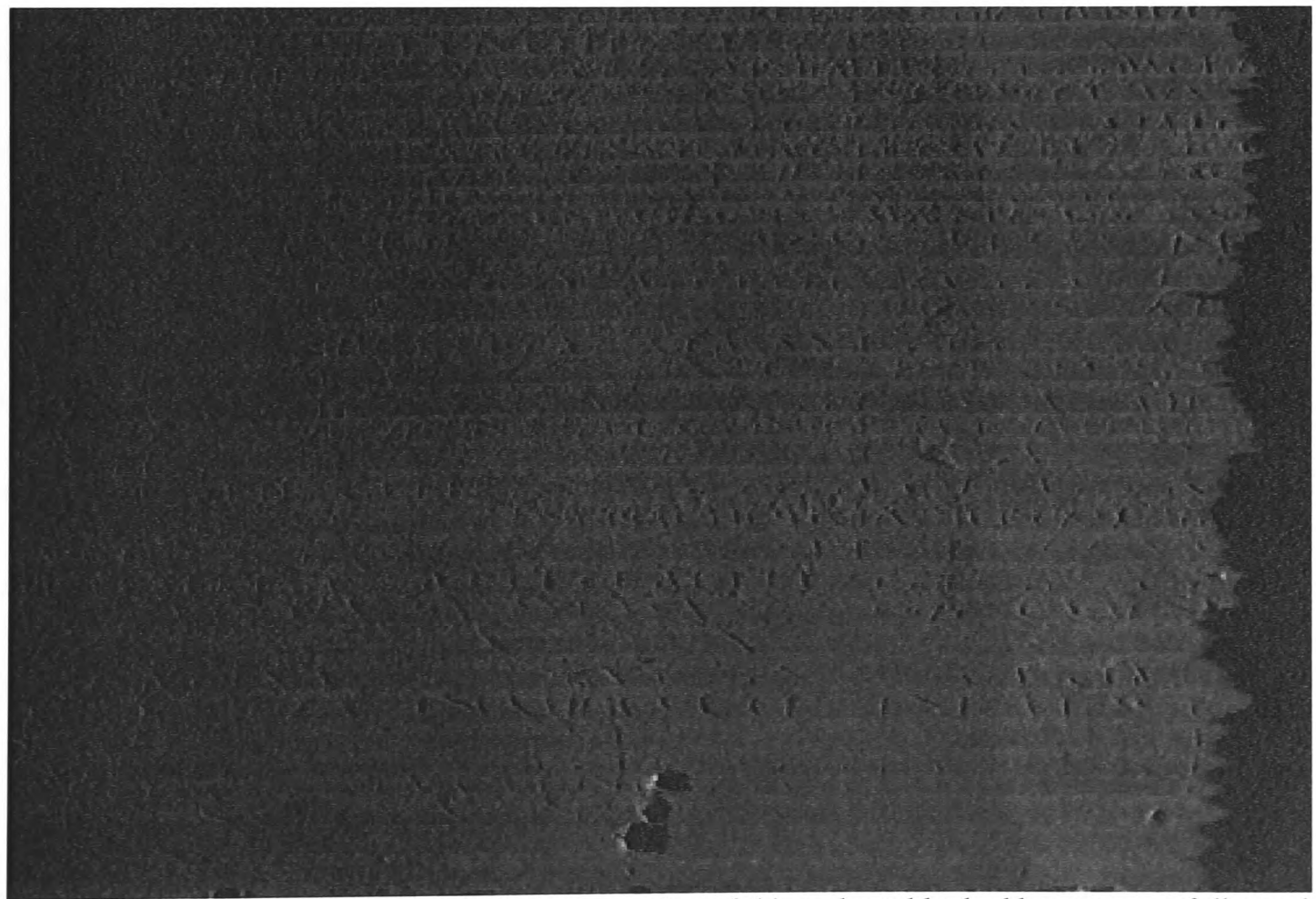


**Figure 3.12:** Stylus tablet 974. A fairly large portion of this stylus tablet had been successfully read, allowing the last few lines of text to provide examples of letter forms contained on the stylus tablets. Comparing this image to the previous images of the ink texts indicates the difficulties the experts face when trying to read such a text.

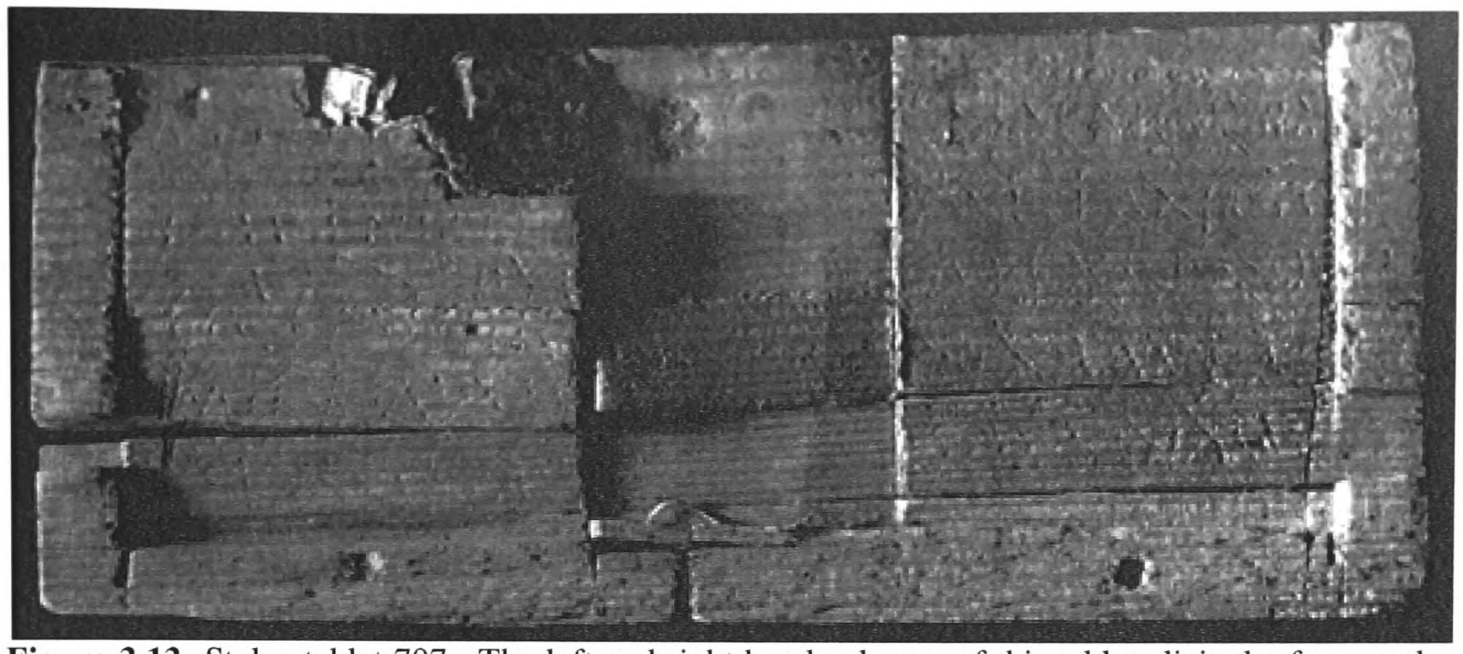


**Figure 3.13:** Stylus tablet 797. The left and right hand columns of this tablet elicited a few words, giving a small sample of characters.

One expert was provided with large images of the chosen texts, on which he traced the individual strokes of each letter as they appear on the ink tablets. These images were married up with the published texts (to confirm the text that had actually been read; letters such as R and A are easily confused) and the information gleaned from the knowledge elicitation exercises regarding character forms and formation to provide the information with which to annotate the images. Although this does involve some interpretation of the data by the knowledge engineer, it was done in as methodological and systematic a fashion as possible. The expert's tracings of the character forms were taken as the main source of information regarding the texts. Care was taken to annotate the images without the addition of personal bias by retaining a distanced stance of annotating exactly what was there, rather than what *should* be expected. The annotations were also double checked by cross referencing, as discussed below (3.7.2).

## 3.6 The Use of the GRAVA Annotator

To be able to annotate the images using this encoding scheme, an annotation program was used. This was an adjusted version of Paul Robertson's GRAVA Annotator Program, a Motif[13] program originally developed to manually segment and label a corpus of aerial satellite images[14] (see Robertson, 2001, Appendix C). The results of the annotation are written to a file in an extended SGML format, to allow easy interaction with other programs. Robertson's software was adapted to incorporate the encoding scheme, detailed above, that was relevant to the Vindolanda texts, to allow these to be labelled in the same manner.

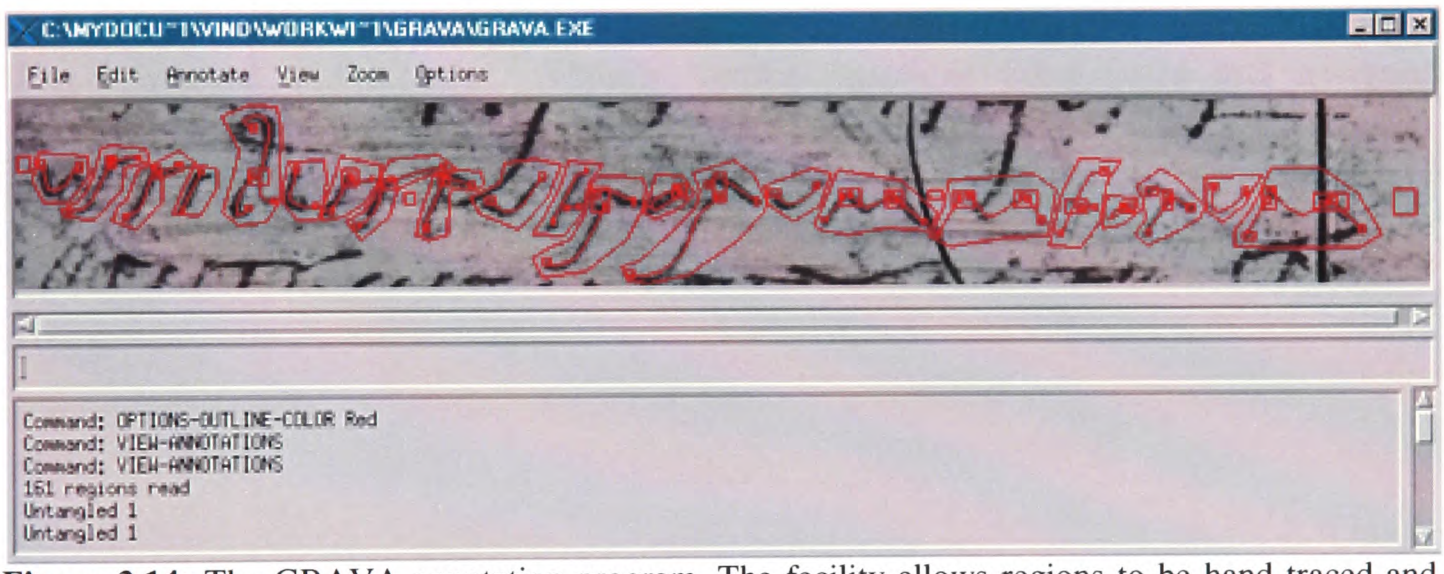



**Figure 3.14:** The GRAVA annotation program. The facility allows regions to be hand traced and assigned a label. The application screen is divided into a menu bar, image display pane, command input area, and message area. Most interaction with the annotator is performed using the mouse.

[13] Motif is the industry standard graphical user interface environment for standardising application presentation on a wide range of platforms. Developed by the Open Software Foundation (OSF, see www.opengroup.org) it is the leading user interface for the Unix system. In this case, we utilized Exceed, an X-Server for the PC, to run our Motif application on the Windows platform.

[14] The corpus produced in this case was subsequently used to train a system to segment, label, and parse aerial images so as to produce an image description similar to that produced by a human expert.

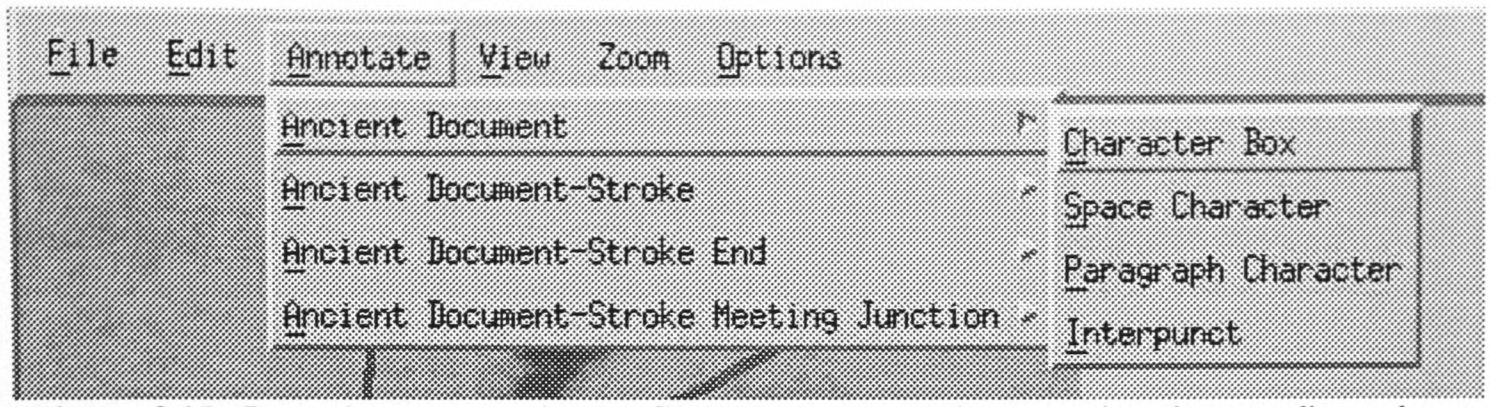



**Figure 3.15:** Drop down menus in the GRAVA Annotator, incorporating the encoding scheme determined from the knowledge elicitation exercises.

## 3.7 Annotating the Characters

Characters were annotated by firstly drawing around the outline of a character with the mouse, and assigning it a character label. Individual strokes were then traced, and numbered by selecting the option from the drop down menu. Stroke ends were then identified and labelled. Finally, stroke junctions were noted and assigned labels. An example of this is shown below, using the letter S from the start of 311 as an example.

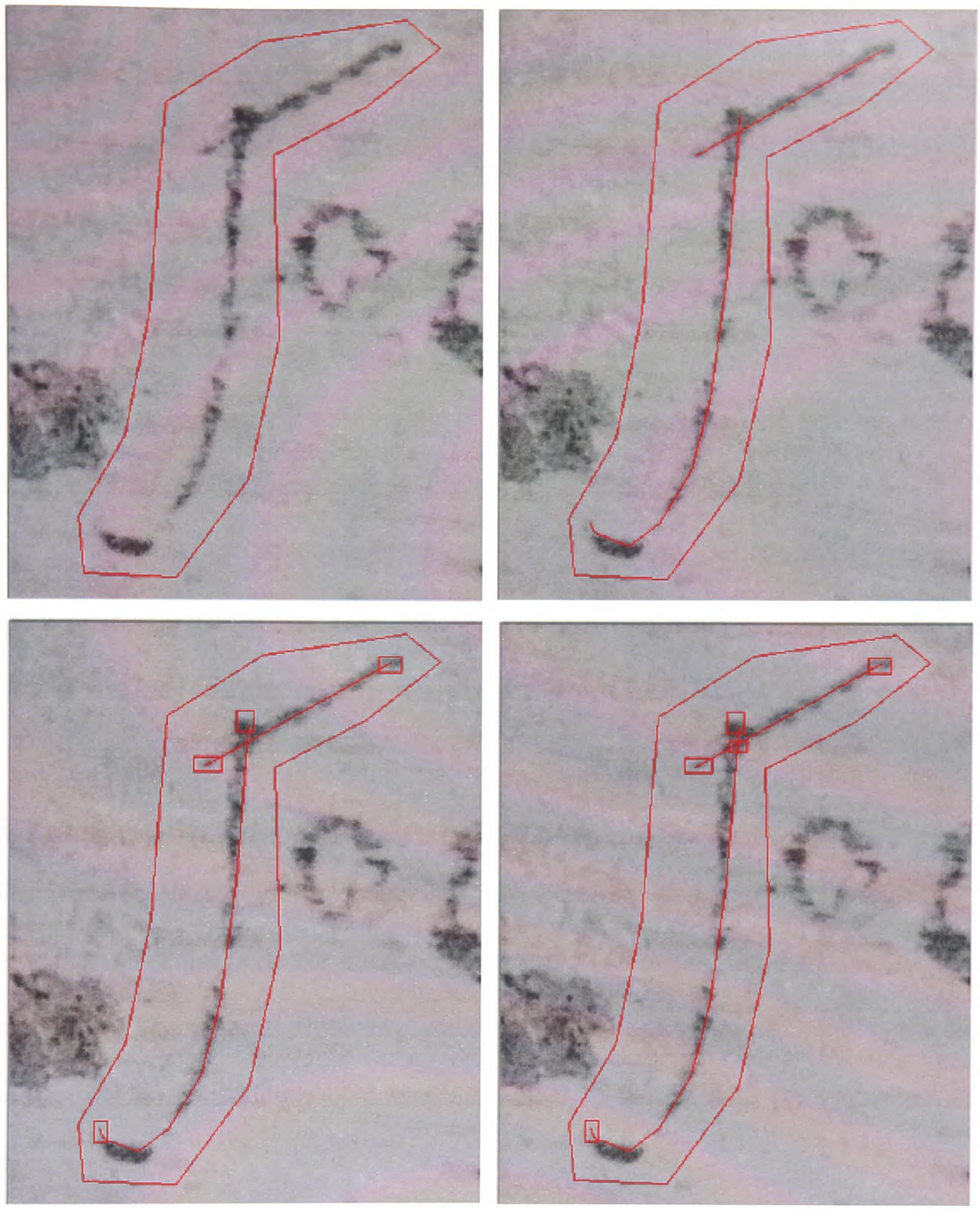


**Figure 3.16:** Steps taken in annotating a letter. The outline is traced at first, followed by individual strokes, stroke endings, and stroke meeting junctions. All are assigned labels from the drop down menus in the program.

These annotations are preserved in an SGML file which textually describes the annotated image. For example, this is the SGML which describes the bounding character box of the letter 'S' in the above example:

```
<GTRegion author="Melissa Terras" regionType="ADCHAR0" regionUID="RGN0"
regiondate="04/04/02 18:07:16" coordinates="112, 142, 106, 223, 87, 317, 52,
377, 56, 420, 126, 423, 180, 349, 203, 244, 199, 108, 269, 71, 323, 28, 298,
7, 190, 23, 118, 69, 112, 142"></GTRegion>
```

The region type tag "ADCHARO" defines this region as a character box which surrounds a collection of strokes, the region ID "RGNO" gives the annotated region a unique identification number, and the co-ordinates preserve the shape of the character box. Each individual region that is annotated (characters, strokes, stroke endings, and stroke meetings) has its own similar line in the SGML file, with each having a unique identifying Region ID, and a region type which specifies what type of annotation has been made. This file is structured hierarchically; all strokes, stroke endings, and stroke meetings "belong" to an individual character.

The structure of the SGML files generated is fully described in Appendix A.2. A table of Region Identifier codes is available in Appendix A.3.

## 3.7.1 Additional Annotations

The additional tags described above in 3.4 were included into the annotation by assigning a textual code for each region in the "comments" field of the GRAVA annotator. The letter S from 311, above, was deemed to be of large height and width, and so the additional comments added to the SGML file regarding the character box were:

```
comments="*s, SHl, SWl"
```

"*S" identifies the character as the letter S, SHl indicates that the height is large, and SWl indicates that the width is large. This gives the final resulting SGML output for this region as

```
<GTRegion author="Melissa Terras" regionType="ADCHAR0" regionUID="RGN0"
regiondate="04/04/02 18:07:16" coordinates="112, 142, 106, 223, 87, 317, 52,
377, 56, 420, 126, 423, 180, 349, 203, 244, 199, 108, 269, 71, 323, 28, 298,
7, 190, 23, 118, 69, 112, 142" comments="*s, SHl, SWl"></GTRegion>
```

All comment tags are provided in Appendix A.1. An example of a full SGML file which describes this letter S is presented in Appendix A.2.

### 3.7.2 Practicalities

Due to the fact that the images were very large, each document was split into a series of smaller images to allow easier annotation, on a line by line basis. Splitting the images in this way was also necessary due to the processing power needed to run the annotation program. There were 89 lines of ink text, and 21 lines of stylus text to annotate, resulting in 110 images in total.

At a later date the files of annotations were cross referenced with the texts printed in the published volumes to ensure that the annotations were correct. Any mistakes or areas of confusion were corrected, and the corpus updated. For example, the text "USSIBUS" had been incorrectly annotated as USSIBUSS" in a section of 255. This was corrected. In this case, a version with the incorrect annotation was retained to allow testing on the system, as described in 4.8.2. It is acknowledged that the annotations could contain some further examples of human error, but the annotations were carried out as systematically as possible, and this cross-referencing ensured that the majority of mistakes would be identified.

## 3.8 Results

Each image was annotated in the way described above. Care was taken to be as methodological as possible whilst undertaking this task. In total, there were 1506 individual characters from the ink tablets annotated, and 180 characters from the stylus tablets. The 9 completed sets of annotated images represented approximately

300 hours of work, with an average of around 6 or 7 characters being completed in an hour. The images were annotated and checked over a period of three months.

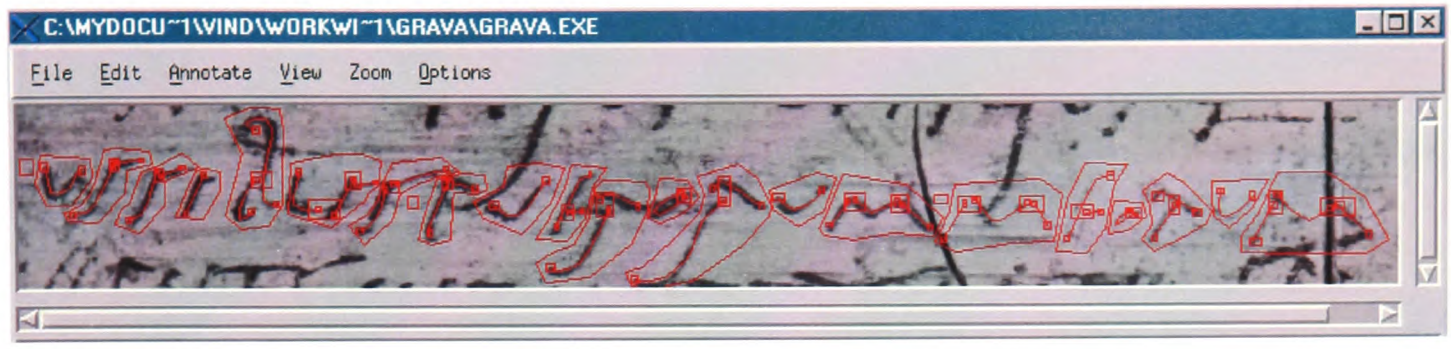



**Figure 3.17:** Completed line of annotated text, from ink tablet 255.

The resulting corpus can be accessed on the accompanying CDROM, as described in Appendix A. It will also eventually be displayed as part of the Vindolanda texts web site, along side readings and interpretations of the texts, which is currently under development at the Centre for the Study of Ancient Documents[15], University of Oxford.

Each individual character from the corpus is also represented in Appendix B, giving a full palaeographic representation of all of the character forms found within these nine texts.

## 3.9 Representativeness of Corpus

Due to the fragmentary nature of the Vindolanda corpus, it is difficult to be sure how representative the coverage, of the words and letters in the text corpus (see 4.2), and the characters within the image corpus, is. There are relatively few texts that the Vindolanda corpus can be compared with on a word level, as textual evidence from this period of the Roman Empire is rare. It is also difficult to obtain

the necessary statistics to allow any comparisons. However, on a broader level, it is possible to compare the Vindolanda Corpus with the largest corpus of written Latin, that of the Perseus Project[16].

The Perseus Project maintains a textual corpus of almost 7.8 million characters which provides comprehensive coverage of classical Latin. The corpus is primarily comprised of classical commentaries and histories[17]. Although the "vulgar" Latin in the Vindolanda corpus differs semantically and grammatically from classical Latin, on a letter by letter basis the frequencies of characters present should be very similar (as indicated by Zipf in his comparison of letter frequencies in the English Language (Zipf 1935, Reprinted 1965). A comparison between the letter distribution in the Perseus data set and the Vindolanda corpus can then be taken as a rough indication that the coverage of the image, and the textual, corpora are adequate. Statistics regarding the letter frequencies in the Perseus Corpus can be found in Mahoney and Rydberg-Cox (2001). These were compared to the coverage of the Vindolanda image corpus.

[15] www.csad.ox.ac.uk
[16] www.perseus.tufts.edu
[17] The database consists of the following texts: Plautus, Caesar (BG), Catullus, Cicero (orations and letters), Virgil, Horace (Odes), Livy (books 1-10), Ovid (Metamorphoses), Suetonius (Caesars), the Vulgate, and Servius's commentary on Virgil.

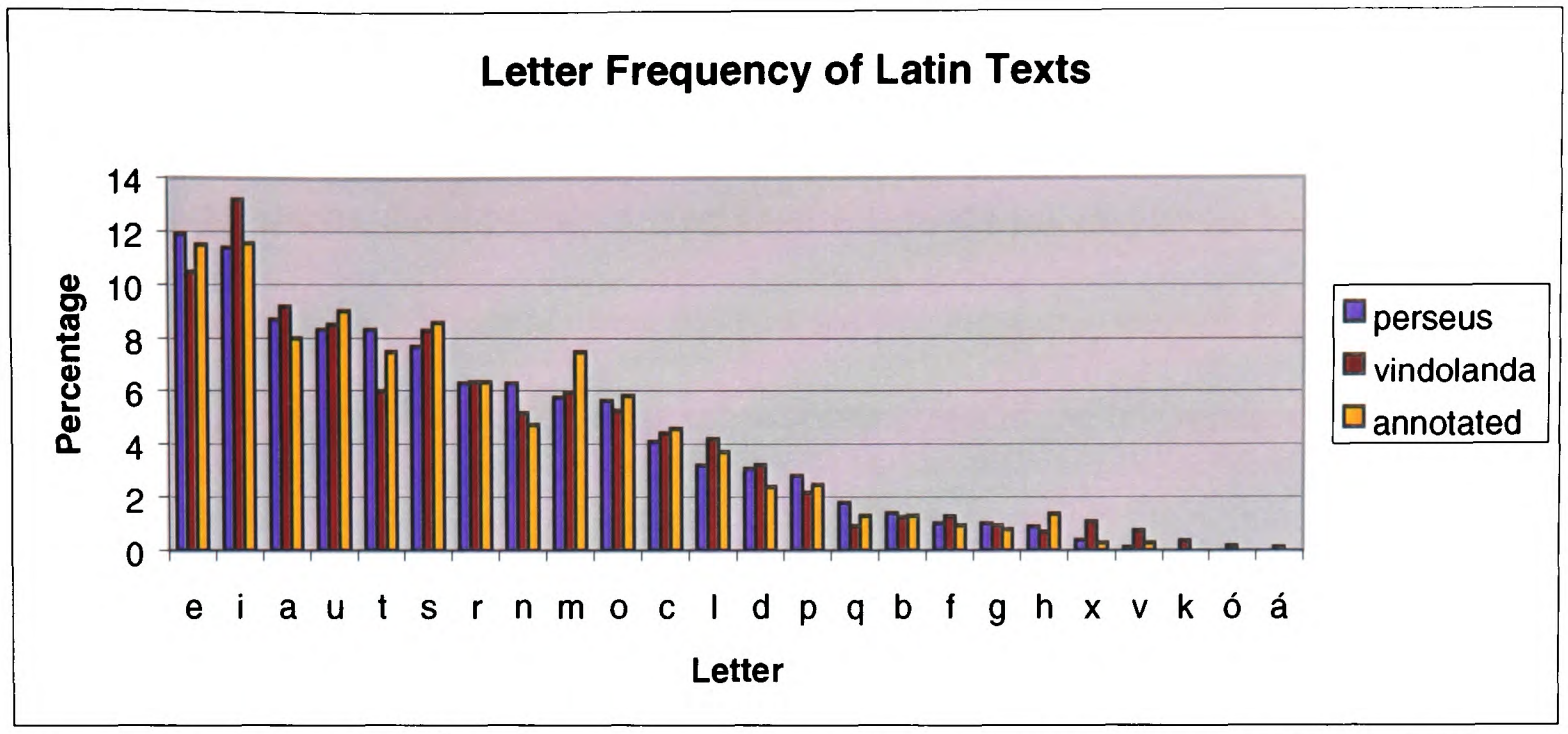



**Figure 3.18:** Comparisons[18] of letter frequency in the Perseus text corpus, the Vindolanda text corpus, and the annotated set of images from Vindolanda[19].

Even though the Perseus corpus has 7.8 million characters, the Vindolanda text corpus (see 4.2) 27,000 characters, and the annotated image corpus only 1,700 characters, the frequencies are remarkably similar. This can be taken as an indication that the annotated characters give a good distribution of the letter forms used in the documents.

## 3.10 Letter Forms

It was suggested in 3.1.2 that the letter forms contained within the Vindolanda stylus texts should be similar to those contained within the ink documents (this being the assumption that the palaeographers work from). The data set constructed during this part of the project contains enough information to evaluate whether this is the case. Appendix B contains all the representations of the characters used within this project, and every instance of the characters in the ink and stylus tablets. Although the author of this thesis is not a palaeographer, there are some

[18] Further statistical analysis needs to be done on this data to assess its validity.

observations that can be made regarding the variations on letter forms between the ink and stylus texts from the data given in this Appendix.

- **A**. On the whole, the character A contained within the ink and stylus texts seems to be very similar. In some cases the two strokes meet at a sharper angle in the stylus A, with the first stroke being more upright than in the ink version.
- **B**. The ink B is fairly fluidly made, whilst those found in the stylus texts are less curved. The second, longer stroke in the stylus B tends to be more straight and does not have the pronounced loops to each end of the stroke as present in the ink B.
- **C**. The letter C is very similar in both texts. In some cases of the stylus text the loop can be less full than in the ink text, with the length of stroke also being shorter in the stylus C.
- **D**. The longer stroke of the stylus D tends to slope top left to bottom right, whereas that of the ink D can be more variable. As with the letter B, there tends to be less curvature in the long stroke of the stylus D than the ink D.
- **E**. As already noted (3.2), the ink and stylus texts contain a different form of the letter E. This can be seen clearly in the examples given. There are no examples of the stylus form being used in the ink texts, or the ink form being used within the stylus texts, in the data set that was constructed. That is not to say that this would never be the case, however.
- **F**. There were no examples of the letter F in the stylus image corpus.
- **G**. As there was only one example of a letter G in the stylus image corpus, it is impossible to make any generalisations regarding the differences between the

[19] The data from which this figure is derived can be seen on the accompanying CDROM in Chapter3/Perseus Comparison.

ink and stylus letter G. However, this one instance is very similar to those found in the ink texts.

- **H**. There were no examples of the letter H in the stylus image corpus.
- **I**. The stylus I shows less use of ligatures and serifs than that found in the ink texts. The models generated from the ink and stylus texts show that the ink I tends to slope from bottom left to upper right, whilst the stylus I slopes from upper left to bottom right. This is pronounced in the models due to the fact that the strokes in each case were almost vertical, and they have been stretched to fit into the canonical 21 by 21 grid. Nevertheless, this does indicate that in general the stylus I tends to slope a little more in the opposite direction to the ink I.
- **L**. The ink L shows a lot of variation in direction, angle, and use of serifs. The small number of examples of stylus L available indicate that this variation is also present in the stylus tablets. However, there seems to be less use of serifs in the stylus tablet forms.
- **M**. The ink M shows considerable variation in its form, width, and angle of stroke meetings. This variation is echoed in the stylus tablet form. The stylus M seems to be less fluid, and the individual strokes more separated than those in the ink M, which can often run together smoothly.
- **N**. The ink N shows considerable variation in its size and angle of stroke meetings. This is reflected in the forms of the stylus tablet N.
- **O**. The ink text O is commonly made in one stroke, which makes a loop. There are a couple of instances where it is made with two strokes in the corpus. The stylus tablet O is made with two combining strokes in every example in the corpus, and never made in one single stroke.

- **P**. The ink tablet P shows some degree of variation. There is only one example of a stylus letter P, and so comparisons between the two are limited. However, the stylus letter P presented does seem to be very similar to the ink letter P.
- **Q**. The longer stroke of the ink letter Q tends to be more upright than that of the stylus Q, which tends to slope more from upper left to lower right. The bow of the ink Q tends to be more of a more elongated width and smaller height than that of the stylus Q.
- **R**. The ink R shows some variation, but the first stroke seems to always be the longer, with the second stroke sitting above the first. In the stylus tablets, there is considerable variation in the letter form. Sometimes both strokes are of the same length, and can meet end to end, rather than overlapping. The first stroke tends to be straighter in the stylus tablets than that of the ink tablets.
- **S**. The ink S shows some variation in form, but in general the form found in the stylus tablets echoes that found in the ink texts.
- **T**. Again, the form found in the ink texts shows some variation, and this is the case with those found in the stylus tablets, but in general the forms found on both texts are very similar.
- **U**. There were no examples from the stylus tablets to compare the forms from the ink texts with.
- **V**. There were no examples from the stylus tablets to compare the forms from the ink texts with.
- **X**. There were no examples from the stylus tablets to compare the forms from the ink texts with.

On the whole, the letter forms used on the stylus text do seem to be similar to those contained within the ink documents. The main difference between the two regards strokes which are looped or curved: it being much more difficult to curve the stylus through the wax than make the same fluid motion with ink on wood. However, the forms of the character remain the same (apart from the letter E, as demonstrated), and so models of the ink characters, and the stylus characters available, should provide an adequate means of reading the unknown characters contained within the stylus texts in the future. This is explored in 4.8.

## 3.11 Conclusion

In this chapter, a review of palaeographic research regarding Old Roman Cursive was presented, and a series of knowledge elicitation exercises carried out, to enable the encapsulation of the types of information experts utilise when discussing and identifying letter forms. This enabled an encoding scheme to be developed so that a corpus of annotated images could be constructed. The resulting data set is the only such amalgamation of palaeographical information regarding Old Roman Cursive handwriting in existence, and as such, could prove to be a unique resource of this period for papyrologists and palaeographers. This corpus has already provided the means to investigate the difference between the letter forms as found on the ink texts and the letter forms found on the stylus tablets. The corpus was utilised as the means to train a system to attempt to read the ink and stylus tablets, as discussed in Chapter 4.

The Knowledge Elicitation techniques used in this chapter were time consuming, and the amalgamation of knowledge depended on the thoroughness of the

knowledge engineer. However, the resulting encoding scheme is comprehensive. Although it refers to Old Roman Cursive in particular, it could be used as the basis of a scheme to annotate any stroke based written text. There were some elements of the encoding scheme that were felt to be not as relevant as others when it came to annotating the Vindolanda texts (for example, stroke width is fairly standardised throughout the corpus) but the encoding scheme provides the means to encapsulate different types of graphical information. Often, the relevance of the information only becomes apparent after the annotation has taken place, and this encoding scheme and annotation tool enables as much information as possible to be captured in the image corpus.

Annotating the corpus was a tedious, time consuming task. Although it was carried out in the most systematic way possible, there may very well be some undetected human error included in the annotations. One way to resolve this problem would be to annotate the files again, and to compare the resulting annotations: any areas of difference would be highlighted, and the confusion resolved. Fortunately for the author, there was not enough time available during the project to enable this to happen.

# CHAPTER 4

## Image to Interpretation

## Using a Stochastic MDL Architecture to Read the Vindolanda Texts

"Now – here we go!" He reached up and pulled a switch on the panel. Immediately, the room was filled with a loud humming noise, as a crackling of electric sparks ... and sheets of quarto paper began sliding out from a slot to the right of the machine ... They grabbed the sheets and began to read. The first one they picked up started as follows "Aifkjmbsaoegweztpplnvoqudskigt&, -fuhpekanvbertyuio, lkjhgfdsazxcvbnm,peru ,trehdjkg munb, wmsky..." They all looked at the others. The style was roughly similar in all of them. Mr Bohlen began to shout. The younger man tried to calm him down.

"It's all right, sir. Really it is. It only needs a little adjustment. We've got a connection wrong somewhere, that's all. You must remember, Mr Bohlen, there's over a million feet of wiring in this room. You can't expect everything to be right first time."

"It'll never work," Mr Bohlen said.

Roald Dahl, The Great Automatic Grammatizator. (Dahl 1997)

To implement a system based on the way the papyrologists approach and read ancient texts, an appropriate architecture was firstly identified. The GRAVA system, developed by Dr Paul Robertson, solves interpretation problems by using Minimum Description Length as a unifying means of comparing and passing information between different semantic levels. Successfully used to build a system that could effectively "read" a hand written phrase, the GRAVA system was adopted as the means by which to implement a system to read the Vindolanda texts, as it provided the architecture to mobilise different types of knowledge to solve an interpretation problem. Because of the need to eventually segue with the work done on image processing on these documents, the original GRAVA system was adapted to incorporate more sophisticated character and detection agents. Information regarding the letter forms contained in the Vindolanda texts was ascertained from

the annotated images (Chapter 3), and statistics regarding the language used were generated from the corpus of Vindolanda texts read to date. The system was implemented, and tested firstly on hand annotated ink tablet data, before testing on hand annotated stylus tablet data. As a final test, automatically generated annotations from the image processing algorithms were used as the test set. This investigated whether an implementation of a system in this manner could provide the means to dovetail with the research done on propagating possible stroke identifications on the Vindolanda stylus tablets through the use of Phase Congruency (see 1.2).

## 4.1 The GRAVA System

In his thesis (2001) Paul Robertson utilises an agent based system to read a hand written phrase, implementing a multi-level hierarchical model. This is akin to the Interaction Activation and Competition Model for Word Perception as proposed by Rumelhart and McClelland (McClelland and Rumelhart 1986, see also 2.8.2) where a phrase is read by identifying features, characters, and, finally, words. However, Robertson's GRAVA (Grounded Reflective Adaptive Vision Architecture) system does not use Parallel Distributed Processing architecture. PDP remains difficult to implement because the approach, which aims to develop neural network architectures that implement useful processes such as associative memory, throws out traditional models of computing completely, requiring a new computational model based on neuronal systems. An additional problem with PDP is that the behaviour of such a network is highly nonlinearly related to the parameter values; a small change in those values may lead to very different behaviour (although this may also be the case with MDL). In order to overcome this intrinsic difficulty it is

generally considered that PDP requires a large training set of data to "fine tune" the parameters used within the system; there is not enough data available to us regarding the letter forms of Vindolanda to be able to do this. Instead Robertson implements an architecture where the atomic elements are implemented as agents (in YOLAMBDA, a dialect of LISP), using familiar programming practices, which retains a more conventional programming model than the PDP approach.

The primary purpose of an agent "is to fit a model to its input and produce a description element that captures the model and any parameterization of the model" (Robertson, 2001 p.59). The GRAVA system[1], developed by Robertson, manipulates agents, and builds programs from them. Agent co-operation can span semantic levels, allowing hierarchical stacking in the same way that is described in PDP. This enables the building of systems that exhibit semantic interaction, a well understood hierarchical concept that allows the behaviour and performance of systems to be closely monitored and understood (using techniques such as convergence analysis). Such an architecture is well suited to interpretation problems, which can be solved by fitting an input to a series of models and generating descriptions of the likelihood of these matching. Many interpretation problems have more than one possible solution, and by using such a system many solutions can be propagated; the best solution being the ultimate target. Of most relevance to *this* thesis, Robertson shows how his architecture is an effective way of rendering hierarchical systems by demonstrating how his software can "read" a hand-written phrase. The text in this case was a nursery rhyme. The example given

[1] For a full specification of the GRAVA system architecture, see Robertson (2001), Appendix B.

provides the basis on which to develop a system to read the Vindolanda ink, and eventually, stylus tablets.

### 4.1.1 System Architecture

Robertson's example comprises of three levels of agents. The first level agents detect low level features of the written text. The second level agent builds character models from the database of character forms, and compares these models with the evidence presented by the first level feature detectors to output a possible character identification. The highest level agent finds evidence of words in the input by looking at data generated from the character finding agent and comparing this to a given corpus of words. The corpus used is the complete nursery rhyme, from which statistical information is collected and from which models are constructed regarding characters and words.

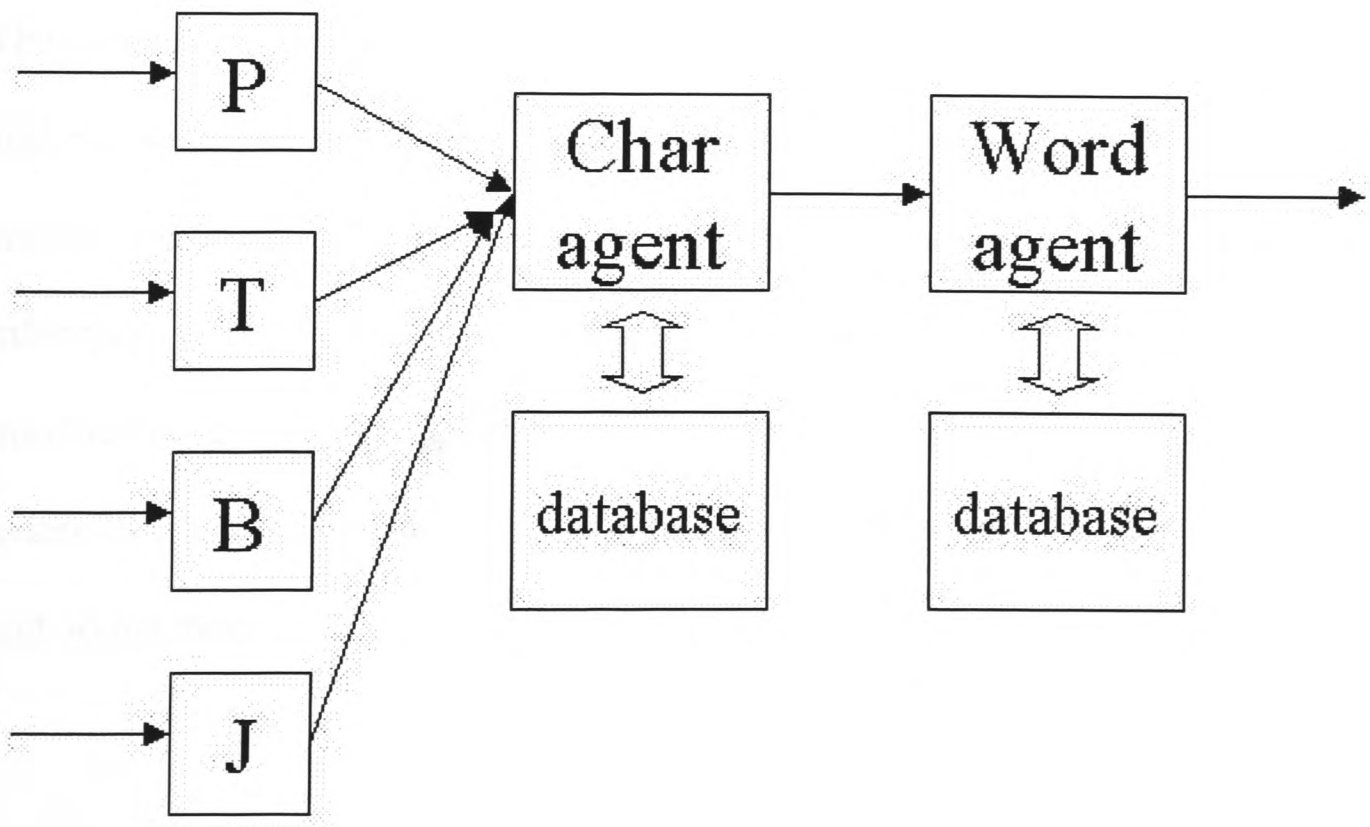



**Figure 4.1:** Hand Written Phrase Recogniser Program. From Robertson (2001, p.65).

There were four separate agents which combined to make the lowest level feature detection section of the model. Each agent reported on features discovered within a character position:

1. Top stroke endpoints (T, above). This agent reported on the number of stroke endpoints at the top of the character. For example, the letter 'N' has one stroke endpoint on the top and 'W' has two.
2. Bottom stroke endpoints (B). This agent reports on the number of stroke endpoints at the bottom of the character. For example the letter 'A' has two endpoints at the bottom of the character and letter 'I' has one.
3. Stroke Junctions (J). This agent reports on the number of line junctions formed from three or more lines. For example the letter 'A' has two such junctions. The letter 'N' has none.
4. Character present (P). This agent detects whether the character position contains anything at all. Everything but the space character contains something. (Robertson, 2001, p.64)

The character boxes are segmented into "top" and "bottom" so that the top stroke and bottom stroke features can be determined from a simple endpoint filter. Such a simple encoding system allows features to be flagged in order to pass the information up to the next character identification level: however, they are insufficient to identify unambiguously a character themselves. For example, the letters 'S', 'C', 'I', 'L', and 'N' all have one endpoint at the top, one at the bottom, and no junctions.

#### 4.1.1.1 Description Length

The character level and word level both contain single agents to calculate the probability that the data in the corpus and the data presented matches. GRAVA utilises Description Length (DL) as a means of comparison (See Robertson (2001, p.50)). Description Length is the theoretical code length required to transmit an object in the form of a message over a noiseless channel, and is a standard method for the comparison of probability. It provides a fair basis for cost computation as it captures the notion of likelihood directly: $DL= -\log_2(P)$. The Minimum Description Length (MDL) generated when matching an input to a set of models yields the "best fit" of that data to an individual model, and so presents the most likely match. The global Description Length generated by matching a series of inputs to models can be simply calculated by adding the DL of each unit. This can give a measure of the overall probability of the sequence, allowing easy comparison with other possible interpretations (the interpretation with the lowest global Minimum Description Length being the most likely). MDL is used in GRAVA as a unifying method of comparing data from different levels of abstraction: thus providing a solution to the problem of how to compare the probabilities of both image and linguistic data matching the input.

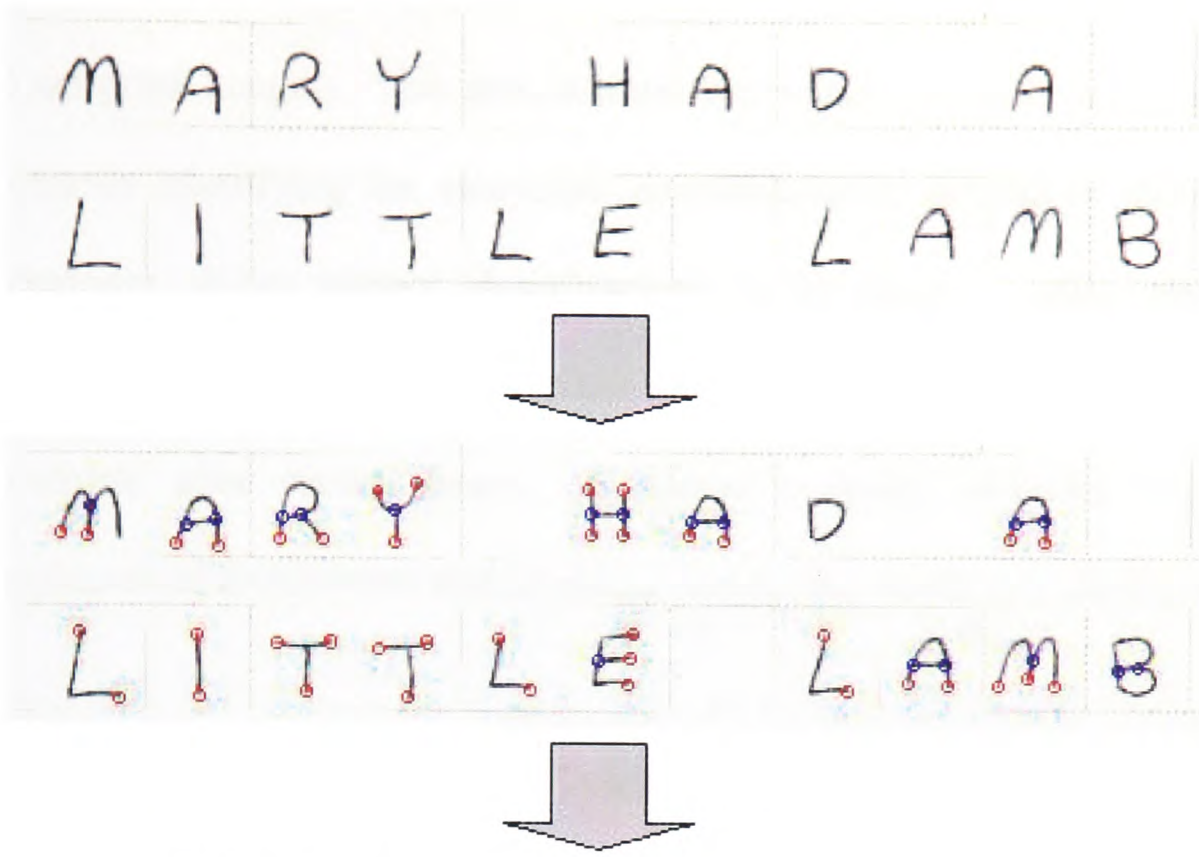
#### 4.1.1.2 Monte Carlo Method

The system also employs a Monte Carlo method, a means of providing approximate solutions by performing statistical sampling, to randomly choose which data is passed upwards between levels. This is a "weighted random" selection process which picks likely data much more frequently than less likely examples (see Robertson (2001, p.48)). If only the data with the lowest Description Length was

passed up between levels, the correct answer may never be found: the data with the locally Minimum Description Length may not be the correct selection on the global level. The stochastic nature of this method of sampling ensures the generation of different results, and also means that the system rapidly generates possible solutions without relying on exhaustive search (cutting search time). The system generates possible solutions on each iteration; the more iterations, the better the chance that a match to the solution is generated. Convergence on an ideal solution is then asymptotic: the system finds approximate solutions, and the more iterations that occur, the better the approximation that is reached. In practice, the system tends to find the exact solution in a short number of iterations, meaning that performance times are acceptable. This is shown in 4.8.1.1, below.

### 4.1.2 The System Demonstrated

Robertson demonstrates how the low level agents start out with a description of features based on the tops, bottoms, and junctions of characters. The system compares these with the models computed earlier from the corpus (on a character level), computing for each symbol a Description Length. The character that is passed upwards to the next level is determined at random by Monte Carlo sampling. The system then compares the resulting "words" with those in the corpus, generating Description Lengths for each. A global Description Length for the phrase is calculated by summing the Description Lengths of each of the symbols and words. In subsequent iterations different possibilities are generated, and the system searches for the best fit by looking for the configuration which gives the lowest global Description Length. This global Minimum Description Length corresponds with a correct "reading" of the text.




```
=> (runCycles #t)

Description Length=306.979919
Description=(t=0 b=0 j=0 p=0 t=0 b=2 j=0 p=1 t=0 b=2 j=2 p=1
        t=0 b=2 j=2 p=1 t=2 b=1 j=1 p=1 t=0 b=0 j=0 p=0
        t=2 b=2 j=2 p=1 t=0 b=2 j=2 p=1 t=0 b=0 j=0 p=1
        t=0 b=0 j=0 p=0 t=0 b=2 j=2 p=1 t=0 b=0 j=0 p=0
        t=1 b=1 j=0 p=1 t=1 b=1 j=0 p=1 t=2 b=1 j=1 p=1
        t=2 b=1 j=1 p=1 t=1 b=1 j=0 p=1 t=2 b=1 j=1 p=1
        t=0 b=0 j=0 p=0 t=1 b=1 j=0 p=1 t=0 b=2 j=2 p=1
        t=0 b=2 j=0 p=1 t=0 b=0 j=2 p=1)

Description Length=116.690002
Description=(M A A E HAD A  IL T E ST  I R M B)

Description Length=65.699996
Description=(M R A Y HAD  R LITTLE LAMB)

Description Length=61.749996
Description=(M R A E HAD A LITTLE LAMB)

Description Length=41.649997
Description=(MARY HAD A LITTLE LAMB)
```

**Figure 4.2:** Input data, input data with junctions and endpoints identified, resolved into an encoding scheme which the character and word agents compare to the models generated from the corpus data. The result with the shortest Description Length is the correct solution. From Robertson (2001, p.24).

In this way, Robertson shows that MDL formulation leads to the most probable interpretation, and also that global MDL mobilises knowledge to address ambiguities. (In the second iteration of this example fully half of the characters

were "guessed" wrongly by the system, but by the fifth iteration all ambiguities were resolved, because the correct choices were what led to the global Minimum Description Length). This demonstrates that whilst the input data is inadequate for correctly identifying the characters unambiguously, the use of global MDL as a constraint allows correct identifications to be made. MDL also provides a "reasonable" description in a relatively rapid time: exhaustive searches can fail to complete after several hours. Robertson's system provides a robust, easily implemented architecture that produces convincing results in a short time frame.

### 4.1.3 Application of the System

Robertson uses this MARY HAD A LITTLE LAMB example to illustrate the power of his new architecture on a simple problem, before developing special purpose agents for aerial image understanding (the development of a self adaptive architecture for image understanding being the focus of his thesis). However, this example provided a useful starting point for a system to read the stylus tablets as it provided the architecture to construct, develop, and adapt a system to read ancient texts in the manner described in the model of 2.8.3. The use of such features as end points, and junctions also provides a useful introduction to the problem, as this information has already been captured in the database (see 3.5), and so a first run of the system on the data was easy to implement. Firstly, however, it was necessary to collect various statistics regarding the language used in the Vindolanda ink and stylus texts to provide information for the character and word agents.

## 4.2 Gathering Corpus Data

The Vindolanda ink tablet corpus is the only contemporaneous resource to the Vindolanda stylus tablets. The word list generated from this corpus may be different from that produced from any other available source material regarding Latin, due to the temporally and geographically distinct nature of the corpus[2]. Although the language used in the texts is not strictly Vulgar Latin[3] "they contain... a stock of terms... of the type which are rarely, if at all, attested in literary genres" (Adams 1995, p.120). The texts contain lexical and syntactic "errors"[4], and utilise new words, new meanings of known words, the first attestations of abnormal forms of words, Celtic loan words, and anticipations of Romance language. Most other large sources of Latin words and grammar deal with Classical Latin: a version of the language that was written by a few authors, and spoken by almost no-one; Vulgar Latin was spoken by millions, and the Vindolanda texts are one of the only large sources available on which to base any conclusions regarding the language. For these reasons, the entire known corpus of ink texts from Vindolanda (as of December 2001) was the only source used to generate lexicostatistics to aid in the reading of the Stylus texts. In the future, it will be possible to generate other sets of statistics from different corpora, which could be used to analyse how similar the

---

[2] "We can even talk of regional "dialects" of Latin. We can suppose this partly on the basis of generalities that have been discovered to be empirically true for the study of all languages; when a language is used of a wide and disparate geographical area, influenced by widely varying external factors of an ethnic and socio-cultural type, geographical variants can arise that are noticeably different from each other despite the fact that they all form part of a single linguistic system ... It seems likely that in imperial times a slight amount of geographical variation did slowly arise in Latin, affecting pronunciation in particular, but perhaps also a few morphological details (ignoring, of course, the wide differences we find in personal and place names, due to the different ethnic origins of the populations of different areas.) This kind of divergence posed no threat to the fundamental unity of the language, which is hardly surprising in view of the centralising power of the empire itself and the strength of its traditions" (Herman 2000, p. 116-9).

[3] "By *Vulgar Latin* is meant primarily that form of the language which was used by the illiterate majority of the Latin-speaking population" (Coleman 1993, p2). For a discussion of the relation of the language of the Vindolanda texts to Vulgar Latin see Adams 1995, p.131.

[4] When compared with literary Latin.

Vindolanda corpus is to other texts available (as these documents are correspondence from the Roman Army, there must be standardisation in the language used. The statistical comparisons made between letter frequency from the Vindolanda text and the Perseus corpus (see 3.9 and below) would seem to indicate that this was the case.) This additional data could also being used to provide alternative information sources for the system described in this thesis. However, the Vindolanda corpus was the only source used here as it was readily available, is contemporaneous to the texts found on the stylus tablets, and is of a large enough size to provide statistics with which to test the implementation of the system.

The Vindolanda ink corpus used comprised of the 230 Latin texts published in Bowman and Thomas (1994), plus 56 new texts that had been read in preparation for the publication of the next volume of the Vindolanda texts (see 2.5.5.3). There were 27364 characters (excluding space characters) in total, comprising 6532 words, or word fragments. In the corpus there were 2433 unique word tokens (1801 words appeared only once, the rest were repeated). This should provide an adequate corpus on which to base any conclusions about the language used at Vindolanda[5].

[5] The representativeness of a Corpus has been much discussed in the field of Corpus Linguistics (Kenny 1982; Biber 1983; OOstdijk 1988; Biber 1990), the consensus being that "small is beautiful" and that "there is every reason to make maximal use of these corpora [of 2000 word length] for analysis of linguistic variation until larger corpora become readily available" (Biber 1990, 269) The major corpus of American English, The Standard Corpus of Present Day Edited American English (known as the Brown Corpus) uses 2000 word samples to represent its various genres. Zipf's "law" (Zipf 1935, Reprinted 1965), which predicts an "ideal" set of frequencies (and hence probabilities) for lexical items given a particular vocabulary size, maintains that a dictionary size of 1000 words will give a percentage cover of 56.220% of the language, although this is based on an analysis of English language texts. In the case of this project, the corpus used is 100% of the known corpus of Latin for this period.

This corpus was analysed using WordSmith (see 2.4.2) and TACT[6], two common corpus linguistics computer programs, to generate the following data:

- Frequencies of letters
- Letter combinations (bi-graph analysis)
- Word list
- Frequencies of words[7].

The representativeness of the data can be demonstrated by comparing the frequency of letters in the Vindolanda corpus to that of the Perseus Corpus (see 3.9). Although the Perseus corpus is significantly larger, with 7.8 million characters to the Vindolanda 27,000, and deals with a more standardised type of Latin, the frequencies of letters are markedly similar[8]:

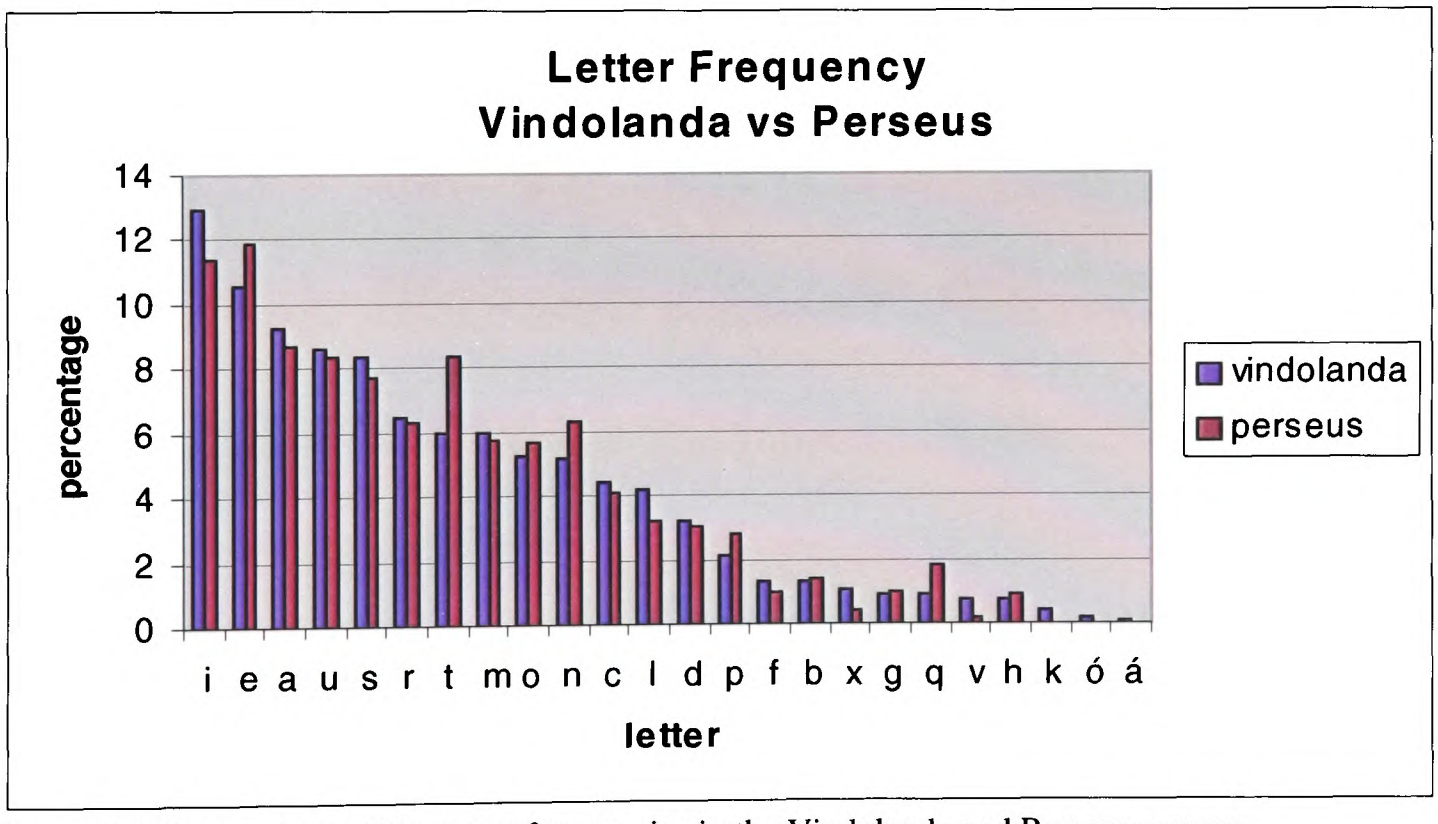



**Figure 4.3:** Comparison of the letter frequencies in the Vindolanda and Perseus corpora.

[6] "Text Analysis Computing Tools": http://www.chass.utoronto.ca/cch/tact.html
[7] Spreadsheets containing these data sets can be found on the accompanying CD-ROM in Chapter 4/lexicostatistics/.
[8] Much more statistical analysis needs to be done on this data to illustrate its validity.

## 4.3 Preliminary Experiments

First trials were done, utilising the annotated ink tablet images, where endpoints of the characters were used to provide the identifying features which were passed upwards to the character level (in exactly the same way as the example above, 4.1.1) but only utilising endpoints as the source of character data instead of endpoints and junctions).

Firstly, model data was generated using the majority of the annotated images of the ink stylus texts (the test set, a section of tablet 255, being excluded from the training set to allow a fair comparison). Secondly, the annotated test section of 255 was read into the system, and the system attempted to resolve the features into words within 15 run cycles. Results were encouraging[9].


**Figure 4.4:** A section of ink tablet 255 used as the test set. The test section reads "ussibuss puerorum meorum[10]".

[9] These results were generated with the help of Dr Paul Robertson, from the MIT Artificial Intelligence Laboratory.

[10] There was some confusion when annotating the images at first as to whether the text read "ussibuss" or "ussibus". It was firstly annotated as "ussibuss", as shown here, then changed to "ussibus" when cross referenced with other data. The first (incorrect) set of annotations were retained to provide a means of comparing the system, as described in 4.8.2. The first trial, however, was only applied to the "ussibuss" spelling.

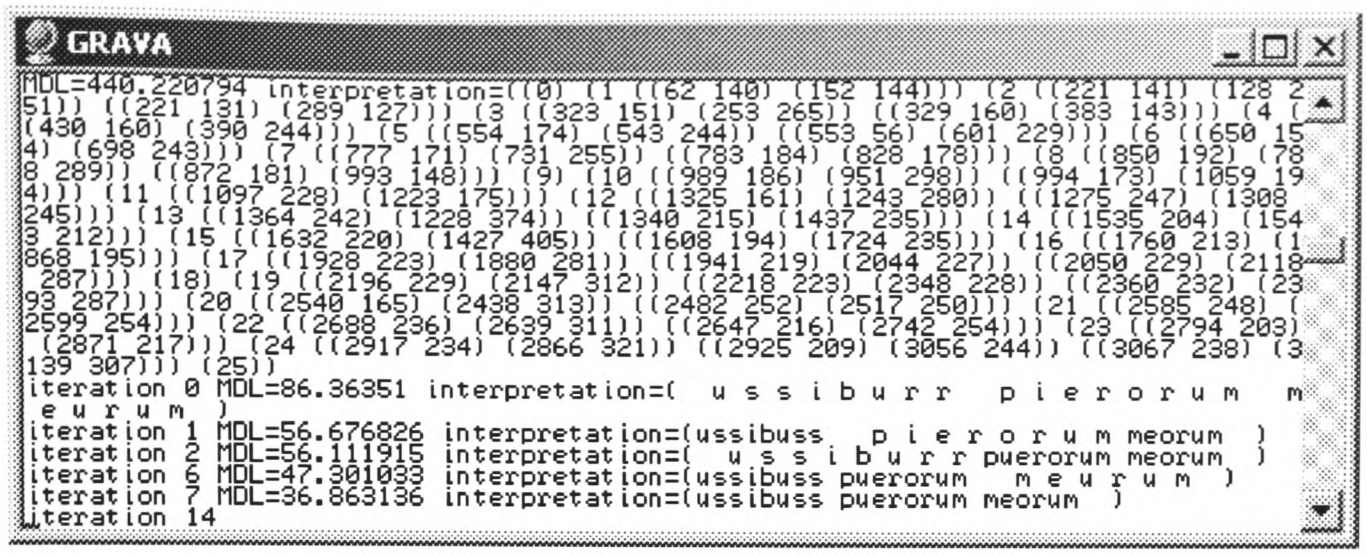



**Figure 4.5:** The screen output of the GRAVA system, showing the Description Length of different iterations. The text was correctly interpreted in the seventh iteration. The first section of output lists, for each stroke, a set of co-ordinates. The subsequent section outputs the summed Description Length for each resulting string (which is a possible solution to the problem). A space between characters indicates that character is being read individually, characters together are identified as words. The system only prints output on iterations where the MDL is shorter than that of the previous output. The number of run cycles can be chosen by the user: in this case it was 15 iterations. The output with the shortest MDL is the correct interpretation.

However, there is a significant limitation to this approach in using the data to "read" unknown characters. Ultimately, this system aims to dovetail with the work done using Phase Congruency based local energy measurement to identify possible candidate handwriting strokes within the stylus tablets (see 1.2, also Schenk 2001; Molton, Pan et al. Forthcoming (2003); Pan, Brady et al. Forthcoming (2003); Robertson, Terras et al. Forthcoming (2003); Brady, Pan et al. Forthcoming, (2003)). To investigate whether the system could cope with automatically generated representations of characters (or "computer" annotated images as opposed to "hand" annotated images), the same section of ink tablet 255 as above was analysed[11], utilising the techniques developed for the analysis of the stylus tablets.

### 4.3.1 Automatically Annotating Stroke Data

The analysis had three stages, culminating in the generation of a representation of a character as a collection of strokes. Firstly, feature detection was carried out using

phase congruency based local energy measurement. Secondly, the image was segmented. Finally, a stroke description was generated, which described each stroke by its location, shape outline, central line, end points, and junction points. This description was output to an SGML file, using the same markup as the GRAVA annotator program (see 3.6), allowing it to be read in and processed by the GRAVA architecture in the same way as a hand annotated image. Unfortunately, the resulting data is significantly different, in some ways, than that of a hand annotated image, making comparisons regarding end points (and/or junctions) alone worthless.

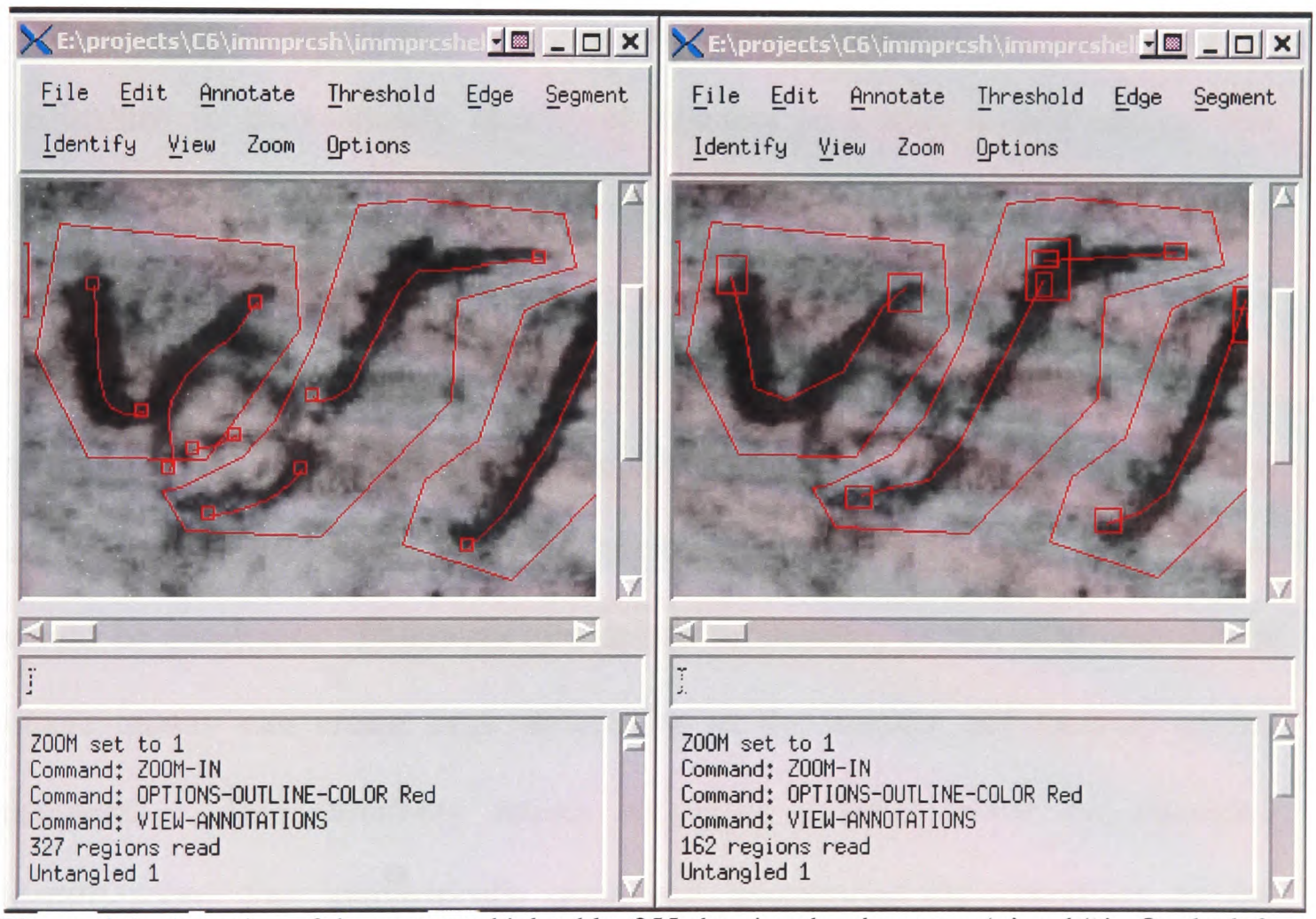



**Figure 4.6:** A section of the annotated ink tablet 255 showing the characters 'u' and 's'. On the left is the image annotated by the automated process, on the right the hand annotated images.

It can be seen that there are numerous occurrences in these two characters alone where the strokes and endpoints are identified completely differently. With the hand annotated U, the character is shown as one stroke with two endpoints, both in

[11] This analysis was carried out by Dr Xiabo Pan, formerly of the Robots Group, Engineering

the upper part of the character. In the automatically recognised character U, the character is split into two large strokes, with one smaller stroke to the right also being included in the character. This gives a total of six endpoints. The S is also significantly different. The hand annotated S is shown as comprising of two strokes, one long, one short, with a junction between them in the upper left of the character. The automatically recognised S also has two strokes, but the longer is at the top, and the shorter at the bottom. There is no junction, and there is a significant break between the two. Although both have four endpoints, two of these are in very different places.

This difference is hardly surprising, due to the noisy nature of the tablets, and the difficulties in automatically identifying features accurately within images (see Gonzalez and Woods (1993) for an introduction). It is often difficult for a human to segment overlapping areas in an image, never mind a computer program (witness the mistake made when annotating "ussibus" as "ussibuss" due to the unclear nature of the final characters, see 3.7.2, 4.3 and 4.8.2). However, this test indicates that the use of endpoints as the sole identifiers of characters in this application of the system would be unsuitable. Endpoints turn out to be unstable, in that small changes in image quality can create large differences in the number and location of the endpoints. This instability makes endpoints a poor choice for character identification. The automatically generated annotations are compared to the character models built from the hand annotated characters, and so they have to hold a certain level of similarity for the system to function. The automatically annotated data generated from the test set was tried in the system, and it did not succeed in producing the correct reading. A more stable, graphical means of encoding and

---

Science Department, Oxford University, now at Mirada Solutions Ltd.

modelling the letter forms was needed, as a means of employing the available data in the feature level of the system. This would have to compare the most important data regarding the characters: their constituent strokes.

## 4.4 System Development and Architecture

As before, the character database formed by annotating images of the ink tablets, as discussed in 3.5, is the main source of information regarding the character forms. The models derived from this data are used to compare the unknown characters in the test set (which may be annotated by hand, or generated automatically, as described above (4.3.1)). The difference between the original system (4.1) and the new system lies in the way that character models are developed, and how the test data is compared to this set of character models. Whereas the original system relied on endpoints, this final version relies on data regarding the strokes themselves. The end point agents in the feature level of the original system were replaced by a stroke detection agent. This results in models of characters that are less sensitive to the feature detection process (i.e. the generation of end points, which is problematic when dealing with noisy images such as those of the stylus tablets). It also means that the feature level agents depend on information which is much more easily propagated from automatic feature detection, allowing for easier amalgamation with the stroke detection system, as discussed in 4.8.5. Most importantly, stroke information is much more stable than endpoint data. Small changes in image quality cause only small changes in the stroke features detected. A schematic of the final system that was developed is shown below (figure 4.7), incorporating all elements of the resulting process.

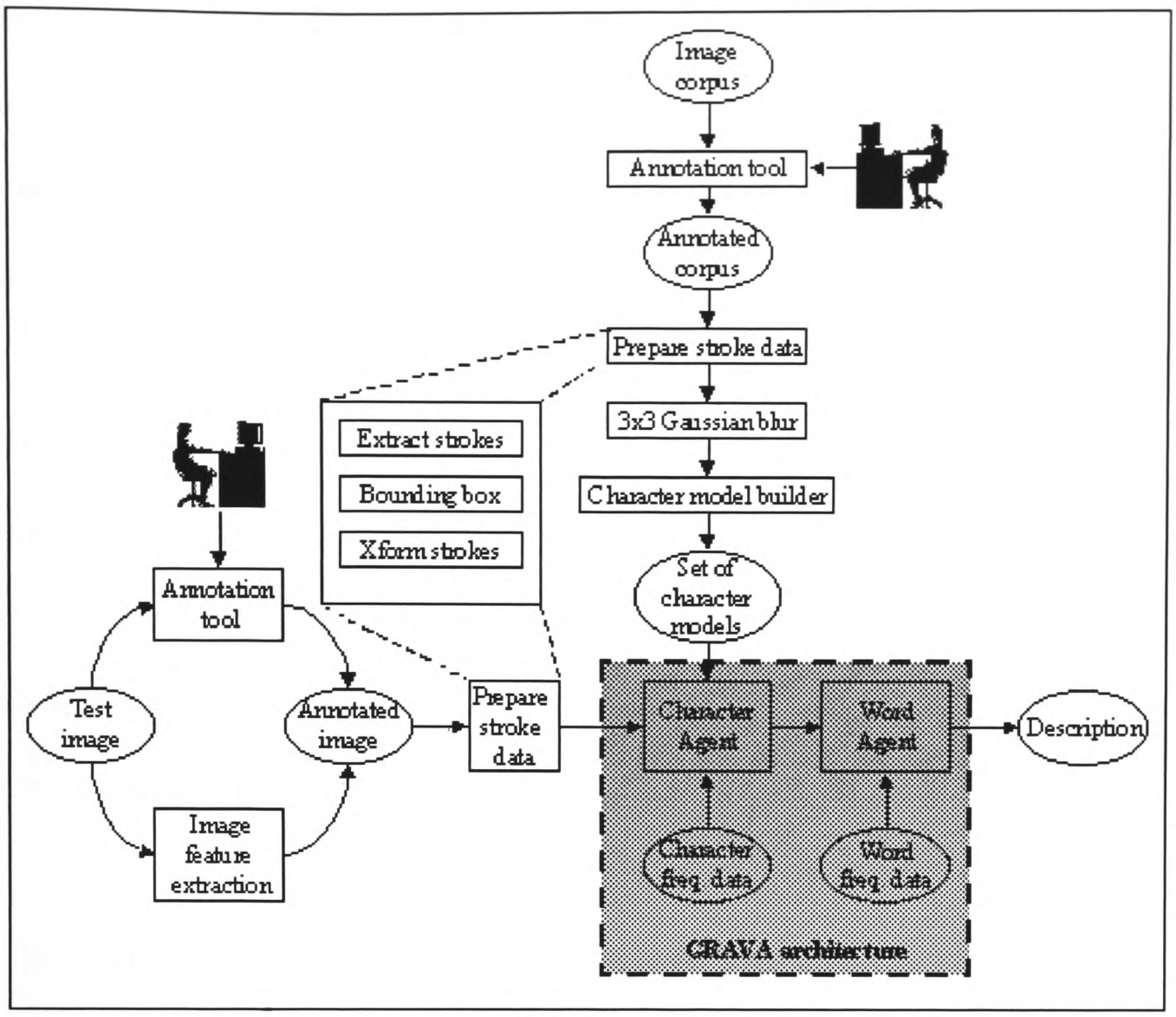



**Figure 4.7:** Schematic of final system. Robertson's GRAVA architecture is highlighted to indicate the processes which are carried out as part of the final run of the system. The "Annotated image" can be generated manually by an "Annotation tool", or automatically by "Image feature extraction".

Before being used by the character agent, both the test set of data and the annotated corpus must be prepared in order to analyse their stroke data. This is done by extracting the strokes, drawing a bounding box around each of the characters to preserve groupings of strokes, and transforming these strokes onto a canonical sized grid to allow easy comparison.

Character models are built from the annotated set of images from the corpus by applying a Gaussian blur operator, and summing every instance of each individual character to build up a character model. The character agent then compares the unknown characters from the test data with those in the character models, also utilising frequency information about each character to generate a Description

Length for each model. One of these likely characters is then selected using the Monte Carlo sampling method and passed up to the Word agent. This ensures that, over successive iterations, a fair, representative amount of each candidate character is selected and passed onto the next level.

The Word level takes in the data from the Character agent, combines them to form words, and compares them with the words from the corpus – the word "models" in this case being the words found in the corpus. A Description Length for this comparison is noted. A selection is made from the possible words utilising Monte Carlo sampling methods, and the final word output is generated.

The system then adds the Description Lengths for all the words in the phrase (or string of words) together, giving a global Description Length for that combination of characters and words.

The system repeats this process as often as the user dictates, and keeps track of the lowest global Description Length generated by each successive run. The Minimum Description Length produced corresponds with the most likely answer: or the best fit answer available.

The preparation of the character models, and the way that both the character and word agents work, is discussed in detail below, before demonstrating results from this system.

## 4.5 The Construction of Character Models

A character model is defined as a probability field that indicates the likely placing of one or more strokes of a two-dimensional character, producing a general representation of a character type. Unknown characters can then be compared to a series of these models, and the probability that they are an instance of each one calculated, the highest probability indicating a match. Whilst the first implementation of the system relied on an end point agent, this was replaced by a stroke detection agent that builds up character models based on the actual strokes of the character.

On a conceptual level, the (stroke-based) character model is constructed by taking an image of an individual character, finding its bounding box (identifying the rightmost x co-ordinate, and the leftmost x co-ordinate, and the highest and lowest y co-ordinates), and transforming this into a standardised (21 by 21 pixel) grid. The stroke data is convolved with a Gaussian Blur operator to reduce over-fitting. Each standardised representation is accumulated onto a generalised matrix for each character type: resulting in a generalised representation of each type of character. These are subsequently used as the models to which unknown characters are compared.

### 4.5.1 Finding the Bounding Box

The minimum and maximum x and y points of the character are noted, drawing a "bounding box" around the letter. It does not matter how large or small this is, or how wide or narrow, as the representation is standardised by transforming this bounding box into a regimented size.

## 4.5.2 Calculating the Transform

The matrix presented by the bounding box is transformed into a standardised size to allow easy comparison of representations. This is achieved by simply translating the area within the bounding box to a canonical 21 by 21 pixel region at the origin.

$X_{min}$, $Y_{min}$ and $X_{max}$, $Y_{max}$ of the bounding box are noted. $M_{width}$ and $M_{height}$ are the model width and height respectively that the box is to be scaled to (in this case 21 by 21 pixels). A scaling matrix is then computed, giving the ratio to which the height and width will be scaled. $S_1 = M_{width}/ (X_{max}-X_{min})$. $S_2 = M_{height}/(Y_{max}-Y_{min})$. A translation vector is computed so that all characters are based at the origin. The scaling matrix is then applied to the translated stroke pixels, $X_t$ and $Y_t$ being the new scaled co-ordinates:

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \left( \begin{pmatrix} X \\ Y \end{pmatrix} - \begin{pmatrix} X_{min} \\ Y_{min} \end{pmatrix} \right)$$

The choice of a 21 by 21 pixel grid was arrived at through a process of experimentation. If the grid is too large, then the data is too sparse and it is difficult to make any generalisations about the letter forms. The bigger the grid, the more examples are needed to make the generalised model applicable. The smaller the grid, the closer together the data, giving a trade-off regarding size, accuracy, and usability. At first, an 11 by 11 array was used, but this was gradually increased, which improved the accuracy of this part of the system. The 21 by 21 sized grid seemed to give the best results for the training sets that were used.

### 4.5.3 Applying Gaussian Blur

The stroke data was convolved with a three by three Gaussian Blur operator. This creates intermediary points of data that reduces over-fitting of the model, to increase generalisation from the examples given.

### 4.5.4 Calculating the Final Model

The first character is drawn onto a "blank" grid. A second instance of the character (if there is one available) is drawn over this, the values of this being added to the first instance. Additional instances of the character are laid over the grid, and the values summed as they go. This results in a composite model of all available character instances from the corpus, showing the path the strokes are most likely to make.

An example of how these steps combine to generate a character model is given below, where a small corpus which contains three 'S' characters is used to generate a character model of an S.

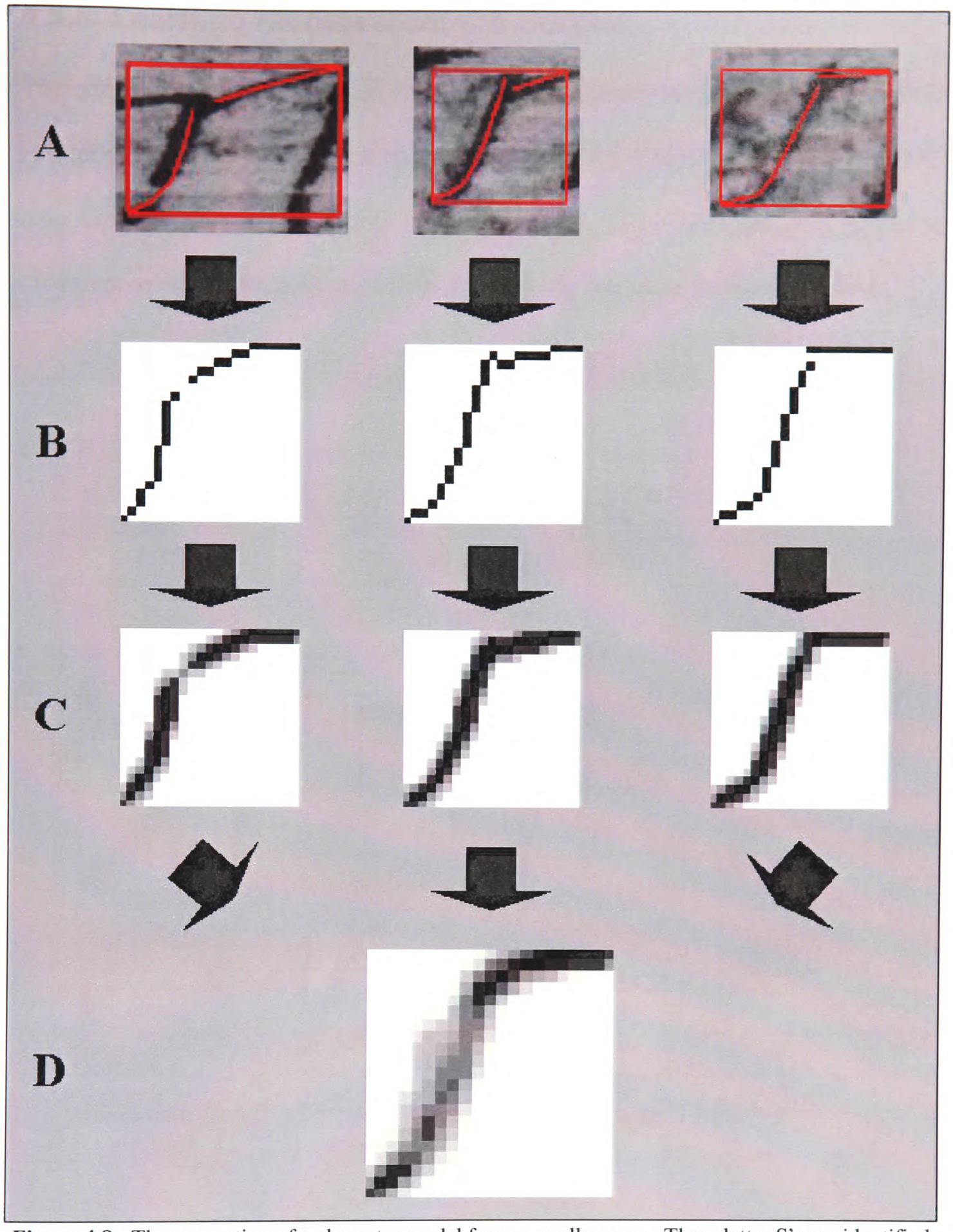



**Figure 4.8:** The generation of a character model from a small corpus. Three letter S's are identified, and bounding boxes drawn around them (A). The stroke data is then transformed into a 21 by 21 pixel grid (B). A Gaussian Blur is applied (C). The composite images generate a character model (D). The darker the area, the higher the probability of the stroke passing through that pixel. In this way, the probabilities of the stroke data occurring are preserved implicitly in the character models.

### 4.5.5 Learning Models from the Corpus

The corpus of annotated images (see 3.5) was randomly divided into a test set and two training sets: that of the ink and the stylus tablets. Keeping the test and training data separate allows fair results to be generated. Two sets (ink and stylus) of character models were generated from the training set; these are shown below.

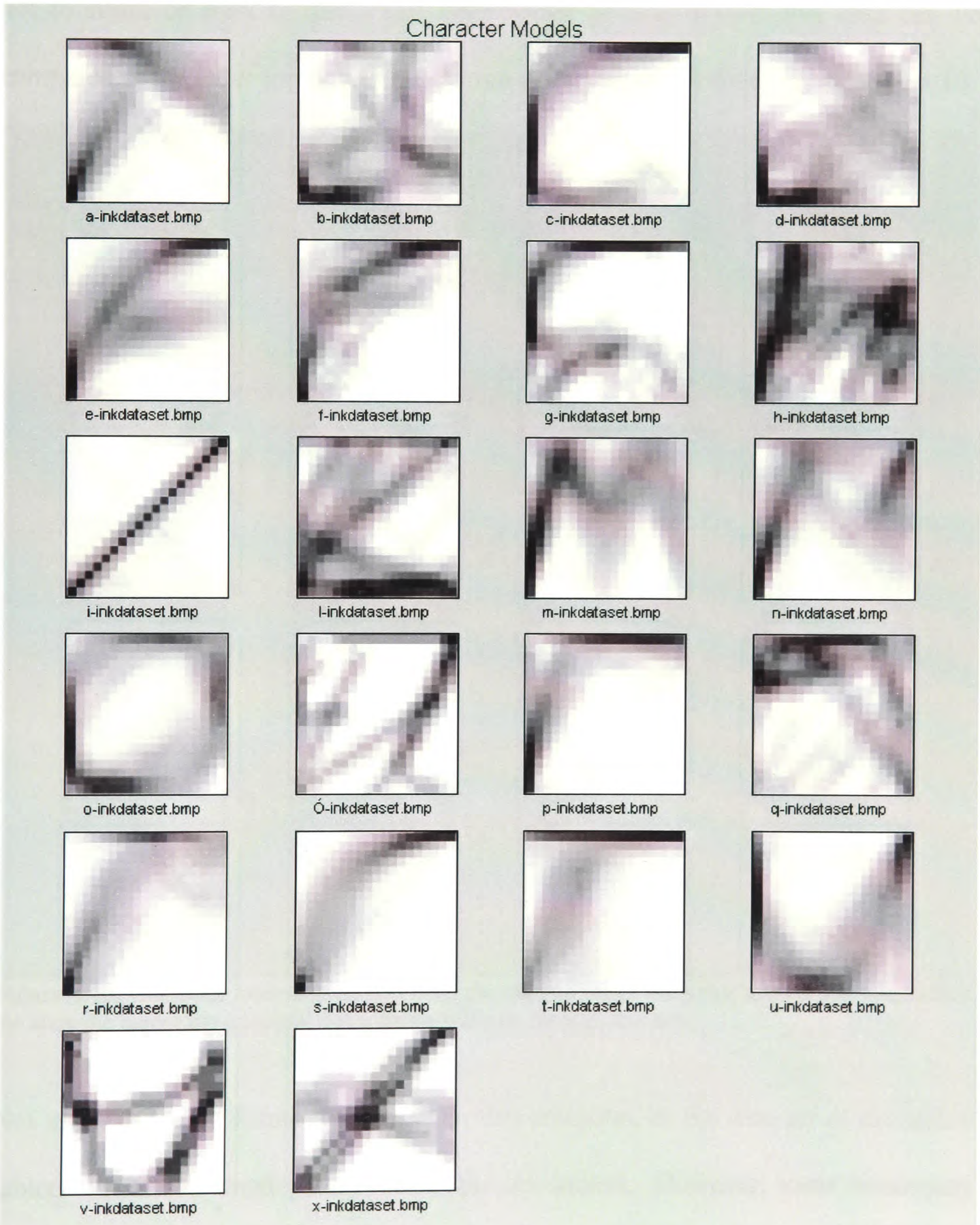



**Figure 4.9:** Character models generated from the training set of the ink text corpus. The darker the area, the higher the certainty that a stroke will pass through that box.

The character models generated from the ink data show that some of the letter forms, such as A, C, I, M, N, and T are fairly standardised throughout the test data. Other characters are more problematic. H, L, and V have a more confused appearance, indicating greater variability in the appearances of instances of the characters. D and Q are problematic, as the long strokes can either slope diagonally left to right, or right to left. The letter forms generated from this data can be compared to the letter forms generated from the stylus tablet data (see Appendix B).

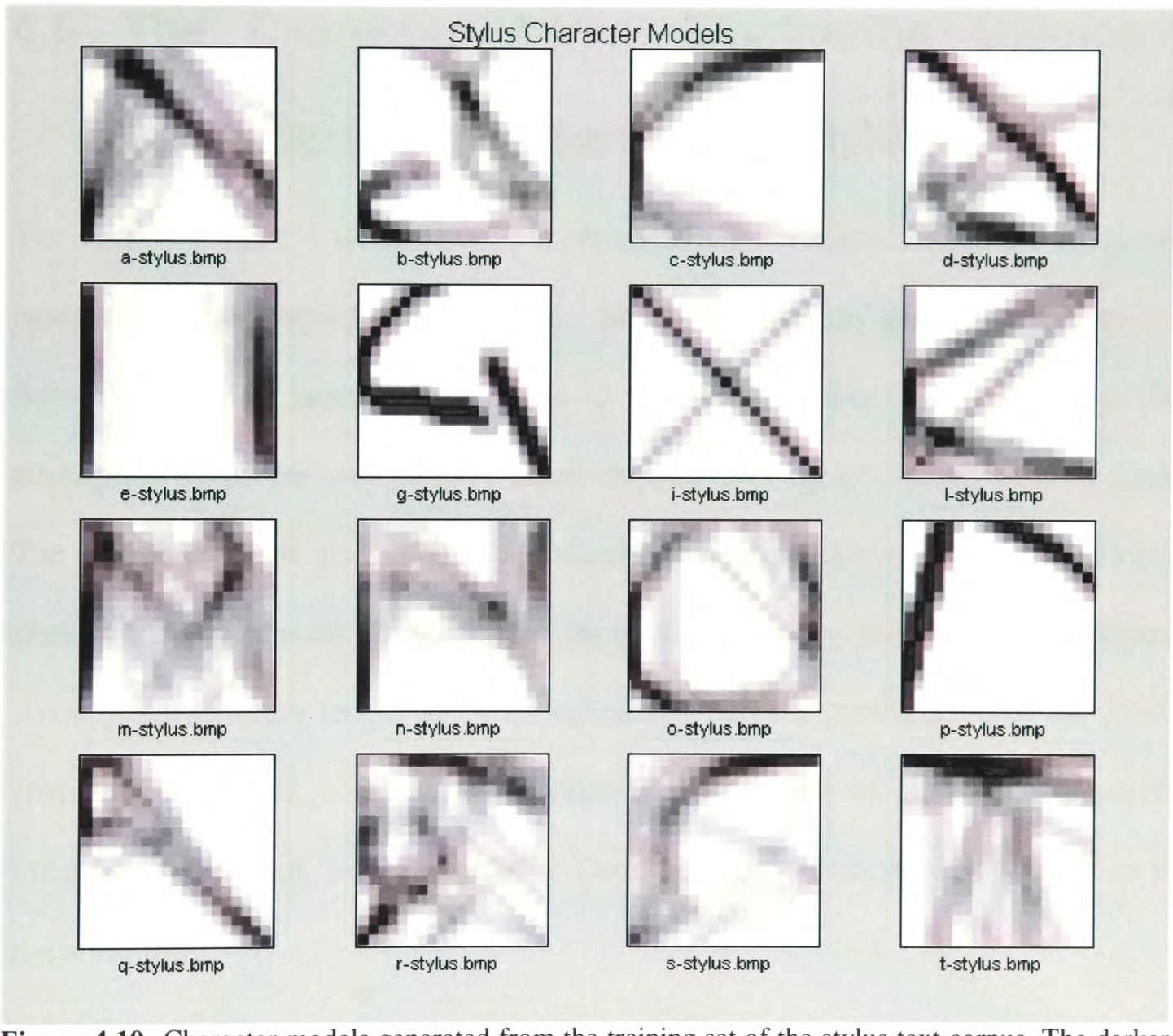



**Figure 4.10:** Character models generated from the training set of the stylus text corpus. The darker the area, the higher the certainty that a stroke will pass through that box.

Not all of the letter forms are present in this selection, as the data set of the stylus tablets was much smaller than that of the ink tablets. However, some interesting comparisons can be made between the two. The letter E shown is clearly a different form than that of the ink tablets (as discussed in 3.1.2). The letter I slopes in the

reverse direction to that of the ink tablets (although the forms of the characters are distorted due to the application of the transform, it still gives an indication of differences in the way characters are written). M, N, R, S, and T are much less standardised than in the ink tablets, perhaps because of the difficulties of writing in the different medium, and the small sample set available. Nevertheless, both sets of models provide adequate representations for comparison with unknown characters.

## 4.6 The Character Agent: Comparing Unknown Characters to the Character Models

The character agent's function is to compare unknown characters to the character models composed from the training set. This is achieved by extracting the strokes from the test data, transforming them to the standard size, then calculating the description length for matching the unknown character to each model in the data set. The character agent also utilises statistical information about the likelihood of a character being present, derived from the letter frequency analysis of the corpus. After the description length has been calculated (from a combination of the MDL frequency and MDL comparison of stroke evidence), one of the likely characters identified is selected, using the Monte Carlo sampling methods, and passed up to the word level.

### 4.6.1 Sizing and Transform

Unknown characters are transformed to the standard size in the same way as described in 4.1.1 and 4.1.2: the stroke data is extracted, a bounding box is drawn

around these strokes, and this matrix is transformed into a 21 by 21 pixel grid. No Gaussian Blur is applied to the stroke data in this case.

### 4.6.2 Calculating the Description Length

Given the probability field representation of a character, we can calculate the probability that a line will cross through any given point. To generate possible identifications of the unknown character, we calculate the probability of how the evidence (stroke data) relates to each of the character models that have previously been constructed. Identification can never be certain, but it is possible to assign probability values, to calculate how the unknown data matches each character model.

We want to find the probability that the character is *a*, given the stroke evidence:

$$P(char = a \mid strokes)$$

The stroke data is assumed to be a legal character, so the sum of the probabilities for all the models should be 1:

$$\sum_{a \in \text{models}} P(char = a \mid strokes) = 1$$

However, $P(char = a \mid strokes)$ is not directly available to us. We have to calculate it utilising Bayes' Theorem:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

Therefore:

$$P(char = a \mid strokes) = \frac{P(char = a)P(strokes \mid char = a)}{P(strokes)}$$

$P(char = a)$, the probability of a character occurring, is already available to us from the lexicostatistics. $P(strokes \mid char)$ can be calculated from the model.

The model was made up by laying $n$ different representations of a character over a standard grid, and adding up the occurrences of when the strokes passed through each box in the grid. To find the probability of one box in the grid being used as part of the character, simply take the value of the box in the composite model, and divide it by the number of characters which make the composite model.

$$P(Box_{xy} \mid char = a) = \frac{Model_a(x, y)}{nchars}$$

The stroke data can be viewed as the collection of boxes that the strokes pass through. Therefore, the total probability of the stroke data for each character is the product of the probability of the conjunctions of all the boxes passed through. If the stroke data goes through 'n' boxes, the probability of the stroke evidence given that the character is $a$ is then:

$$P(stroke \mid char = a) = P(Box_1 \cap Box_2 \cap ...Box_n \mid char = a)$$

If we assume conditional independence[12],

$$P(stroke \mid char = a) = P(Box_1 \mid char = a)P(Box_2 \mid char = a)...P(Box_n \mid char = a)$$

Which can be expressed as

$$P(stroke \mid char = a) = \prod_{i=1}^{n} P(Box_i \mid char = a)$$

Giving

[12] This assumption is not entirely valid because the boxes are not strictly independent: there is some relationship between boxes as the strokes pass through them due to the trajectory of the pen or stylus stroke. However, the assumption works well in practice.

$$P(char = a \mid strokes) = \frac{P(char = a)\prod_{i=1}^{n} P(Box_i \mid char = a)}{P(strokes)}$$

$P(strokes)$ is a value such that

$$\sum_{a \in \text{models}} P(char = a \mid strokes) = 1$$

Therefore,

$$P(strokes) = \sum_{a \in \text{models}} \left\{ P(char = a)\prod_{i=1}^{n} P(Box_i \mid char = a) \right\}$$

This gives the final equation:

$$P(char = a \mid strokes) = \frac{P(char = a)\prod_{i=1}^{n} P(Box_i \mid char = a)}{\sum_{a \in \text{models}} \left\{ P(char = a)\prod_{i=1}^{n} P(Box_i \mid char = a) \right\}}$$

So the description length of the strokes using the model for 'a' is given by:

$$DL(strokes, a) = -\log_2 \left( \frac{P(char = a)\prod_{i=1}^{n} P(Box_i \mid char = a)}{\sum_{a \in \text{models}} \left\{ P(char = a)\prod_{i=1}^{n} P(Box_i \mid char = a) \right\}} \right)$$

The description length is computed for a comparison between all available character models and the unknown character. This is added to the description length of the probability of that character occurring in the sequence (generated from the lexicostatistics). Any character with a probability of less than 0.01 is discarded. This cuts down on processing time, and discards any truly unlikely character matches. The cut off value of 0.01 was arbitrarily chosen as an outlier rejection

method.) The Monte Carlo selection algorithm then chooses, at random, which of the remaining characters will be passed upwards to the next level, the description length of this character is, of course, passed up too.

### 4.6.3 An Example

This process can be demonstrated by showing the results given when an unknown character (the second S from "ussibus", above), is compared to all of the available models generated from the ink tablets. In this case, the models most likely to fit the character were identified as S and R, shown graphically below.

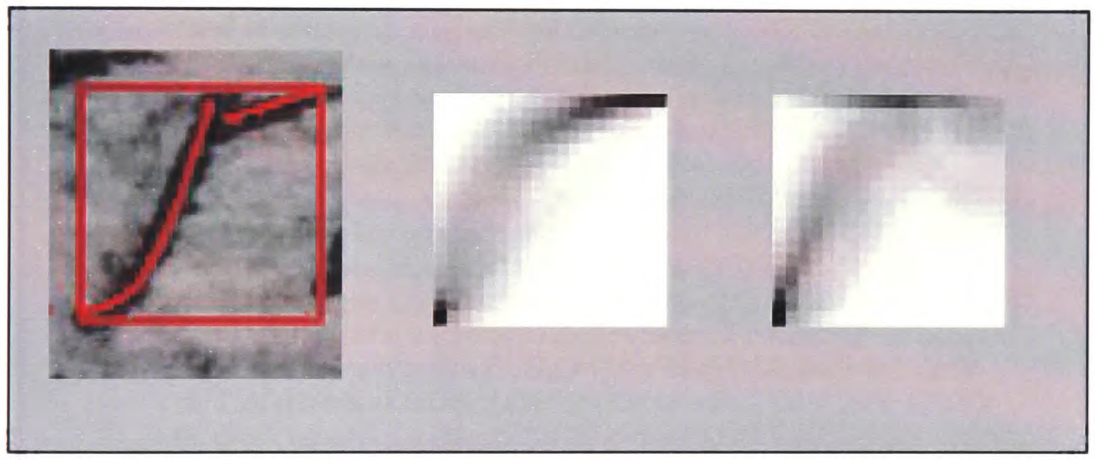


**Figure 4.11**. Letter S from ink tablet 255, compared to all the models is most likely to be S (left) or R (right).

The strokes were identified in the unknown character, a bounding box drawn around them, and the matrix was transformed and compared to all available models, resulting in a description length. This is added to the description length generated from the probability of that letter occurring (computed from the lexicostatistics), to give the overall description length of the letter. In this case, it can be clearly seen that the letter is most likely to be S or R (those results highlighted in blue for clarity): the shorter the description length, the more likely the identification is.

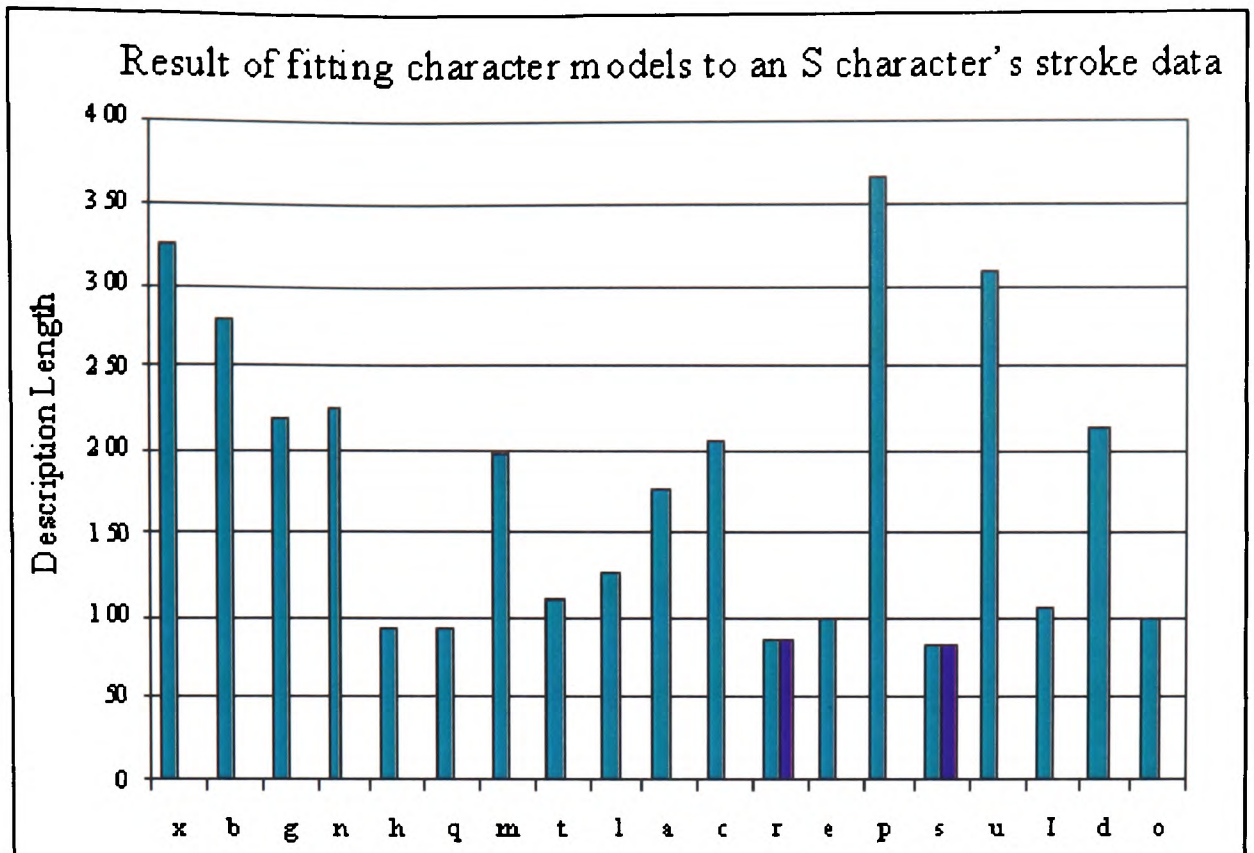



**Figure 4.12.** Fitting character models to the letter S: comparing an unknown character to model data. The order of the character models in the bar chart is of no significance. The textual portion shows the text output from the model fitting program in debug mode.

The last line in the above diagram indicates the result of the match: The character data has been estimated to be an S with probability 0.858364 or an R with probability 0.140206. The remaining probability of 0.00143 is divided out among the other characters, and the probability of those occurring (even the nearest contenders, H and Q) is so small as to not be worth considering. The Monte Carlo

method is then used to choose one of the possible identifications, R or S, to pass up to the next level: the Word Agent.

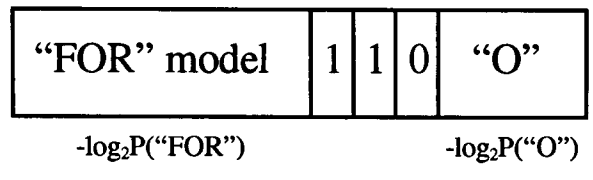
## 4.7 The Word Agent: Comparing Unknown Words to the Word Corpus

The Word agent's function is to compare strings of possible characters to the word models in the corpus, in much the same way as the character agent compares stroke data to models generated from the corpus. However, the word agent's task is considerably simpler than that of the character agent, as there is no need to represent stroke data. The word "models" are the words in the corpus themselves, and the probability of them occurring is known from the word frequency data.

At the word level, word models consist of character strings (delimited by a space character at either end). The input passed up from the character levels combine to make strings of characters, and these are compared to the word models on a letter by letter basis. If the letter matches that of the same position in the word model, it is flagged as a match. If the letter does not match, it is flagged as a non-match, and the description length from the character level is inserted in its place. The total description length of comparing that string of characters to the individual model is calculated as negative log of the probability of the word occurring, plus the sum of the description lengths of the individual characters within the string (the DLs only being present if they did not match the associated character in the model directly):

$$DL = -\log_2 P(Word) + \sum_{i=1}^{n} \left\{ -\log_2 \left( \frac{2^{-DL(strokes,CH_i)} P(CH_{i-1} \mid CH_i}{P(CH_{i-1})} \right) \right\} \text{if } CH_i \text{ doesn't match, 0 otherwise}$$

This is best understood by considering a possible message representing the fit of a word model to a character sequence. Consider fitting the word "foo" to the word model "for". The message begins with a representation of the model for the word "for", then for each character there is either a "1" bit, indicating that the character in that position matched the model, or a "0" bit indicating that the character didn't match, followed by a representation of the character that failed to match. For this example the message stream will look like this:




The description length of the match is therefore the code length for the model used, one bit for each matching character, and one bit plus the code length for each non-matching character.

The system compares the string of characters to each individual word model in the corpus. The comparison with the lowest DL generates the most likely word identification.

However, what if the string of letters represents a word which is *not* in the corpus? The system compares the string to all available models, and also generates the total description length from the sum of all of the characters' description lengths. If there is no match between the sequence and the existing models, this DL will be the lowest, and the solution is presented as a string of characters. However, due to the fact that bi-graph information is included in the word agent, the string with the minimum DL will be statistically the most likely combination of characters, even

though a word has not been matched to the input. This can be seen in the results section, 4.8.2.

### 4.7.1 An Example

The string of characters "U S S I B U S" was fitted to all 342 of the word models which had a string length of 7 characters. Only the results with a description length less than 36 bits are presented here (these being the lowest 8 description lengths computed.) It is clear that the string USSIBUS, present in the corpus, is the closest match, as this produces the lowest description length. Four other words (or word fragments present in the corpus) are identified as being as probable as the string of characters itself, SCRIBUS, SCRIBAS, VEHIMUS, UIDEDUN. This indicates how the system propagates best fit solutions, which are approximate to the correct solution.

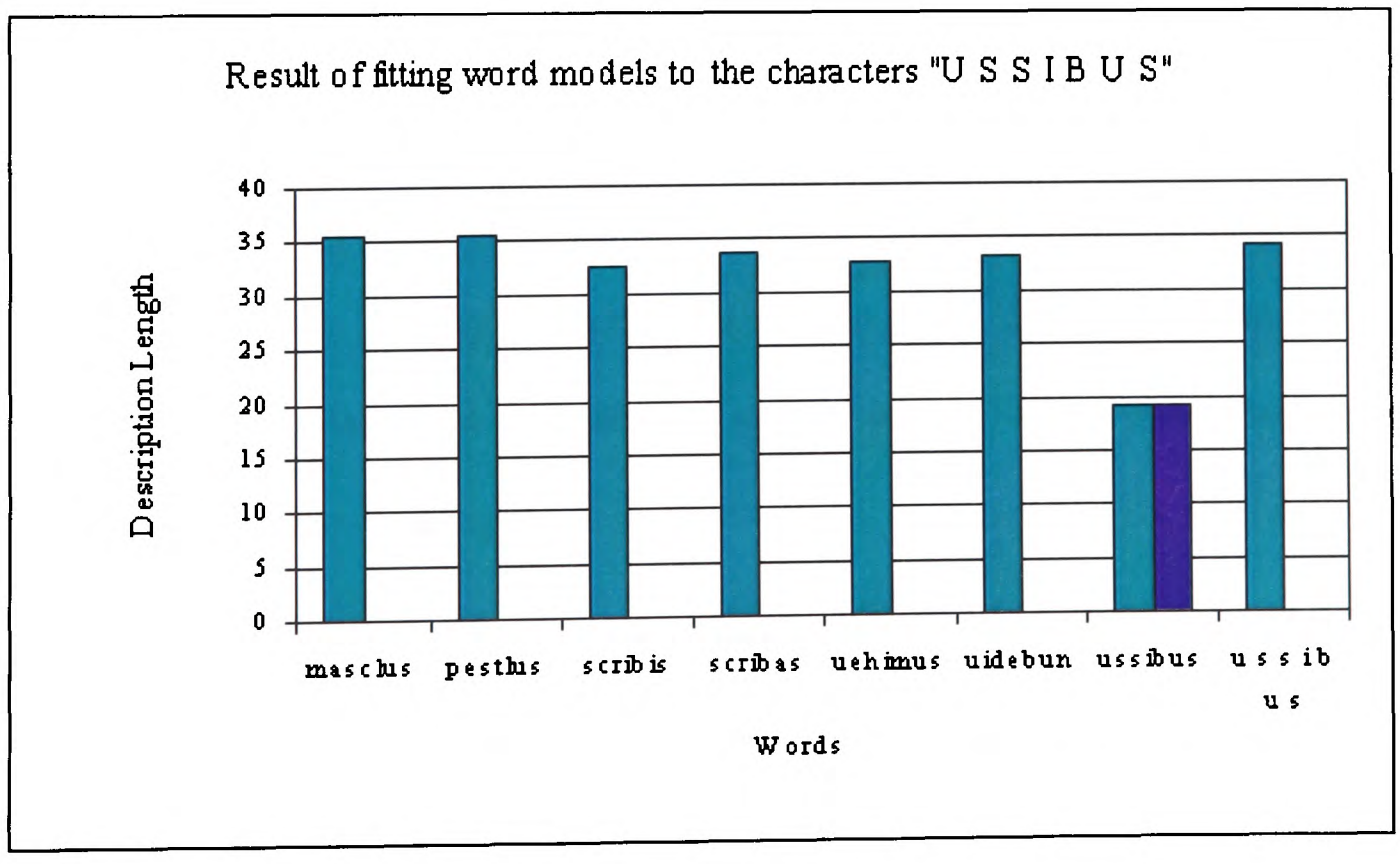



**Figure 4.13:** Result of fitting word models to the characters USSIBUS, most likely fit. The word USSIBUS is by far the most likely contender.

Again, the system chooses which word to select as a solution using a Monte Carlo selection algorithm. The system does not incorporate any other data regarding word sequencing or grammar, at this stage. The MDL for a string of words is calculated by simply summing the description length for each word. Subsequent iterations produce different sequences of characters and words. The most probable solution is that with the lowest MDL after a number of successive runs.

## 4.8 Results

This version of the system was applied to various sets of test data, to see how effective it could be in producing the correct "reading" of a text. Firstly, a section of tablet 255 was analysed, using the character models ascertained from the ink tablet corpus as the basis of comparison. This gave encouraging results, and also shows the asymptotic nature of the system's convergence on a solution. This experiment was then repeated with the same section of tablet, which in this case had been annotated in a different (wrong) manner, to see how the system coped with more difficult data. A section of stylus tablet was then analysed, using firstly the set of character models derived from the ink corpus, and secondly the set of character models derived from the stylus corpus, to indicate how successfully the system operates on the stylus texts. Finally, a section of ink tablet which had been automatically annotated (using the techniques described above in 4.3) was analysed, to indicate if this implementation of the system provided a possible solution to the problem of incorporating data generated from the image processing algorithms into a knowledge based system.

### 4.8.1 Using Ink Data for Ink Tablets

The first complete run was done on the section of tablet 255 previously used as a test set for the system (as in 4.3 above, although in this case the updated, correctly annotated version was used).

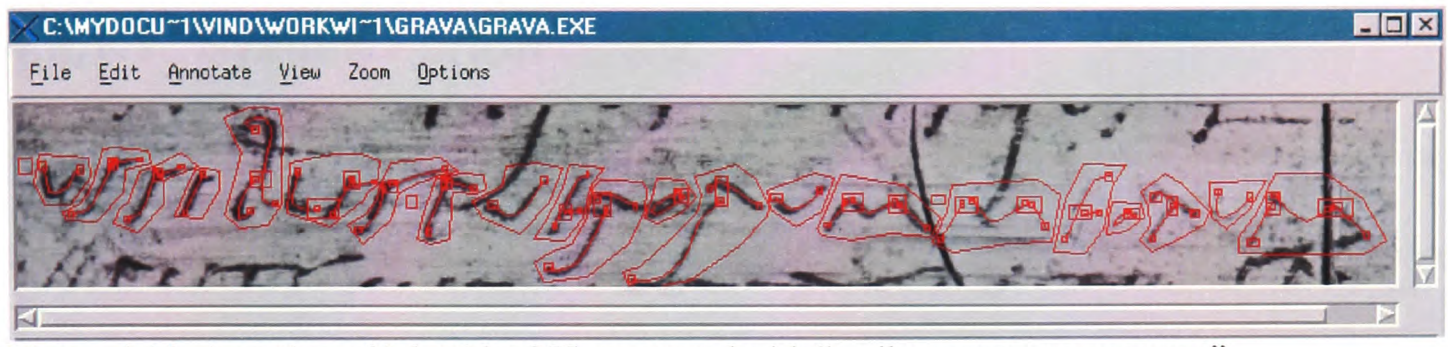



**Figure 4.14:** A section of ink tablet 255, annotated with "ussibus puerorum meorum"

The set of character models used in this run was that derived from the ink tablet corpus. The output of the system is shown below.

```
Grounded Reflective Agent Vision Architecture (GRAVA) Version 2.0.
Yolambda listener pushed. Type :exit to return to GRAVA.

=> ... load the system and the data ...

=> (runCycles 25)
iteration 0 DL=440.220794 interpretation=( ... ((2482 252) (2517 250))) ... )
iteration 1 DL=64.085075 interpretation=(  u r s i b u s puerorum  m n o a u m  )
iteration 2 DL=49.374412 interpretation=(ussibus puerorum  m n o r u m  )
iteration 3 DL=48.831413 interpretation=(ussibus puerorum  m n o a u m  )
iteration 5 DL=47.816696 interpretation=(ussibus puerorum  m e o a u m  )
iteration 8 DL=36.863136 interpretation=(ussibus puerorum meorum  )
iteration 25

=> :exit
```

**Figure 4.15:** Output from first successful run on the section of 255, using the ink tablet character models.

In this run of the system, 8 iterations took place before the correct answer was generated. (Although the system carried out 25 iterations on this data, the Minimum Description Length generated occurred in iteration 8, and so this is the last data shown. It should also be stressed that, with all of these results, the experiment was run a number of times, and the results presented are the best case, where the system generates the correct result in the fewest iterations.) Previous outputs from iteration

1, 2, 3, and 5, had been possible solutions, but that from iteration 8 proved the best, given the data provided. This was also the correct solution. The GRAVA system successfully reconstructs the correct reading of the text in a short time, on this occasion.

#### 4.8.1.1 System Performance

Although, above, the correct output was generated in merely 8 iterations, because of the stochastic nature of the process there is a possibility that the correct answer may never be found. If it is generated (in practise the correct output is generated within a few iterations) the number of iterations taken to reach this answer will be different on each run. This can be shown by determining the average description length that is generated over a variety of runs on the same data. The example, 225front7, as used above, was run 200 times, with 25 iterations specified in each run. By plotting the average description length generated from each of the 25 iterations over the 200 runs, it becomes obvious that the system converges on a result. The MDL of the "correct" result in this case was 36.863143, whilst after 25 iterations the system averaged 37.08221. This is due to the fact that the average asymptotically approaches the perfect value because of the Monte Carlo sampling methods utilised (see above 4.1).

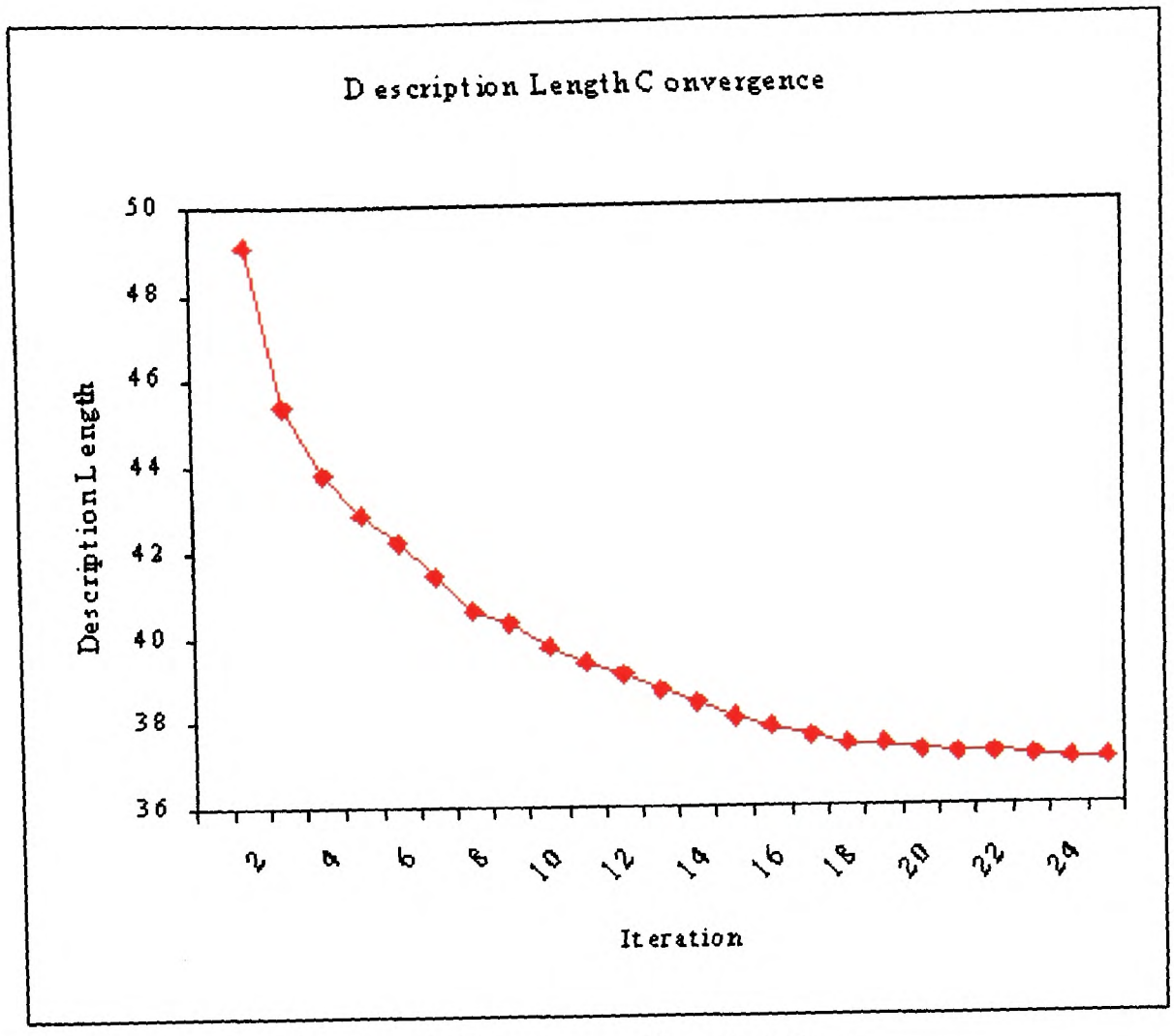



**Figure 4.16:** Description Length Convergence Over Iterations

Of course, this example only pertains to the section of 255 used as the test set. The more complex the input data, the longer the system will take to converge on a solution (which will have a different MDL). This example, however, demonstrates that the system is effective at generating likely solutions within a relatively short time frame.

### 4.8.2 Using Ink Models for an Unknown Phrase

A second section of ink tablet was analysed, this time to see how the system coped with a more confusing section of strokes, and also how it could identify a word that was not in the corpus. The first version of the annotation of 255front7 was used for this: where it had been unclear where one letter ended and another began (see 3.7.2). The difference between the two versions is that the first was annotated USSIBUS,

whilst the eventual, correct, annotation is USSIBUS. (The change was made when cross-referencing the various information ascertained from the papyrologists, and this example kept to see how the system coped with a more difficult segmentation problem).

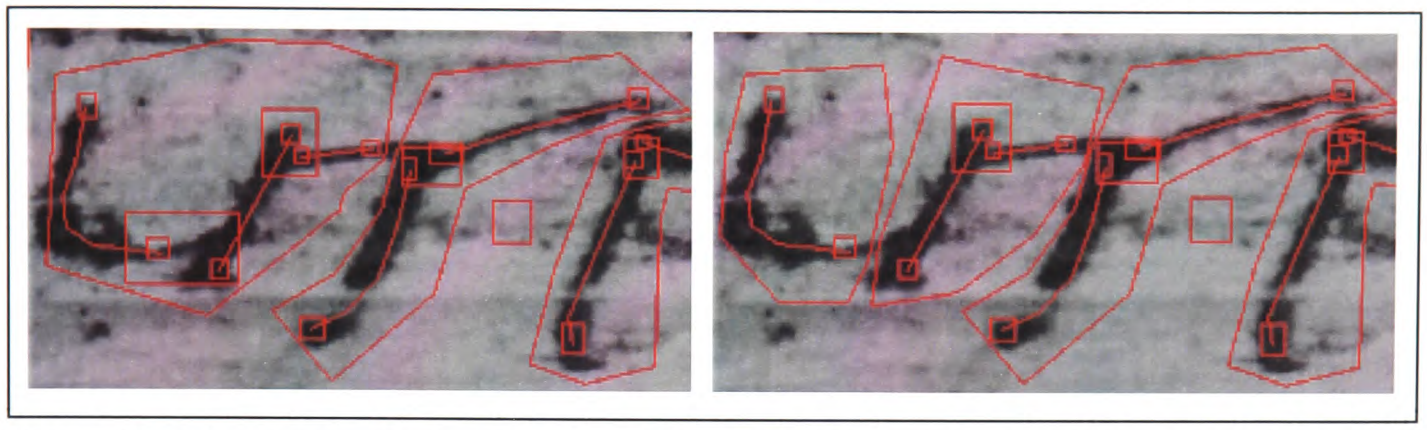


**Figure 4.17:** On the left, the section of 255 correctly annotated as US. On the right, the same section annotated incorrectly as USS.

Steps were taken to ensure that the word fragment USSIBUSS did not occur in the word corpus for this iteration, to see how the GRAVA system coped with this difficult input.

```
Grounded Reflective Agent Vision Architecture (GRAVA) Version 2.0.
Yolambda listener pushed. Type :exit to return to GRAVA.

=> ... load the system and the data ...

=>(runCycles 200)
iteration 0 DL=440.220794 interpretation=( ... ((2482 252) (2517 250))) ... )
iteration 1 DL=69.437759 interpretation=( u r s i b l n s puerorum  m n o a u m  )
iteration 3 DL=69.077354 interpretation=( u s s i b l n s puerorum  m n o a u m  )
iteration 5 DL=68.062644 interpretation=( u s s i b l n s puerorum  m e o a u m  )
iteration 6 DL=57.469482 interpretation=( u r s i b l n s puerorum meorum  )
iteration 18 DL=57.109081 interpretation=( u s s i b l n s puerorum meorum  )
iteration 25 DL=56.903217 interpretation=( u s s i b l t s puerorum meorum  )
iteration 199
#f

=>
```

**Figure 4.18:** Output from the system, analysis of awkwardly annotated segment of 255, using the ink tablet character models.

These results are interesting on a number of levels. Firstly, the system is confused by the last few characters in USSIBUSS, indicating that it is having problems identifying them. The first problem character is identified (not unreasonably) as an L, the second as either an N or a T. This shows how an unclear character can be

assigned a number of possible solutions. Secondly, although the sequence of characters (USSIBUSS) is not in the word corpus, the system does a good job at reconstructing a possible string of characters, resulting in USSIB**S. This is partly due to the use of the character models, and also the use of letter frequencies and bi-graph frequencies. This approximate solution should be enough to give some indication to a human user of what the correct word may be (the system will eventually have to interact with experts in this manner, see 5.3) Finally, the MDL generated from this solution to the problem is 56.903217. The MDL generated from the alternative (correct) annotation of the characters in 4.8.2 was 36.863143. This shows that the most likely solution to identifying the letters will have the lowest MDL, and also that there is some need, in the future, to encapsulate the opportunity to re-annotate difficult characters in a way that will eventually produce the lowest MDL to generate possible solutions.

### 4.8.3 Using Ink Models for Stylus Tablets

It was suggested in 3.1.2.1 that the letter forms from the ink tablets should correspond to those from the stylus tablets closely enough to allow models from the ink tablets to aid in readings of the letters of the stylus tablets, and this was demonstrated in 3.10. In this experiment, a section of stylus tablet 797 was analysed, firstly using models derived from the ink tablets, and in the subsequent section, using the small set of models derived from the stylus tablet corpus. This section of tablet contained fairly common words, NUNC QUID (although it had taken the experts a substantial length of time to come up with this reading.)

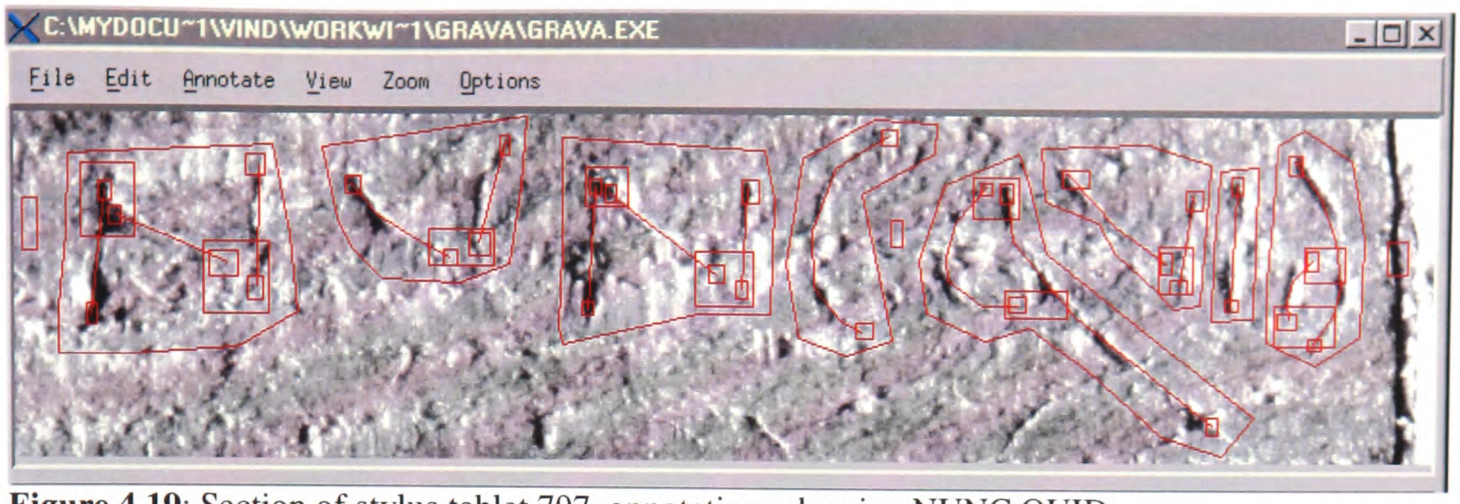



**Figure 4.19**: Section of stylus tablet 797, annotations showing NUNC QUID.

Although a fairly small sample of text, it contains a few difficult characters (the curvy letter D, for example, and the large C): the results demonstrating how the system will cope with the differences in character forms between the ink and the stylus texts.

```
Grounded Reflective Agent Vision Architecture (GRAVA) Version 2.0.
Yolambda listener pushed. Type :exit to return to GRAVA.

=> ... load the system and the data ... (Ink Models)

=> (runCycles 200 #t)
iteration 0 DL=34.567626 interpretation=(  n n n c   d u i d  )
iteration 1 DL=31.070133 interpretation=(  n n u i   d u i m  )
iteration 2 DL=31.070131 interpretation=(  n u u i   d n i m  )
iteration 5 DL=22.481197 interpretation=(  u u u i quid  )
iteration 82 DL=21.051862 interpretation=(nunc quid  )
iteration 199
#f

=>
```

**Figure 4.20**: Output of GRAVA system, section of 797 utilising ink character models.

The system took 82 iterations to reach the correct interpretation, quite a high number of run cycles, probably due to the differences in forms between the character sets. The first few iterations, as suspected, show that the system had problems with the letter D, interpreting it as an M, and the large letter C, interpreting it as an I. However, the correct reading was eventually generated when enough run cycles were allowed to sort through the various hypothesis thrown up by the data.

### 4.8.4 Using Stylus Models for the Stylus Tablets

The same section of tablet 797 was analysed, this time using the small selection of character models generated from the annotated stylus tablets.

```
Grounded Reflective Agent Vision Architecture (GRAVA) Version 2.0.
Yolambda listener pushed. Type :exit to return to GRAVA.

=> ... load the system and the data ... (Stylus Models)

=> (runCycles 25 #t)
iteration 0 DL=35.304573 interpretation=(  u u n c   q n i d  )
iteration 1 DL=34.447456 interpretation=(  u u u c   q u l m  )
iteration 5 DL=30.377746 interpretation=(unc   q n i m  )
iteration 12 DL=24.857675 interpretation=(  u u n c quad  )
iteration 16 DL=24.145236 interpretation=(  u u u c quid  )
iteration 18 DL=21.051862 interpretation=(unc quad  )
iteration 24
#f

=>
```

**Figure 4.21:** Output from section of 797, utilising the stylus character models set.

This run of the system identified the correct response in only 18 iterations, making it much more successful than the run, above, where data from the ink character models were used. There are a number of reasons why this is the case. Firstly, the character models generated from the ink and stylus tablets are *slightly* different, and this small difference must have a large effect on the comparisons. Secondly, although the stylus models character set is impoverished due to the lack of available data (see 4.5.5), it contains almost all of the characters based in this sample (save for the letter U. The model for the character U was borrowed from the ink tablet models in order to be able to try this test. The letter U seems to be preferred by the recogniser over other characters in this run. This is probably because the model for U borrowed from the ink tablet models was based on a large number of samples whereas the models for the stylus characters were based on a small sample of characters). However, the stylus character model set does not contain all the available characters, which would make it difficult to identify further examples

where less common letters predominantly feature. Although comparisons with the stylus character models appear better than the ink character models, there is still a place for the ink character models. Future implementations of the system could have more than one model for each type of letter, as discussed below (4.9).

### 4.8.5 Analysing Automated Data

It has been shown, above, that the system can correctly interpret annotated images from the corpus which were generated by hand. This is all well and good, but in essence the system needs to work effectively alongside the image processing algorithms that have been developed in tandem to this research (1.2 and 4.3.1). The section of 255 used as test data was analysed and annotated automatically[13].

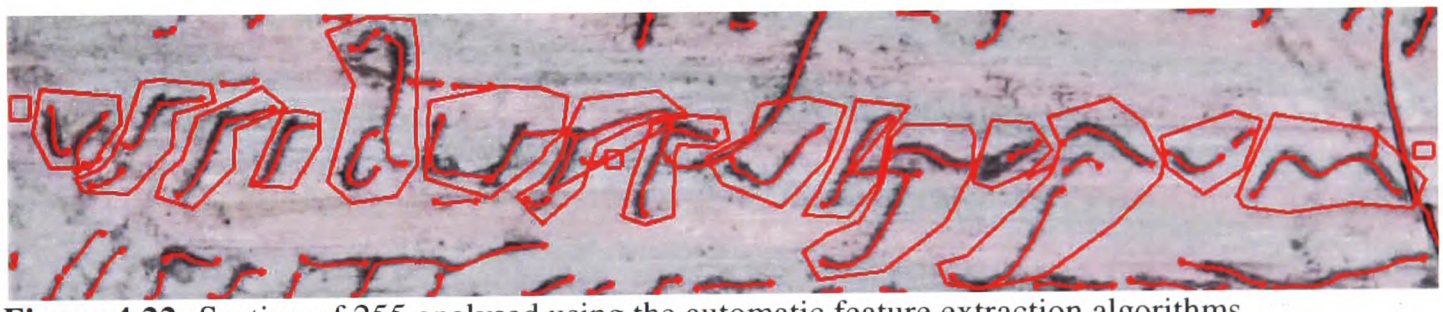


**Figure 4.22:** Section of 255 analysed using the automatic feature extraction algorithms.

This was then used in the system as test data. Due to the image data of the third word being incomprehensibly faint, this was elided from this test. The results are presented below.

[13] Again, this was carried out by Dr Xiabo Pan, from Mirada Solutions Ltd.

```
Grounded Reflective Agent Vision Architecture (GRAVA) Version 2.0.
Yolambda listener pushed. Type :exit to return to GRAVA.

=>... load the system and the data ...

=> (runCycles 200 #t)

DL=326.142322 interpretation=( . )
iteration 0 DL=24.575424 interpretation=(ussibus puerorum )
iteration 2 DL=23.990461 interpretation=(ussibus solearum )
iteration 199
#f

=>
```

**Figure 4.23:** Output from an analysis of 255, using automatically annotated images and the character models from the ink tablets.

Although the system initially finds the correct result, it goes onto find a "better" result that was wrong. This is because in the second word it identifies **ER*RUM, where * is an error. Eventually, the Monte Carlo search finds characters that match SOLEARUM well enough to get a lower description length than PUERORUM. However, the system did get the first word 100% correct at the character level, and the second word was almost recovered despite a rather poor interpretation at the character level. This shows promise towards generating possible interpretations of images through utilising this architecture. Further work needs to be done on the stroke extraction algorithms used in the automated annotation process, to encapsulate stroke data more fluidly, to enable this to be utilised by the character agent. The current algorithms have problems with the grouping of strokes, and identifying the continuation of strokes where the traces of them are faint. When these problems are overcome, and the automated output is in a clearer format, it will be easier to generate possible interpretations of the images using the GRAVA system. Nevertheless, these results show promise towards a possible solution to the problem of how to segue the image processing algorithms into an information based system to aid the papyrologists in the reading of the texts.

## 4.9 Future Work

This research has demonstrated how an MDL architecture such as the GRAVA system can be appropriated to generate possible solutions to interpretation problems encountered whilst trying to read the ink and stylus tablets. Although the examples shown were fairly straightforward, they indicate that attempting further analysis of the stylus tablets in this way will be a worthwhile endeavour. However, much can be done to improve the system, or to try and incorporate other types of information into the architecture to improve its functionality.

Firstly, it was shown that the stylus tablets were "read" easiest when the stroke data was compared with the stylus tablet models, but that some character forms were not contained in the stylus character models data set. It would be easy to amalgamate the two sets of character models into one: not combining the character models themselves, but just adding the different character forms to the existing set, so that there were two (or more) forms of different characters in the model base. (It does not matter to the Word agent if there are two different forms of the letter X at the character level, as long as one character gets passed up.) Also, the different letter forms of Q and D could be separated (where their descenders go right to left, or left to right), so that there were two character models of this letter present in the model set. By expanding and refining the character model set in this manner, the identification of unknown characters by the character agent should improve.

Secondly, there is a great deal of information regarding the character forms that was captured in the annotation process (3.5) which has not been utilised in this system. For example, it has been shown that one of the most obvious features that experts

look for when identifying letters were obvious ascenders and descenders, where strokes extend above and below the line. This information was captured in the annotation of the ink and stylus tablets images but has not been put to use in this system. It should be easy to incorporate further information into the character agent, to see if this aids in the recognition of unknown characters. There is a possibility that additional information would only affect the output of the character agent slightly, but it would be worth experimenting to see if it could be integrated into the system in this manner.

Additionally, there is a great deal of work which could be done to expand the database that the Word agent uses. The current word list is derived from the corpus of the Vindolanda texts exactly as it stands, with no changes made, or new words generated from other words in the corpus. Developing a system to output other possible word lists would be possible, but a major undertaking. This is covered in the future work chapter, Chapter 5.

Finally, this research has indicated a proof of concept: that utilising an MDL architecture can provide the necessary infrastructure as a basis to implement a system that can read and interpret images of the Vindolanda texts, and in particular, the Stylus tablets. However, there is a long way to go before an integrated desktop application has been developed which will be able to aid the papyrologists in their day to day tasks. This depends on further development of this system, and further investigation into the image processing algorithms to allow integration with this process. Again, this is covered more fully in Chapter 5.

## 4.10 Conclusion

This chapter has successfully shown how an MDL architecture such as the GRAVA system provides the infrastructure to model the reasoning process the papyrologists go through when reading an ancient document. This provides the means to implement a system that works in a similar manner as the experts, generating possible, reasonable, solutions to interpretation problems. Although not a development of a complete stand alone application, this research provides the means by which to go on and develop such a package which the experts can use to aid them in their task of reading the stylus tablets. It has been successfully demonstrated that information from the Vindolanda corpus regarding character forms and words can be used to generate plausible readings of tablets. It has also been shown that the development of a character agent that relies on data regarding the character strokes, rather than endpoints and junctions, is more effective at producing possible identifications, and that this character agent can more easily work with test information presented by the image processing algorithms.

Although much more work needs to be done on both this system and the image processing techniques before a desktop application can be implemented, this research has provided the first steps towards that aim. A possible method has been identified, implemented, and demonstrated, which enables the amalgamation of the image processing and procedural information into a cohesive whole.

# CHAPTER 5

# Future Work

And if the world were black or white entirely
And all the charts were plain
Instead of a mad weir of tigerish waters ...
We might be surer where we wished to go
Or again we might be merely
Bored but in brute reality there is no
Road that is right entirely.
Entirely, Louis MacNeice (1979).

Due to the interdisciplinary nature of this research, many topics have been covered, however briefly, in this thesis. There is scope for further research in almost every facet. Considerations of future work will focus on two main areas: other possible approaches to Knowledge Elicitation to enable further understanding of how experts read ancient documents, and the enhancement and development of the system described in Chapter 4 to increase its accuracy, and eventually deliver an application to the papyrologists.

## 5.1 Knowledge Elicitation

Although the study undertaken in Chapters 2 and 3 considering how experts read ancient texts and identify individual letter forms was as comprehensive as time and facilities would allow, there remains a considerable amount that could be undertaken in this area. Firstly, using the techniques applied in Chapter 2, it would be useful to look at the techniques used by experts who deal with other linguistic systems. This would indicate whether the findings of this research can be applied to the field in general or whether they merely relate to the reading of the Vindolanda

texts. Secondly, although the data set constructed in Chapter 3 is, again, comprehensive, there are a few more steps that could be taken to understand the palaeography of Vindolanda, and to improve the quality of the data set. Finally, little work has been done on how the experts utilise visual clues to build up an understanding of the texts; the research in this thesis mainly deals with verbal and written evidence, and does not investigate the type and quantity of data the experts visually study whilst trying to read a text. By carrying out a series of tests to investigate oculomotor action, it may be possible to see which features of the texts are the most important, and problematic to identify, when trying to piece together a reading of ancient documents.

### 5.1.1 Reading Ancient Texts

This study has focussed exclusively on how experts read the Vindolanda ink and stylus tablets. It would be possible to easily extend the remit of the survey, utilising the techniques used in this thesis, to investigate the readings of different types of ancient text. This could include such texts as papyri, ostraca, inscriptions, and lead tablets from similar and different periods than that of the Vindolanda texts. Similarly, the reading of texts that contain different letter forms, such as Capitalis and New Roman Cursive, as well as different linguistic systems, such as Cuneiform, could be investigated. Investigation into the work of experts who read and translate bilingual texts may also be fruitful in detailing how such a task is carried out (see Boswinkel and Pestman 1978; Adams, Janse et al. 2002, for an introduction to bilingual texts). A comprehensive study would contribute greatly to the literature available regarding how humans read in languages other than their own native tongue, and also how experts cope with ambiguities in texts. Although it is

suspected that the findings of the research contained within this thesis would be applicable to other areas, a larger scale investigation would show whether this was the case. Such a study would be time consuming and labour intensive, and depend on the co-operation of experts from related fields, but could prove to be an excellent research topic.

### 5.1.2 Collecting Further Data

As much data as possible, concerning how the experts operate, was collected given the time and resources at hand. However, if they were asked to discuss their reading of further documents from Vindolanda, and the resulting transcripts were analysed in the same manner as in Chapter 2, the results presented in this thesis could be statistically verified. This would also give a larger sample set from which to generate statistics regarding the relationships between different types of information used in the discussions regarding the Vindolanda texts (see 2.5.2.1). The two ink texts that were used in the initial experiment were chosen because some sections were easily readable and some were more abraded. It would be possible to set the experts different tasks using more abraded images to see how that effected the type and amount of information they discussed regarding the documents.

### 5.1.3 Collecting Character Information

The data gathered regarding the type of information used when identifying characters in Chapter 3 was more than enough to be able to construct a reliable data set of annotated images. However, it may be useful, at some stage, to use a Knowledge Elicitation tool such as WebGrid (see 3.2.3) with the experts, to directly capture information regarding each individual letter form, and each letter's

characteristics. This could provide another substantial source of information that could be used as an alternative data set with which to compare unknown letter forms, and also highlight any areas not covered by the knowledge engineer. Use of this program by more than one expert would also allow a direct comparison of the information each individual utilises (the program provides statistical tools to analyse the different types of information entered by different users). This could provide further data for Chapter 2, showing how individuals make use of similar or different types of information when reading an ancient text.

The character corpus built up by annotating the ink images provides a representative sample of the characters found within the ink documents. However, the coverage of the character forms found on the stylus tablets was poor, due to the small sample size of these texts. This raises an interesting opportunity. As more stylus tablets are read, images of them can be annotated and added to the corpus, thus improving the training set on which to base further readings. Character models will therefore be refined as more documents are read and their characters added to the existing corpus: making the feedback mechanism the papyrologists instinctively use an explicit part of the computer system[1]. Also, in the future, when a document has been successfully "read" using the system, other data regarding that document, for example words, can be fed back into the data set, thus increasing its relevance, and improving system performance.

As mentioned in 3.10, to check the quality of the annotations made in the corpus, it would be prudent to undertake a further program of annotation, where the manually

added codes could be re-annotated, and the resulting data compared to the initial annotations. Any differences in the data would then be easily detectable.

### 5.1.4 Studying Oculomotor Action

The majority of the Knowledge Elicitation exercises carried out during this research were based on the collection and analysis of textual data. There was no investigation done into how the experts read the ancient texts on a physical level, to see which features of the text they visually concentrated on, or, alternatively, features they easily identified and did not have to focus on. One way to investigate this would be to study the pattern of eye movements used by the experts as they attempt to read an ancient text. This type of research would have been infeasible until recently, as oculomotor technology has only just become cheaply and readily available.

There is a widely held assumption that there is a close correlation between pattern of eye movements and mental processes undertaken (Liversedge, Paterson et al. 1998). However, the relationship between vision, perception, reading, and language is still unknown (Gregory 1994, p.204). The study of oculomotor action is critical for the efficient and timely acquisition of visual information regarding complex visual-cognitive tasks, such as reading. How humans acquire, represent and store information about the visual environment is a critical question in the study of perception and cognition, and data regarding eye movements provides an unobtrusive measure of visual and cognitive information processing (Henderson and

[1] As papyrologists learn more about the letter forms and words used in certain types of documents, the accuracy and speed at which they can read them improves, see Bowman and Thomas for an example (Forthcoming 2003).

Hollingworth 1998). How we move our eyes when we look at a picture depends on what perceptual judgement we are asked to make (Yarbus 1967). There is a growing body of literature about the role of eye movement in scene perception, and that which exists suggests that informative areas receive more fixations than others (Underwood and Radach 1998). The role of eye movement in reading has been the focus of many experimental studies in the last twenty years (Rayner 1998), including the role of eye movement in reading music (Furneaux and Land, 1997). Techniques developed could be used to analyse how experts read ancient texts. Various oculomotor measures, such as the duration of fixations and saccades whilst reading a document, could be undertaken. Although untangling the process of reading a document from start to finish from the resulting data would be a complex task, the results would show the features of a text the experts have most difficulty identifying, or are most important for the identification of individual characters. This, in turn, could affect the focus of the work that is being carried out on feature detection and image processing, as it would indicate the areas of importance for the experts when reading an ancient text.

## 5.2 Expanding the Existing System

Although there remains a lot of research that could be done regarding understanding how experts read ancient texts, the main focus of further work should be towards expanding the system; to increase the amount and type of data it relies on and to improve its functionality. There are a number of ways in which this could be accomplished. Firstly, more statistics could be generated from the Vindolanda textual corpus, and these integrated into the system. Secondly, the word list generated from the textual corpus could be expanded by analysing and developing

the existing corpus. Further linguistic work could be undertaken on a grammatical and semantic level. Character information that was captured during the annotation process could be integrated with the existing system, which may improve its functionality. Additional testing is required, and further work done with those developing the image processing algorithms to ensure that the automatically generated annotations are of suitable form and quality to be used in the system.

### 5.2.1 Expanding Statistics from the Extant Corpus

The current system relies on data extracted from the Vindolanda ink tablet textual corpus, namely letter frequency information; a bi-gram analysis; a word list; and word list frequencies. There are other types of information that could be easily derived from this corpus and integrated into the system as it stands to possibly improve its functionality. For example, it would be simple to carry out a tri-gram, and quad-ram analysis, to investigate further the intrinsic statistical qualities of the letter sequences in the Vindolanda ink texts, but this information may not have any effect on system performance. Of more relevance would be the generation of common word endings from the existing corpus, and the incorporation of this information into the existing system. Additionally, expanding the available word list from the Vindolanda ink text corpus would provide more information for the system to utilise, and increase the likelihood of a word match being found.

### 5.2.2 Expanding the Word List

No attempt has been made to modify the word list derived from the ink tablets corpus (see 4.2): the word list of the system is currently limited to *exactly* what is featured in the Vindolanda textual corpus. This word list is presented as a "left-to-

right" or "finite state" model of vulgar Latin at Vindolanda, a type of language model which has been repeatedly shown to be deficient and inadequate since the early phases of Psycholinguistic research (Chomsky 1957). The inability to generate possible new words from the existing data is, at present, a major fault in the system as "It is important to bear in mind that the creation of linguistic expressions that are novel but appropriate is the normal mode of language use" (Chomsky 1972, p.100). It would be possible to utilise the existing corpus to generate other instances of text that could be found on the documents, thus expanding the data set available. This could be done by:

- Splitting existing words into fragments, and presenting these as possible letter sequences. This would be simple to implement.
- Developing some understanding of the underlying grammatical models at play during the formation of words. By doing so, it would be possible to generate new instances of words, for example by identifying the root of a word, and suffixing a different ending to present the word as used in a different tense. However, it would be a complex task to implement this computationally. It may be easier to generate a derived word list manually.

It should be easy to write a small routine which provides fragments of the existing words as possibilities to the system (many of the texts that are encountered only have a section of words extant, and the papyrologist often completes the word, as shown in 2.5.5.3). For example, the word USSIBUS may be split into many different fragments, such as USS, USSI, USSIB, USSIBU, SSI, SSIB, SSIBU, SSIBUS, etc. These additional word fragments may very well match text on the stylus tablets, when the whole word is not present due to damage. This is an

important limitation of the system at present, as it will only identify words which are present in the word list. Due to the fragmentary nature of the Vindolanda stylus tablets there is a high probability that complete words from the word list, or the exact match of fragments already present, would not be found.

In order to mimic the papyrologists' use of language, it will be necessary to integrate into the system some simple word creation mechanism. This would rely on the existing word list, statistics, and grammatical representation, to provide the basis of new words. This is a daunting task, given that

> even after constant research by hundreds of great minds, linguistics still lacks an adequate representation of *English*! Linguists know even less about other tongues (Stainton 1996, p.135).

However, due to the way Latin uses word endings, it should be possible to construct a rudimentary model of how these work, and apply them to root forms of words, to generate possible new words. Although there has been some success recently in modelling language systems to create new words using grammatical rules (Plunkett 1995), this would be a complex project. It would also be hampered by the fact that the grammar contained within the Vindolanda texts is not that of standard classical Latin, and that the word corpus itself is perhaps too fragmentary to generate any concrete grammatical rules regarding the use of language at Vindolanda. One possibility is to loosely utilise research which has been done on the syntax and grammar of classical Latin (Mahoney 2000), although it would be debatable how applicable this would be to the words in the database. It is therefore a task that should not be undertaken lightly.

An alternative solution would be to generate a derived word list manually. The experts could be questioned about the word list, and the knowledge engineer could manually work through the corpus, constructing different forms and tenses of verbs, plural nouns, adjectives, etc. This could provide an alternative word list, which would provide additional material for the system (although there would not be any statistical information available regarding the occurrence of words in the corpus).

### 5.2.3 Further Linguistic Work

If the sample set had been larger and less fragmentary, it would have been possible to gain an understanding of the word order, and higher level grammatical models at play, within the language of the Vindolanda documents. This could have been done by grammatically labelling the corpus on a word by word level and studying the patterns that emerge from the text (Lawler and Dry 1998). This qualitative data could be used to predict what *type* of word was needed next in a document, and so could narrow down the search for an appropriate match. However, because of the fragmentary nature of the corpus, no work on grammar was undertaken. It is something that should be kept in mind for the future, but it is not a priority in this research.

A further possibility in regard to providing alternative data sets would be to gather together groups of words that are linked by subject, for example, all the words in the corpus with respect to the Roman Army; food; transport; horses; leather goods; proper names; place names; numerals; dates; medicinal goods; livestock; trading; slaves; etc. Lists of words by subject are a tool commonly used by papyrologists (André 1981; André 1985; André 1991), and there is no reason why this could not

be integrated into a computer system to provide these in electronic format. Locating words which belong to a probable subject could narrow down the search for a possible word fit, as, when directed by a human expert who was able to predict the context of a word, the system could search within a subset that would be relevant to the unknown word. This may or may not be helpful to the experts – perhaps if they already know the context of the word they are searching for, they would be able to identify the unknown section themselves! The construction of such word sets would involve a great deal of research.

### 5.2.4 Including Character Information

The character models currently used within the system are based solely on the stroke data that was captured during the annotation process. There were many other types of information that were captured when annotating the corpus. For example, it was noted when strokes went above and below the normal writing line: ascenders and descenders being one of the key features that the experts rely on to identify individual characters[2]. It should be possible to integrate this information into the present system, which may or may not improve its functionality, but would be worth investigating. Additional character models could be constructed using the manually added textual codes, to build up a textual description of individual characters. These could be compared to the output of the feature detection agents in the same way as present, by generating SGML files containing information regarding the characters.

[2] This has also been shown to be the case when reading clearly printed text: the shape of words (whether they have ascenders or descenders, etc) affects the time taken to recognise them (Rayner 1998).

### 5.2.5 Testing and Evaluation

The system needs to be tested on more complex lines of text, to see how it copes with more complicated data. This is particularly the case with the stylus texts, as the example shown in 4.8 was a fairly simple one. Additionally, the system needs to be tested utilising output from the image processing algorithms generated from images of the stylus tablets. Considerable work on the image processing techniques is necessary to allow clear annotations to be generated automatically so that these can be used by the system.

## 5.3 Application Development

The primary purpose of undertaking this research was to aid in the construction of an application that would assist the papyrologists in their reading of the Vindolanda stylus tablets. Although this thesis has detailed the success so far of this project, there is still significant amounts of work to be done regarding image processing and integration with the GRAVA architecture before a stand alone system can be delivered to the experts. However, when those obstacles are overcome, it will be simple to utilise the architecture developed in this thesis as the basis of such a system. Because the GRAVA system is written in YOLAMBDA, a dialect of LISP, it will be easy to deliver the application using JAVA to end users due to its object oriented and flexible nature. There are a number of features that could be built into the interface to assist the papyrologist, and these are detailed below.

The interface must allow the expert to interact with the system. This has never been an attempt to create a system which can automatically "read" texts without any

input from a human user, although the architecture presented can reason independently from the user. This interaction can be achieved in many ways.

- The system should provide a "one stop shop" so that an expert can load in an image, have the image processing algorithms detect candidate strokes, and then propagate possible solutions based on this data and the statistics held within the system. Interaction should be possible with both processes, to change parameters and highlight information during the image processing, and to guide the system through the interpretation of the annotations. Selection of the data sets will be possible (if more than one word list is available, for example). It is the expert who should be in control of the system, not the other way around!
- Strokes will be able to be traced manually, allowing the program to reinterpret the stroke data and propagate further solutions.
- The image processing parameters will be adjustable in order to re-annotate less clear areas in a different manner, allowing the system to propagate alternative solutions.
- Any suggestions which may seem unlikely to the user could be discarded. The system could then propagate further alternative solutions.
- The expert should be able to see as many possible solutions as needed, not merely the one which has been deemed the closest. This could provide clues with which to obtain a true reading of the text.
- Solutions could be suggested by the user, with the system calculating the probability of these matching the stroke data present.
- The system should keep track of all changes made and tools used. This will provide a record of the reasoning process the expert undertook: something that

is missing from the current documentation. This information should be available in a clear format, and retained for future use.

- An "undo" function should be incorporated into the system to return it to a prior state, all the while keeping track of the reasoning process undertaken.
- Images of the annotated text should be easy to output to retain a visual record of the reading.
- A recursive mechanism would be put in place, where once a character or word has been identified they are added to the data sets to increase the information available for future runs of the system.
- The user should be able to add words to the word list as s/he sees fit, to increase the amount of relevant information available. Records would be maintained of any information that is added in this manner.
- By default the system should be set to run using data from the original word list. Additional corpora will be kept separate to preserve the integrity of the data.
- The user interface should be intuitive. A comprehensive help menu and full documentation should be provided.
- The program should incorporate commonly used image processing tools, such as zoom, inverse, contrast and brightness control, so that the experts can view images in the manner that they are accustomed to. Although these tools are already available in PhotoShop, incorporating them into the package would end the need for switching between different programs.

## 5.4 Conclusion

There remains a significant amount of work that needs to be done to develop this system, to increase the functionality and accuracy of the existing system. In the

future, this could be used as the basis to program a stand-alone application to aid the experts in their reading of the Vindolanda stylus tablets. Various adjustments to the system have been suggested, such as integrating easily obtainable statistical information, and more complex consideration has been made as to how this system may function in the future. Although there remains a considerable amount of work to be carried out, the architecture of the developed system can be adapted easily, and provides the basis for the development of a stand-alone application.

# CHAPTER 6

# Conclusion

"The king cried aloud to bring in the astrologers, the Chăl-dē-ăns, and the soothsayers. And the king spake, and said to the wise men of Babylon, Whosoever shall read this writing, and shew me the interpretation thereof, shall be clothed with scarlet." (Daniel, 5:7, (1953))

This chapter highlights the overall contribution the research has made to both Engineering Science, and Classics. Suggestions for how this type of system could be adopted to aid humans in the interpretation of other types of data are made, and an evaluation of the project is presented.

## 6.1 Contribution

Given the scope of this research, it is not difficult to identify the unique contribution it has made to both the fields of Engineering Science and Classics.

Firstly, by asking how experts operate, it has been possible to understand better the processes the papyrologists go through when reading ancient texts, which may be helpful to other scholars who work on primary source documents. Secondly, through the use of Knowledge Elicitation techniques, this process, which had previously not been studied, has been made explicit, allowing a model to be proposed of how humans read and interpret very ambiguous texts; a topic of interest to psychologists. Additionally, the type of information that is used when identifying individual characters has been made explicit. This allowed a framework to be developed that can annotate written text (there is no reason why this cannot be used

to annotate other forms of writing other than the ORC contained within the Vindolanda Texts). The corpus of ORC characters that were annotated during this project is the only electronic palaeographic resource of its kind regarding this form of handwriting, and so may prove to be of use to scholars. The corpus has already given some insights into the letter forms used at Vindolanda, by providing a means to compare the characters present on the ink and stylus tablets. The statistics derived from the Vindolanda ink text word corpus also provide a resource for scholars in the field.

From a computational angle, the use of an MDL based architecture has demonstrated that it is possible to build a large system that can reason about different types of complex data efficiently, propagating useful solutions to interpretation problems. In effect this has paved the way for the construction of a "cognitive visual system": one that can read in image data, and output useful interpretations of that data. Although there remains work to be done to dovetail this system with the feature detection image processing algorithms, the success of this research indicates that utilising an MDL based architecture in this manner provides the framework necessary to build complex knowledge based image processing systems.

Although this research did not deliver a stand-alone application for the papyrologists to use to aid in reading the stylus tablets, this was not the primary aim of the project. It has provided an understanding of the type of tools required by the experts, as well as implementing a system that can analyse image data and propagate useful interpretations. Further testing and development is necessary

before a completed application can be made available, but the findings presented here provide the basis for the construction of such a system: a fruitful culmination of varied, interdisciplinary research.

## 6.2 Future Directions

The research presented here presents many opportunities for future work. From a humanities angle, this type of computer tool could prove to be instrumental in reading various types of documents that were illegible to the human eye: the joining of image processing and linguistic information allowing many possible interpretations of data to be generated to aid experts in their task. It would not be difficult to adapt this system to other linguistic systems given that the necessary statistics were available, and the primary sources made available for digitisation. Just how useful such a tool actually is to those who read such primary sources remains to be seen, but papyrology as a field has so far embraced computer based tools and resources rapidly.

MDL based architectures could be used on any number of image processing tasks, where complex information from other semantic levels needs to be used to interpret images. It has been shown here that MDL provides the common currency to relate different types of information, and this could be investigated further. An architecture such as the one described in this thesis could be used to read other types of handwriting, but the architecture could be expanded much further, to incorporate other semantic levels, such as grammar and contextual information. MDL architectures could also be used for entirely different image interpretation problems, such as aerial image analysis, sign language interpretation, or medical image

analysis, as long as procedural information from different semantic levels was available or obtainable, to allow complex hierarchical systems to be implemented. So far as to say, MDL architectures could provide the basis for the development of any type of computer based interpretation system, for example: speech recognition (or production), the analysis and interpretation of physical processes (the monitoring of weather, water flow and direction), predicting the outcome in war gaming systems, etc. The scope for the appropriation and development of this type of architecture is almost limitless: the important point being it provides a way of comparing and contrasting semantically different types of information fairly and efficiently to generate the best probable outcome from available data.

## 6.3 In Retrospect

Although this research has had some demonstrable successes, if it were to be carried out again, there would be a few changes made to the direction taken. Less time would have been spent on the least helpful knowledge elicitation tasks (for example the semantic content analysis of the published commentaries), although it is the nature of such tasks that the benefit of carrying out such an analysis does not become apparent until the data has been collected. More effort would have been made in understanding the linguistic corpus of Vindolanda, perhaps enabling the construction of subject based word sets, and an understanding of how word endings operated. The annotation encoding scheme would have been developed further to become TEI compliant. The annotated corpus would have been subject to further checks to establish the correctness of the annotations, and further testing would have been carried out on more complex examples of the Vindolanda stylus texts. Nevertheless, this project has had some demonstrable success in having

implemented a system to aid the historians in carrying out their work. It is hoped that in the future the direction of this work will be continued to deliver a stand-alone application to the experts.

## 6.3 To Conclude

This research presents a novel approach to a complex problem, delivering a system that can generate plausible interpretations from images, in the same way that human experts appear to do, to aid them in their task. In doing so, areas of further research have been presented, offering further opportunities to develop intelligent systems that can interpret image data effectively, to aid human beings in complex perceptual tasks.

# Bibliography

Aalto, P. (1945). "Notes on Methods of Decipherment of Unknown Writings and Languages." Studia Orientalia, Edidat Societas Orientalis Fennica XI.4.

Adams, J. N. (1995). "The Language of the Vindolanda Writing Tablets: An Interim Report." Journal of Roman Studies LXXXV: 86-134.

Adams, J. N., M. Janse, et al., Eds. (2002). Bilingualism in Ancient Society - Language, Contact and the Written Word. Oxford, Oxford University Press.

Aiken, L. R. (1971). Psychological Testing and Assessment. London, Allyn and Bacon.

Aitchison, J. (1998). The Articulate Mammal, An Introduction to Psycholinguistics. London, Routledge.

Anderson, B. (1974). The Quantifier as Qualifier: Some Notes on Qualitative Elements in Quantitative Content Analysis. Gothenburg, University of Gothenburg.

André, J. (1981). L'Alimentation et la Cuisine à Rome. Paris, Belles Lettres.

André, J. (1985). Les Noms de Plantes dans la Rome Antique. Paris, Belles Lettres.

André, J. (1991). Le Vocabulaire Latin de l'Anatomie. Paris, Belles Lettres.

Asher, R. E., Ed. (1994). The Encyclopedia of Language and Linguistics. Oxford, Pergamon.

Bagnall, R. S. (1997). "Imaging of Papyri: A Strategic View." Literary and Linguistic Computing 12(3): 153-154.

Bately, J., M. P. Brown, et al., Eds. (1993). A Palaeographer's View, The Selected Writings of Julian Brown. London, Harvey Miller Publishers.

Bearman, B. H. and S. Spiro (1996). "Archaeological Applications of Advanced Imaging Techniques." Biblical Archaeologist 59:1.

Biber, D. (1983). "Representativeness in Corpus Design." Literary and Linguistic Computing 8: 243-57.

Biber, D. (1990). "Methodological Issues Regarding Corpus-based Analysis of Linguistic Variation." Literary and Linguistic Computing 5: 257-69.

Bible (1953). The Holy Bible, King James VI's Version. Glasgow, Collins' Clear-Type Press.

Bidwell, P. (1997). Roman Forts in Britain. London, English Heritage.

Bidwell, P. T. (1985). The Roman fort of Vindolanda at Chesterholm, Northumberland. London, Historic Buildings and Monuments Commission for England, 1985.

Birley, E., R. Birley, et al. (1993). The Early Wooden Forts: Reports on the Auxiliaries, the Writing Tablets, Inscriptions, Brands and Graffiti, Hexham : Roman Army Museum Publications for the Vindolanda Trust.

Birley, R. (1977). Vindolanda, A Roman Frontier Post on Hadrian's Wall. London, Thames and Hudson.

Birley, R. (1999). Writing Materials. Greenhead, Roman Army Museum Publications.

Bischoff, B. (1990). Latin Palaeography, Antiquity and the Middle Ages. Cambridge, Cambridge University Press.

Boose, J. and B. Gaines, Eds. (1990). The Foundation of Knowledge Acquisition. London, Academic Press.

Boose, J. H. (1990). Uses of Repertory Grid-Centered Knowledge Acquisition Tools for Knowledge-Based Systems. Foundations of Knowledge Acquisition. J. Boose and B. Gaines. London, Academic Press: 61-83.

Boswinkel, E. and P. W. Pestman, Eds. (1978). Textes Grecs, Demotiques et Bilingues. Papyrologica Lugduno-Batava XIX. Leiden.

Bowman, A. and J. D. Thomas (1983). Vindolanda: The Latin Writing Tablets. London, Society for Promotion of Roman Studies.

Bowman, A. and J. D. Thomas (1994). The Vindolanda Writing-Tablets (Tabulae Vindolandenses II). London, British Museum Press.

Bowman, A. K. (1994). Life and Letters on the Roman Frontier. London, British Musuem Press.

Bowman, A. K. (1997). The Vindolanda Writing Tablets. XI Congresso Internazionale di Epigrafia Greca e Latina, Roma.

Bowman, A. K., J. M. Brady, et al. (1997). "Imaging Incised Documents." Literary and Linguistic Computing 12(3): 169 - 176.

Bowman, A. K. and J. D. Thomas (Forthcoming 2003). The Vindolanda Writing - Tablets (Tabulae Vindolanses III). London.

Bowman, A. K. and R. S. O. Tomlin (Forthcoming 2003). Wooden Stilus Tablets from Roman Britain. Images and Artefacts of the Ancient World, London, British Academy.

Brady, M., X. Pan, et al. (Forthcoming, (2003)). Shadow Stereo, Image Filtering and Constraint Propagation. Images and Artefacts of the Ancient World, London, British Academy.

Breeze, D. J. and B. Dobson (1976). Hadrian's Wall. London, Richard Clay (The Chaucer Press).

Brown, M. S. and W. B. Seales (2001). 3D Imaging and Processing of Damaged Texts. ACH/ALLC, New York University.

Carr, M., J. Durand, et al. (1994). Translation Theory and Practice, a Discourse. Salford, University of Salford.

Casamassima, E. and E. Staraz (1977). "Varianti e Cambio Grafico Nella Scrittura dei Papiri Latini, Note Palaeografiche." Scrittura e Civiltà: 9-110.

Cattel, J. M. (1886). "The Time Taken Up by Cerebral Operations." Mind II: 220-242.

Cencetti, G. (1950). "Note Paleografiche Sulla Scrittura dei Papiri Latini dal I al III Secolo D.C." Acccad. delle Scienze dell'Instituto do Bologna, Cl. di scienze morali, Memorie 5(I).

Charniak, E. (1993). Statistical Language Learning. Cambridge, MA, MIT Press.

Chomsky, N. (1957). Syntactic Structures. The Hague, Mouton.

Chomsky, N. (1972). Language and Mind. New York, Harcourt Brace Jovanovich.

Coleman, R. G. G. (1993). "Vulgar Latin and Proto Romance: Minding the Gap." Prudentia 25.

Connell, J. H. and M. Brady (1987). "Generating and Generalizing Models of Visual Objects." Artificial Intelligence 31(2): 159-183.

Conway, R. S. (1923). An Introduction to Latin, Greek and English Etymology. London, John Murray.

Cordingley, E. (1989). Knowledge Elicitation Techniques for Knowledge Based Systems. Knowledge Elicitation: Principles, Techniques and Applications. D. Diaper. Chichester, Ellis Horwood: 90-103.

Côté, M., E. Lecolinet, et al. (1998). "Automatic Reading Of Cursive Scripts Using A Reading Model And Perceptual Concepts: The PERCEPTO System." International Journal of Document Analysis and Recognition 1(1): 3-17.

Crain, S. and M. Steedman (1985). On Not Being Led Up the Garden Path: The Use of Context by the Psychological Syntax Processor. Natural Language Parsing: Psycholinguistic, Computational and Theoretical Perspectives (Studies in Natural Language Processing). D. Dowty, L. Karttunen and A. M. Zwicky. Cambridge, Cambridge University Press: 320-358.

Dahl, R. (1997). The Great Automatic Grammatizator and Other Stories. London, Puffin Books.

Danks, J. H. and J. Griffin (1997). Reading and Translation, A Psycholinguistic Perspective. Cognitive Processes in Translation and Interpreting. J. H. Danks, G. M. Shreve, S. B. Fountain and M. K. McBeath, Sage Publications.

Danks, J. H., G. M. Shreve, et al. (1997). Cognitive Processes in Translation and Interpreting, Sage Publications.

de Carteret, C. and R. Vidgen (1995). Data Modelling for Information Systems. London, Pitman.

de Groot, A. M. B. and C. Barry (1994). The Multilingual Community: Bilingualism. Hillsdale (USA), Lawrence Erlbaum Associates.

Dermott, J. (1988). Preliminary Steps Toward a Taxonomy of Problem Solving Methods. Automating Knowledge Acquisition for Expert Systems. S. Marcus. Lancaster, Kluwer Academic Publishers.

Diaper, D. (1989). Knowledge Elicitation: Principles, Techniques and Applications. Chichester, Ellis Horwood.

Dodel, J. P. and R. Shinghal (1995). Symbolic/ Neuronal Recognition of Cursive Amounts on Bank Cheques. Third International Conference on Document Analysis and Recognition, Montreal.

Duncker, K. (1926). "A Qualitative (Experimental and Theoretical) Study of Productive Thinking (Solving of Comprehensible Problems)." Pedagogical Seminar 33: 642-708.

Ellegard, A. (1963). English, Latin, and Morphemic Analysis. Gothenburg, University of Gothenburg.

Ellis, R. and G. Humphreys (1999). Connectionist Psychology, A Text With Readings. Hove, Psychology Press.

Ericsson, K. A. and H. A. Simon (1993). Protocol Analysis, Verbal Reports as Data. Cambridge, Massachusetts, MIT Press.

Eysenck, M. W. and M. T. Keane (1997). Cognitive Psychology, A Student's Handbook. Hove, Psychology Press.

Feigenbaum, E. A. (1977). The Art of Artificial Intelligence: Themes and Case Studies in Knowledge Engineering. IJCAI-77.

Fink, R. O. (1971). Roman Military Records on Papyrus, Case Western Reserve University.

Forster, M. R. (1999). The New Science of Simplicity; Simplicity, Inference and Econometric Modelling. H. A. Keuzenkamp, M. McAleer and A. Zellner. Cambridge, Cambridge University Press.

Franzosi, R. (1990). "Strategies for the Prevention, Detection and Correction of Measurement Error in Data Collected from Textual Sources." Sociological Methods and Research 18: 442-471.

Franzosi, R. (1997). Labor Unrest In the Italian Service Sector: An Application of Semantic Grammars. Text Analysis for the Social Sciences. C. W. Roberts. Mahwah, NJ, Lawrence Erlbaum Associates: 131-145.

Fu, K. S. and P. H. Swain (1969). On Syntactic Pattern Recognition. Software Engineering. J. T. Tou. 2: 155-182.

Furneaux, S. and Land, M.F. (1997), The Role of Eye Movements During Music Reading. Proceedings of the Third ESCOM Conference, Uppsala. 210-214.

Gaines, B. R. and M. L. G. Shaw (1997). WebGrid: Knowledge Modeling and Inference through the World Wide Web, Knowledge Science Institute, University of Calgary. 2002. http://repgrid.com/reports/KBS/KMD/

Gao, Q. and L. Ming (2000). "Applying MDL to Learning Best Model Granularity." Artificial Intelligence 121: 1-29.

Gibson, E. J. and H. L. Levin (1976). The Psychology of Reading. Cambridge Mass, MIT Press.

Gonzalez, R. C. and R. E. Woods (1993). Digital Image Processing, Addison-Wesley Publishing.

Goodman, K. S. (1967). "Reading: a Psycholinguistic Guessing Game." Journal of the Reading Specialist 6: 126-135.

Gough (1977). One Second of Reading. Cognitive Theory, 2. N. J. Castellan, D. B. Pisoni and G. R. Potts. Hillsdale, Lawrence Erlbaum Associates Inc.

Gregory, R. L. (1994). Eye and Brain, the Psychology of Seeing. Oxford, Oxford University Press.

Groot, A. M. B. (1997). The Cognitive Study of Translation and Interpretation. Cognitive Processes in Translation and Interpreting. J. H. Danks, G. M. Shreve, S. B. Fountain and M. K. McBeath, Sage Publications.

Henderson, J. M. and A. Hollingworth (1998). Eye Movements During Scene Viewing: An Overview. Eye Guidance in Reading and Scene Perception. G. Underwood. Oxford, Elsevier: 269-293.

Herman, J. (2000). Vulgar Latin. T. B. R. Wright. Pennsylvania, Pennsylvania State University Press.

Hilton, O. (1959). "Characteristics of the Ball Point Pen and its Influence on Handwriting Identification." Journal of Criminal Law, Criminology, and Police Science 47: 606-13.

Hilton, O. (1984). "Effects Of Writing Instruments On Handwriting Details." Journal of Forensic Sciences 29: 80-6.

Hirst, G. (1987). Semantic Interpretation and the Resolution of Ambiguity. Cambridge, Cambridge University Press.

Hoey, M. (2001). Textual Interaction, An Introduction to Written Discourse Analysis. London, Routledge.

Hogaboam, T. W. and C. A. Perfetti (1975). "Lexical Ambiguity and Sentence Comprehension." Journal of Verbal Learning and Verbal Behaviour 16(3): 265-274.

Holsti, O. R. (1969). Content Analysis for the Social Sciences and Humanities. Reading, Mass., Addison-Wesley.

Hornby, P. A. (1977). Bilingualism, Psychological, Social and Educational Implications. New York, Academic Press Inc.

Impevedo, S. (1993). Fundamentals in Handwriting Recogntion. NATO Advanced Study Workshop on Fundamentals in Handwriting Recognition, Château de Bonas, France, Springer-Verlag.

Kelly, G. A. (1955). The Psychology of Personal Constructs. New York, W.W. Norton and Company Inc.

Kenny, A. (1982). The Computation of Style, An Introduction to Statistics for Students of Literature and Humanities, Pergamon Press.

Kiernan, K. S. (1991). "Digital Image Processing and the Beowulf Manuscript." Literary and Linguistic Computing 6: 20-27.

Kiraly, D. C. (1997). Think Aloud Protocols and the Construction of a Professional Translator Self-Concept. Cognitive Processes in Translation and Interpreting. J. H. Danks, G. M. Shreve, S. B. Fountain and M. K. McBeath, Sage Publications.

Krippendorff, K. (1980). Content Analysis: An Introduction to its Methodology. London, Sage Publications.

Lanterman, A. (1998). Minimum Description Length Understanding of Infrared Scenes. Automatic Target Recognition VIII, Orlando, Florida, SPIE Proc. 3371.

Lawler, J. M. and H. A. Dry, Eds. (1998). Using Computers in Linguistics, A Practical Guide. London, Routledge.

Leclerc, Y. G. (1989). "Constructing Simple Stable Descriptions for Image Partitioning." International Journal of Computer Vision 3(1): 73-102.

Levison (1965). "The Siting of Fragments." Computer Journal 7: 275 - 277.

Levison, M. (1967). The Computer in Literary Studies. Machine Translation. A. D. Booth. Amsterdam, North Holland Publishing Company: 173-194.

Levison, M. (1999). The Jigsaw Puzzle Problem Revisited. ACH/ALLC, Charlottesville, Virginia, ALLC.

Liversedge, S. P., K. B. Paterson, et al. (1998). Eye Movements and Measures of Reading Time. Eye Guidance in Reading and Scene Perception. G. Underwood. Oxford, Elsevier: 55-75.

Lundberg, M. J. (2002). "New Technologies: Reading Ancient Inscriptions in Virtual Light", West Semitic Research Project, University of Southern California. 2002. http://www.usc.edu/dept/LAS/wsrp/information/article.html

MacNeice, L. (1979). Collected Poems. London, Faber.

Mahoney, A. (2000). "Overview of Latin Syntax", Perseus Project. 2002. http://www.perseus.tufts.edu/cgi-bin/ptext?doc=1999.04.0022

Mahoney, A. and J. Rydberg-Cox (2001). "A Note on Scrabble in Latin." Classical Outlook 78(2): 58-59.

Mallon, J. (1952). Paléographie Romaine. Madrid, Consejo Superior de Investigaciones Científicas, Instituto Antonio de Nebrija, de Filología.

Manguel, A. (1997). A History of Reading, Flamingo.

Marcus, S., Ed. (1988). Automating Knowledge Acquisition for Expert Systems. Lancaster, Kluwer Academic Press.

Marichal, R. (1950). "L'Écriture Latine du Ier au VIIIe Siècle: Les Sources." Scriptorium 4: 131-3.

Marichal, R. (1955). "L'Écriture Latine du Ier au VIIIe Siècle: Les Sources." Scriptorium 9: 129-30.

Masson, J. (1985). "Felt Tip Pen Writing." Journal of Forensic Sciences 30(172-7).

Mathyer, J. (1969). "Influence of Writing Instruments on Handwriting and Signatures." Journal of Criminal Law, Criminology and Police Science 60(102-12).

McClelland, J. L. and D. E. Rumelhart (1986). A Distributed Model of Human Learning and Memory. Parallel Distributed Processing: Vol. 2. Psychological and Biological Models. D. E. Rumelhart, McClelland and T. P. R. Group. Cambridge, MA, MIT Press.

McGraw, K., L. and C. R. Westphal, Eds. (1990). Readings in Knowledge Acquisition, Current Practices and Trends. London, Ellis Horwood Limited.

McGraw, K. L. and K. Harbison-Briggs (1989). Knowledge Acquisition: Principles and Guidelines. London, Prentice-Hall International Editions.

Molton, N., X. Pan, et al. (Forthcoming (2003)). "Visual Enhancement of Incised Text." Pattern Recognition (to appear).

Morik, K., S. Wrobel, et al. (1993). Knowledge Acquisition and Machine Learning: Theory, Methods and Applications. London, Academic Press.

Morwood, J. and M. Warman (1990). Our Greek and Latin Roots, Cambridge University Press.

Oakhill, J. and A. Garnham (1988). Becoming a Skilled Reader. Oxford, Basil Blackwell Ltd.

Ogden, J. A. (1969). The Siting of Papyrus Fragments: An Experimental Application of Digital Computers. Mathematics. Glasgow, University of Glasgow. Ph.D. Thesis.

OOstdijk, N. (1988). "A Corpus Linguistic Approach to Linguistic Variation." Literary and Linguistic Computing 3(12-25).

Pan, X., M. Brady, et al. (Forthcoming (2003)). "Enhancement and Feature Extraction for Images of Incised and Ink Texts." submitted.

Parisse, C. (1996). "Global Word Shape Processing in Off-Line Recognition of Handwriting." IEEE Transactions on Pattern Analysis and Machine Intelligence 18(4): 460-464.

Plautus (1965). Pseudolus. The Pot of Gold and Other Plays. Harmondsworth, Penguin.

Plunkett, K. (1995). Connectionist Approaches to Language Acquisition. The Handbook of Child Language. P. Fletcher and B. Macwhinney. Oxford, Blackwell.

Popping, R. (2000). Computer Assisted Text Analysis. London, Sage Publications.

Preece, J., Rogers, Y. and H. Sharp (2002). Interaction Design: Beyond Human Computer Interaction. New York, Wiley.

Prescott, A. (1997). "The Electronic Beowulf and Digital Restoration." Literary and Linguistic Computing 12: 185-95.

Rayner, K. (1998). "Eye Movements in Reading and Information Processing: 20 Years of Research." Psychology Bulletin 124(3): 372-422.

Reichgelt, H. and N. Shadbolt (1992). ProtoKEW: A Knowledge-Based System for Knowledge Acquisition. Artificial Intelligence. D. Sleeman and N. Bernsen. Hove, Lawrence Erlbaum. 6.

Reiner, E. (1973). "How We Read Cuneiform Texts." Journal of Cuneiform Studies 25: 3 - 58.

Renner, T. (1992). The finds of Wooden Tablets from Campania and Dacia as Parallels to Archives of Documentary Papyri from Roman Egypt. Copenhagen Congress paper.

Rissanen, J. (1978). "Modeling by Shortest Data Description." Automatica 14:465-471.

Rissanen, J. (1999). "Hypothesis Selection and Testing by the MDL Principle." Computer Journal 42(4): 260-269.

Robertson, P. (1999). A Corpus Based Approach to the Interpretation of Aerial Images. IEE IPA99, Manchester.

Robertson, P. (2001). A Self Adaptive Architecture for Image Understanding. Department of Engineering Science. Oxford, University of Oxford. Doctor of Philosophy.

Robertson, P. and R. Laddaga (Forthcoming (2002)). Principal Component Decomposition for Automatic Context Induction. Artificial and Computational Intelligence Conference, Tokyo.

Robertson, P., M. Terras, et al. (Forthcoming (2003)). "Reading Ancient Documents: From Image to Interpretation." .

Rogers, Sharp, and Preece (2002). Interaction Design: Beyond Human Interaction. Wiley.

Saunders, L. (1999). The Uses of Greek. The Forward book of Poetry 1999. W. Sieghart: 53 - 57.

Schenk, V. U. B. (2001). Visual Identification of Fine Surface Incisions. Department of Engineering Science. Oxford, Oxford University. D.Phil Thesis.

Seales, W. B., J. Griffioen, et al. (2000). "The Digital Athenuem: New Technologies for Restoring and Preserving Old Documents." Computers In Libraries 20(2): 26-50.

Shadbolt, N. and M. A. Burton (1990). Knowledge Elicitation Techniques - Some Experimental Results. Readings in Knowledge Acquisition, current practices and trends. K. L. McGraw and C. R. Westphal: 17-23.

Sinclair (1991). Corpus, Concordance, Collocation. Oxford, Oxford University Press.

Smagorinsky, P. (1989). "The Reliability and Validity of Protocol Analysis." Written Communication 6(4): 463-477.

Smith, F. (1971). Understanding Reading: a Psycholinguistic Analysis of Reading and Learning to Read. New York, Holt, Rinehart and Winston.

Solange-Pellat (1927). Les Lois D'Ecriture. Paris.

Stainton, R. J. (1996). Philosophical Perspectives on Language. Ontario, Broadview Press.

Stark, J. A. (1992). Digital Image Processing Techniques With Applications In Restoring Ancient Manuscripts. Cambridge, Department of Engineering, University of Cambridge. EIST Project Report.

Stemler, S. (2001). "An Overview of Content Analysis." Practical Assessment, Research and Evaluation 7(17).

Stoppard, T. (1997). The Invention of Love. London, Faber and Faber.

Stubbs, M. (1996). Text and Corpus Analysis, Computer Assisted Studies of Language and Culture. Oxford, Blackwell.

Swinney, D. A. and D. T. Hakes (1976). "Effects of Prior Context Upon Lexical Access During Sentence Comprehension." Journal of Verbal Learning and Verbal Behaviour 15(6): 681-689.

Thomas, J. D. (1976). "New Light on Early Latin Writing." Scriptorium 30: 38-43.

Thomas, J. D. (1992). The Latin Writing Tablets From Vindolanda in North Britain. Bibliologia, Les tablettes á ecrire de l'antiquite á l'epoque moderne. E. Lalou, Brepols-Turnhout. 12: 204-207.

Thompson, E. M. (1912). Introduction to Greek and Latin Palaeography. Oxford, Clarendon Press.

Thoyts, E. E. (1893). How to Decipher and Study Old Documents, Being a Guide to the Reading of Ancient Manuscripts. London, Elliot Stock.

Tjader, J.-O. (1977). "Latin Palaeography, 1975-7." Eranos 75: 131-60.

Tjader, J.-O. (1979). Considerazione e proposte sulla scrittura Latina nell' età Romana. Roma, Edizioni di Storia e Letteratura.

Tjader, J.-O. (1986). "Review of Tabulae Vindolandenses I." Scriptorium 40: 297-301.

Tomlin, R. S. O. (1988). The Finds from the Sacred Springs. The Temple of Sulis Minerva at Bath, Volume 2. B. Cunliffe. Oxford, OU Committee for Archaeology. Monograph 16: 4-277.

Tomlin, R. S. O. (1994). "Vinisius to Nigra: Evidence from Oxford of Christianity in Roman Britain." ZPE(100): 93-108.

Tomlin, R. S. O. (1996). "Review Article: The Vindolanda Tablets." Britannia 27: 459-463.

Tunmer, W. E. and W. A. Hoover (1992). Cognitive and Linguistic Factors in Learning to Read. Reading Acquisition. P. B. Gough, L. C. Ehri and T. Rebecca. London, Lawrence Erlbaum Associates: 175-214.

Turner, E. G. (1968). Greek Papyri, An Introduction. Oxford, Clarendon Press.

Underwood, G. and R. Radach (1998). Eye Guidance and Visual Information Processing: Reading, Visual Search, Picture Perception and Driving. Eye Guidance in Reading and Scene Perception. G. Underwood. Oxford, Elsevier: 1-28.

Van Hoesen, H. B. (1915). Roman Cursive Writing. Princeton, Princeton University Press.

Venuti, L., Ed. (2000). The Translation Studies Reader. London, Routledge.

Wacholder, B.-Z. and M. G. Abegg (1991). A Preliminary Edition of the Unpublished Dead Sea Scrolls, etc, Fasiscle One. Washington, Biblical Archaeological Society.

Wallace, C. S. and D. M. Boulton (1968). "An Information Measure for Classification." Computer Journal 11: 185-195.

Waterman, D. A. (1986). A Guide to Expert Systems. Reading, MA, Addison-Wesley.

Weinreich, M. (1945). "YIVO and the Problems of Our Time." Yivo-Bleter 25(1): 10 - 24.

White, S. (2000). Enhancing Knowledge Acquisition with Constraint Technology. Department of Computer Science. Aberdeen, University of Aberdeen. DPhil Thesis.

Yarbus, A. L. (1967). Eye Movements and Vision. New York, Plenum Press.

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

Youtie, H. C. (1966). "Text and Context in Transcribing Papyri." GRBS 7 (1966): 251-8.

Zhu, S. C. and A. Yuille (1996). "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multi-band Image Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence 18(9): 884-900.

Zipf, G. K. (1935, Reprinted 1965). The Psycho-Biology of Language. Cambridge, MIT Press.

# APPENDIX A

# Annotation

## A.1 Encoding Scheme

The manual encoding scheme of additional tags is as follows. The tags are added, where necessary, to the "comments" field of each annotated region:

Each **character** should have (separated by comma)
- Letter Identification (*)
- Overall Letter size (S)

Each **stroke** should have (in this order, separated by a comma)
- Stroke direction (D)
- Stroke Length (L)
- Stroke Width (W)
- Place on line (P)

Each **stroke meeting** should have
- Angle (A)

These individual fields are expanded, below.

### Letter Identification (*)

Letter (*a, *b, etc)

If unidentified (*?)

If expected, but not present, (*!)

### Overall Letter Size (S)

Height (SH)
- Large (SHl)
- Average (SHa)
- Small (SHs)

Width (SW)
- Large (SWl)
- Average (SWa)
- Small (SWs)

### Direction of stroke (D)

#### Straight(DS)

Down left (DSdl)
Down right (DSdr)
Up left (DSul)

Up right (DSur)
Horizontal (DSh)
Vertical (DSv)

## Curved (DC)

Simple Curve (DCS)
Down to left (DCSdl)
Curve left (DCSdlcl)
Curve Right (DCSdlcr)
Down to right (DCSdr)
Curve left (DCSdrcl)
Curve Right (DCSdrcr)
Up to left (DCSul)
Curve left (DCSulcl)
Curve Right (DCSulcr)
Up to right (DCSur)
Curve left (DCSurcl)
Curve Right (DCSurcr)

## Double Curve (wave) (DCW)

Down to left (DCWdl)
Curve left (DCWdlcl)
Up (DCWdlclu)
Down (DCWdlcld)
Curve Right (DCWdlcr)
Up (DCWdlcru)
Down (DCWdlcrd)
Down to right (DCWdr)
Curve left (DCWdrcl)
Up (DCWdrclu)
Down (DCWdrcld)
Curve Right (DCWdrcr)
Up (DCWdrcru)
Down (DCWdrcrd)
Up to left (DCWul)
Curve left (DCWulcl)
Up (DCWulclu)
Down (DCWulcld)
Curve Right (DCWulcr)
Up (DCWulcru)
Down (DCWulcrd)
Up to right (DCWur)
Curve left (DCWurcl)
Up (DCWurclu)
Down (DCWurcld)
Curve Right (DCWurcr)
Up (DCWurcru)
Down (DCWurcru)
Horizontal (DCWh)
Curve left (DCWhcl)
Up (DCWhclu)
Down (DCWhcld)
Curve Right (DCWhcr)
Up (DCWhcru)
Down (DCWhcrd)
Vertical (DCWv)
Curve left (DCWvcl)
Up (DCWvclu)

Down (DCWvcld)
Curve Right (DCWvcr)
Up (DCWvcru)
Down (DCWvcrd)

### Loop (DL)

To left (DLl)
Open (DLlo)
Closed (DLlc)
To right (DLr)
Open (DLro)
Closed (DLrc)

## Stroke Length (L)

Comparative
Short (Ls)
Average (LA)
Long (Ll)

## Stroke Width (W)

Comparative
Thin (Wt)
Average (Wa)
Wide (Ww)

## Place on line (P)

Within line average (Pw)
Descender (PD)
Below left (PDl)
Below right (PDr)
Ascender (PA)
Above Left (PAl)
Above Right (PAr)

## Stroke Meeting Angle (A)

Open to top (AT)
Obtuse (ATo)
Right (ATr)
Acute (ATa)
Open to bottom (AB)
Obtuse (ABo)
Right (ABr)
Acute (ABa)
Open to Left (AL)
Obtuse (ALo)
Right (ALr)
Acute (ALa)
Open to Right (AR)
Obtuse (ARo)
Right (ARr)
Acute (ARa)
Crossing (AC)

Right Angle (ACr)
Compressed vertical (ACv)
Compressed horizontal (ACh)
Perpendicular (AP)

## A.2 File Format

Each annotation is preserved in an extended SGML file format. The annotation file for each image contains a single annotation tag GTAnnotations, which encapsulates all of the regions in the file, with the following attributes:

- **imageName**. The name of the source image file that this file annotates.
- **author**. The name of the last person to update the annotation file.
- **creationDate**. The date and time that the annotation file was initially created.
- **modificationDate**. The date and time that the annotation file was last modified.

Each individual region that is annotated is represented by a GTRegion tag which has the following attributes:

- **author**. The name of the person who created or last modified this region.
- **regionType**. An identifier (such as "R26") that identifies the labelling of the region. The mapping of region types to human readable names as specified in the region dictionary file: these labels are explained below (A.3).
- **regionUID**. A unique region identifier (such as "RGN93") that names the region.
- **regionDate**. The date and time that the region was created or last modified.
- **co-ordinates**. A list of pairs of numbers that represent the co-ordinates of points along the boundary of the regions. The boundary of the region is defined to be the region contained within the region formed by drawing straight lines between these points.

- **comments**. Additional tags manually added (A.1) to describe each stroke/region further.

(This text is an updated version of that found in Robertson, 2001, Appendix C.1.)

An example of a full sample SGML file is given below. This file describes the letter S, as shown in 3.7, firstly defining a character box (ADCHAR0), then the two individual strokes (SO1, SO2). Stroke ends are then identified (SE4, SE1), and the junction is then identified (SMJ3). A full description of these codes is given in A.3. Comments included in the file, below, are those which correspond to the list above (A.1).

```
<GTAnnotations imageName="C:\grava\3111.tif" author="Melissa
Terras" creationDate="09/10/02 15:23:58" modificationDate="09/10/02
15:27:30">
<GTRegion author="Melissa Terras" regionType="ADCHAR0"
regionUID="RGN0" regiondate="09/10/02 15:24:36" coordinates="338,
22, 298, 4, 186, 30, 106, 62, 94, 196, 36, 356, 46, 408, 140, 420,
184, 288, 198, 106, 282, 68, 338, 22" comments="*s, SH1,
SW1"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SO1" regionUID="RGN1"
regiondate="09/10/02 15:25:22" coordinates="170, 76, 162, 184, 152,
276, 124, 350, 102, 382, 74, 372" comments="DSd1, L1, Wa,
PD1"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SO2" regionUID="RGN2"
regiondate="09/10/02 15:26:11" coordinates="146, 94, 286, 18"
comments="DSur, La, Wa, PAr"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SE4" regionUID="RGN3"
regiondate="09/10/02 15:26:38" coordinates="62, 354, 94, 354, 94,
386, 62, 386, 62, 354"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SE1" regionUID="RGN4"
regiondate="09/10/02 15:26:44" coordinates="162, 62, 178, 62, 178,
88, 162, 88, 162, 62"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SE1" regionUID="RGN5"
regiondate="09/10/02 15:26:50" coordinates="134, 80, 154, 80, 154,
104, 134, 104, 134, 80"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SE1" regionUID="RGN6"
regiondate="09/10/02 15:26:56" coordinates="270, 14, 294, 14, 294,
36, 270, 36, 270, 14"></GTRegion>
<GTRegion author="Melissa Terras" regionType="SMJ3"
regionUID="RGN7" regiondate="09/10/02 15:27:12" coordinates="154,
70, 184, 70, 184, 98, 154, 98, 154, 70" comments="ARo"></GTRegion>
</GTAnnotations>
```

## A.3 Region Type Identifiers

Each region is given a region type identifier, to specify whether it is a type of character, stroke, end point, or junction. These identifiers are specified below.

| Identifier | Definition |
|---|---|
| ADCHAR0 | Character Box |
| ADCHAR1 | Space Character |
| ADCHAR2 | Paragraph Character |
| ADCHAR3 | Interpunct |
| SO1 | Stroke – First Stroke |
| SO2 | Stroke – Second Stroke |
| SO3 | Stroke – Third Stroke |
| SO4 | Stroke – Fourth Stroke |
| SO5 | Stroke – Fifth Stroke |
| SO6 | Stroke – Sixth Stroke |
| SO7 | Stroke – Seventh Stroke |
| SE1 | Stroke End - Blunt |
| SE2 | Stroke End – Hook – Down Left |
| SE3 | Stroke End – Hook – Down Right |
| SE4 | Stroke End – Hook – Up Left |
| SE5 | Stroke End – Hook – Up Right |
| SE6 | Stroke End – Ligature – To Left - Down |
| SE7 | Stroke End – Ligature – To Left - Up |
| SE8 | Stroke End – Ligature – To Right - Down |
| SE9 | Stroke End – Ligature – To Right - Up |
| SE10 | Stroke End – Serif – To Left - Down |
| SE11 | Stroke End – Serif – To Left - Up |
| SE12 | Stroke End – Serif – To Right - Down |
| SE13 | Stroke End – Serif – To Right - Up |
| SMJ1 | Stroke Meeting – Close Meet |
| SMJ2 | Stroke Meeting – Exact Meet |
| SMJ3 | Stroke Meeting – Cross Meet |
| SMJ4 | Stroke Meeting – Midpoint - Close Meet |
| SMJ5 | Stroke Meeting – Midpoint – Exact Meet |
| SMJ6 | Stroke Meeting – Midpoint - Cross Meet |
| SMJ7 | Stroke Meeting – Crossing |

**Table A.1:** Region Identifier Codes.

## A.4 Viewing the Annotated Corpus

The annotated corpus can be easily viewed using a Java enabled web browser (preferably Netscape Version 6.0 or above). The corpus is available on the accompanying CD-ROM in data/Chapter 3/Annotated Corpus.

To view the annotations, open the file vindolandacorpus.htm. The user is presented with a list of all 110 annotated images (the images of the documents being split into line by line sections, to allow for easier annotation. The last number in the name of the file indicates the line it represents in that document). By double clicking on one of these files, the annotated section becomes visible.

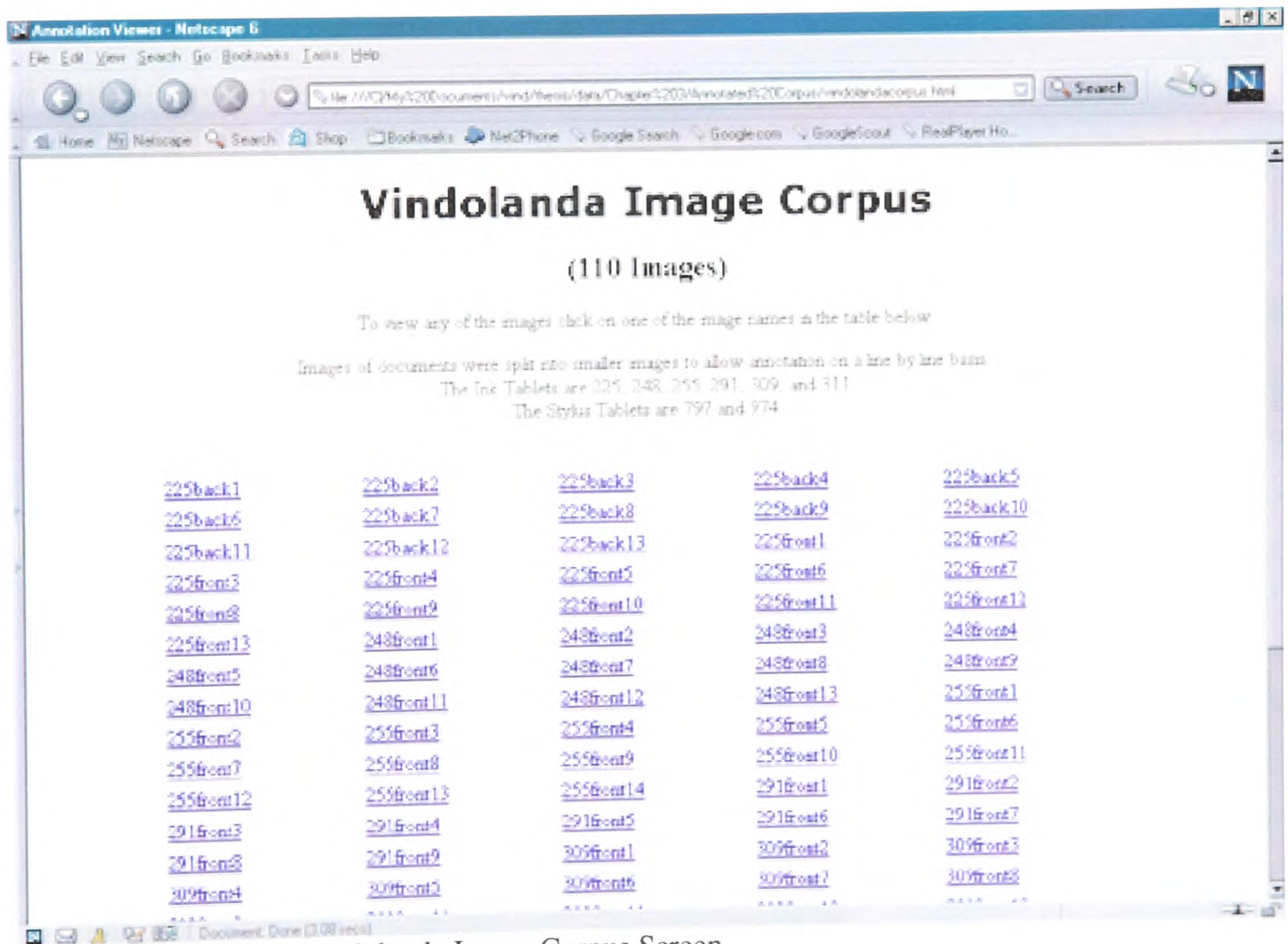



**Figure A.1:** Opening Vindolanda Image Corpus Screen.

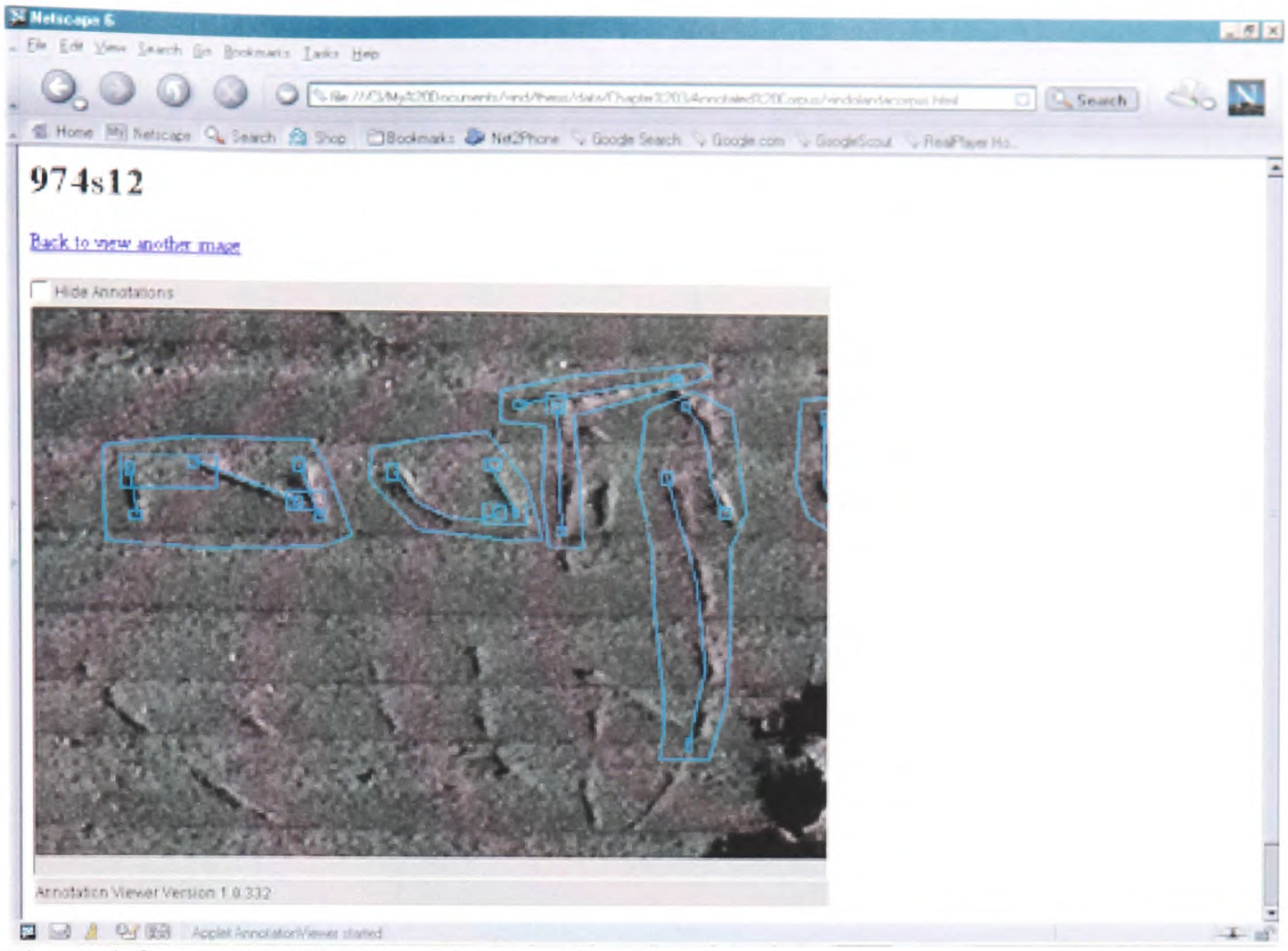



**Figure A.2:** Annotated Section of Stylus tablet 974.

It is possible to toggle the annotations on and off to compare the underlying images to the annotations.

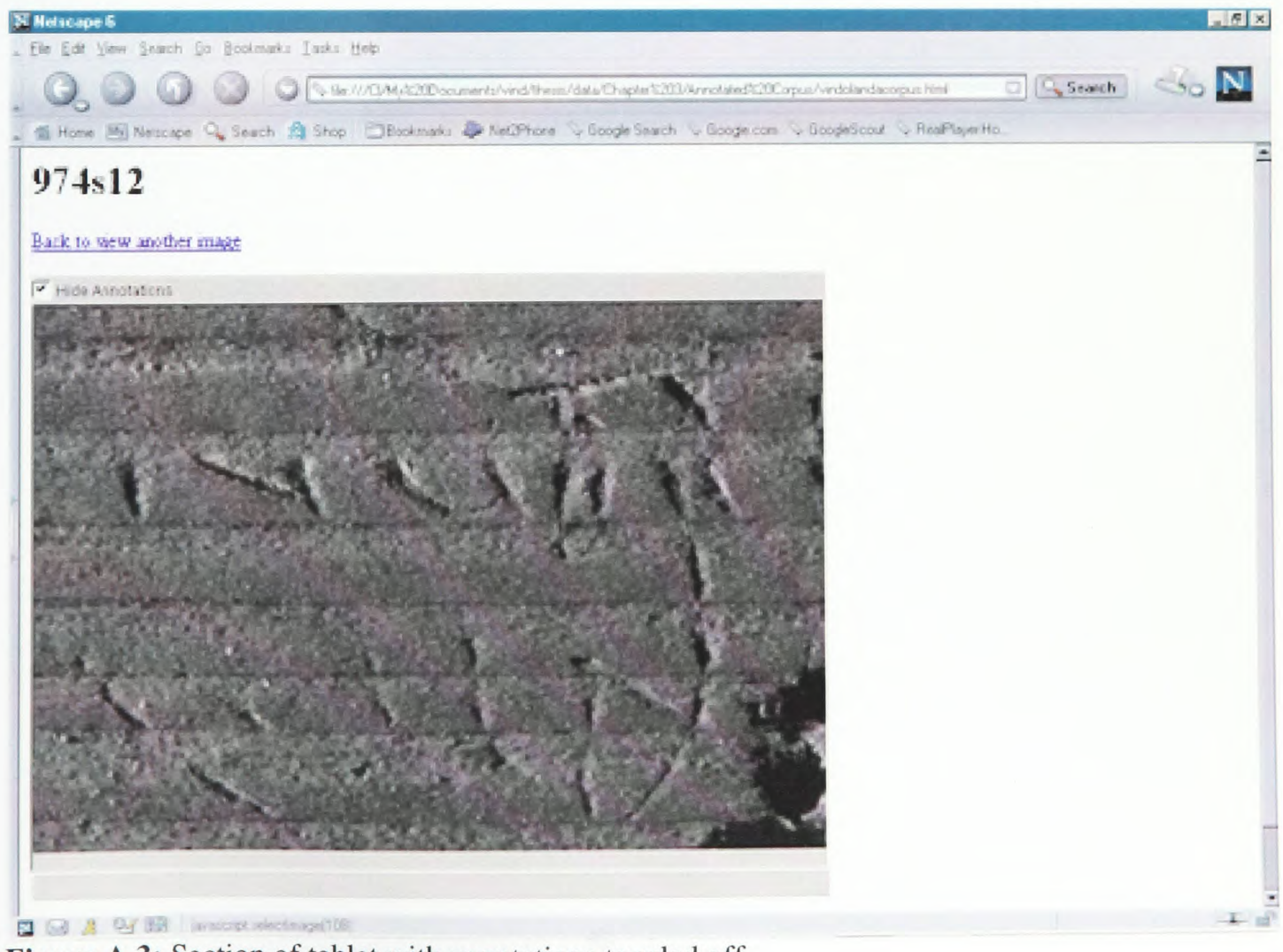



**Figure A.3:** Section of tablet with annotations toggled off.

These annotated images will eventually be presented on the web site regarding the Vindolanda texts, currently under development at the Centre for the Study of Ancient Documents, St Giles, Oxford.

# APPENDIX B

## Vindolanda Letter Form Corpus

This appendix contains a visual representation of the stroke data that was captured when annotating the images of the Vindolanda ink and stylus tablets. Each character is presented individually, first showing the standard letter form as identified by the papyrologists, then the character models that were generated from the ink and stylus tablet data, to allow comparison with the standardised forms. Finally, every instance of the characters in the data set is shown here, to give a indication of the form of the characters that were found in the documents. Of course, the complete set of annotated images can be viewed using a web browser, as described in Appendix A.

# B.1 A

**Figure B.1, 2, 3 :** The standard representation of the character A given in Bowman and Thomas (1983). The character model of the character A generated from all the ink characters in the corpus, and the character model of the character A generated from all the stylus characters in the corpus.

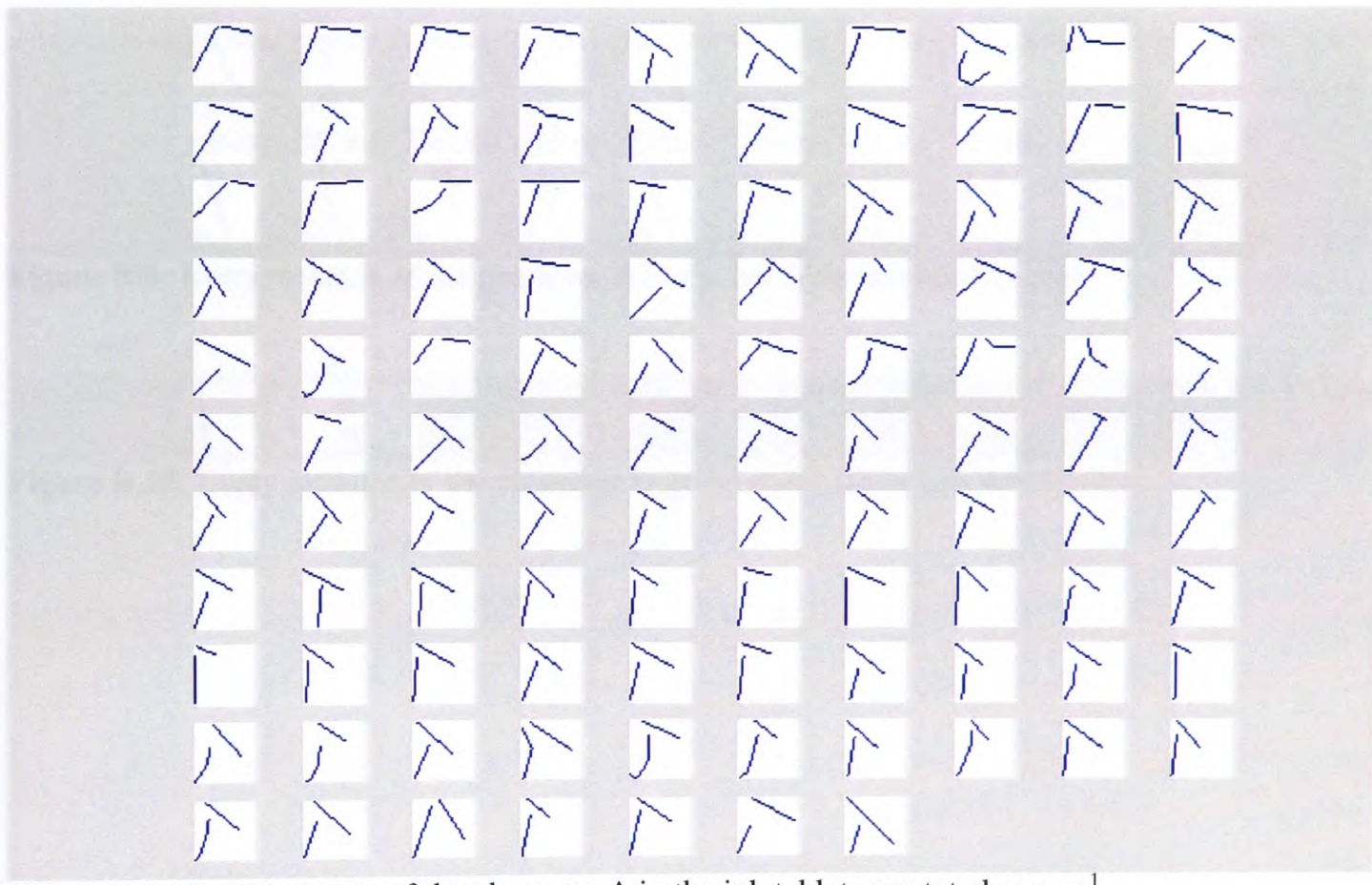


**Figure B.4:** Every instance of the character A in the ink tablet annotated corpus[1].

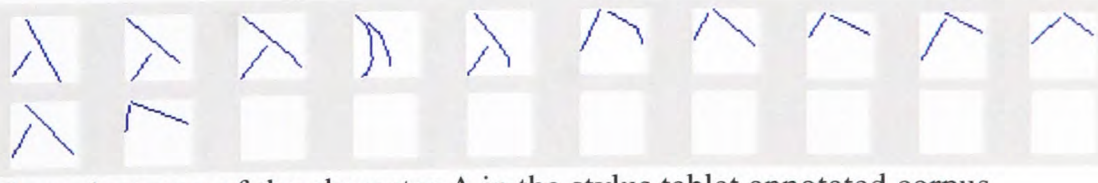


**Figure B.5:** Every instance of the character A in the stylus tablet annotated corpus.

[1] These images were created with the aid of Dr Xiabo Pan.

## B.2 B

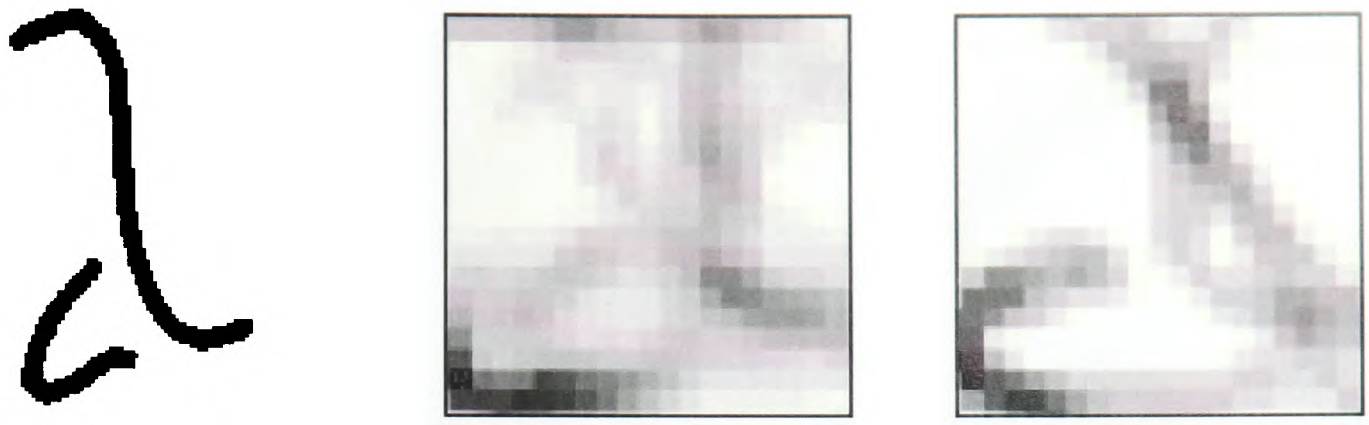


**Figure B.6, 7, 8:** The standard representation of the character B given in Bowman and Thomas (1983). The character model of the character B generated from all the ink characters in the corpus, and the character model of the character B generated from all the stylus characters in the corpus.

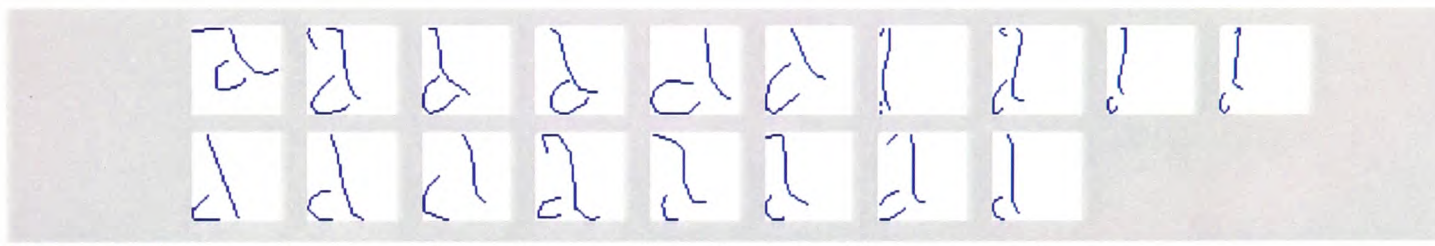


**Figure B.9:** Every instance of the character B in the ink tablet annotated corpus.

**Figure B.10:** Every instance of the character B in the stylus tablet annotated corpus.

## B.3 C

**Figure B.11, 12, 13:** The standard representation of the character C given in Bowman and Thomas (1983). The character model of the character C generated from all the ink characters in the corpus, and the character model of the character C generated from all the stylus characters in the corpus.

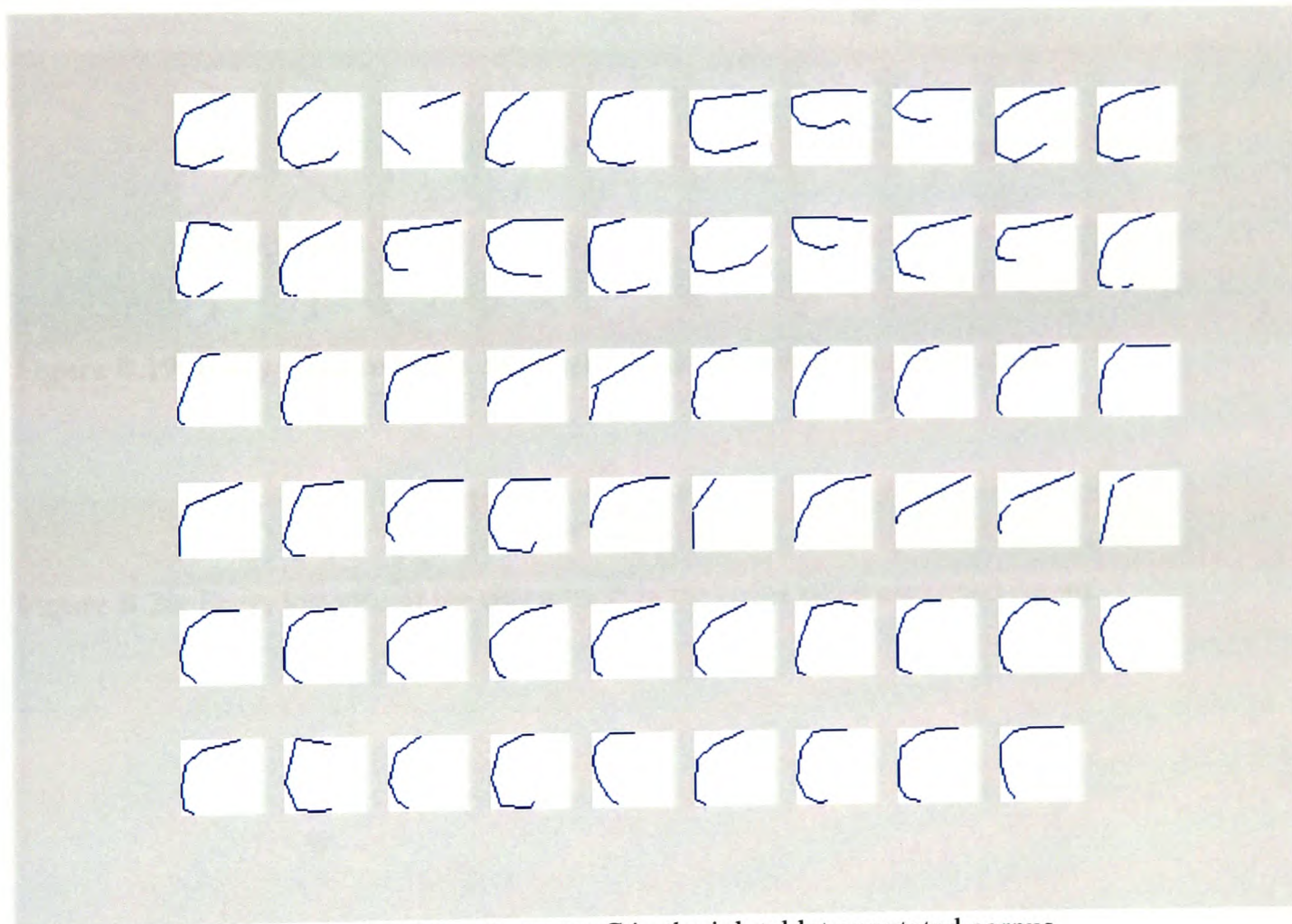


**Figure B.14:** Every instance of the character C in the ink tablet annotated corpus.

**Figure B.15:** Every instance of the character C in the stylus tablet annotated corpus.

## B.4 D

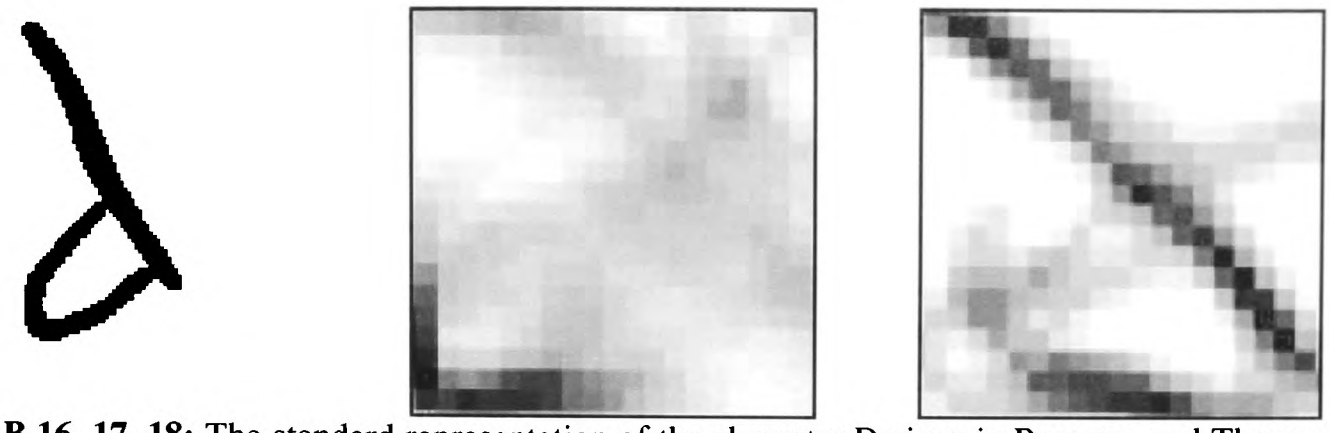


**Figure B.16, 17, 18:** The standard representation of the character D given in Bowman and Thomas (1983). The character model of the character D generated from all the ink characters in the corpus, and the character model of the character D generated from all the stylus characters in the corpus.

**Figure B.19:** Every instance of the character D in the ink tablet annotated corpus.

**Figure B.20:** Every instance of the character D in the stylus tablet annotated corpus.

## B.5 E

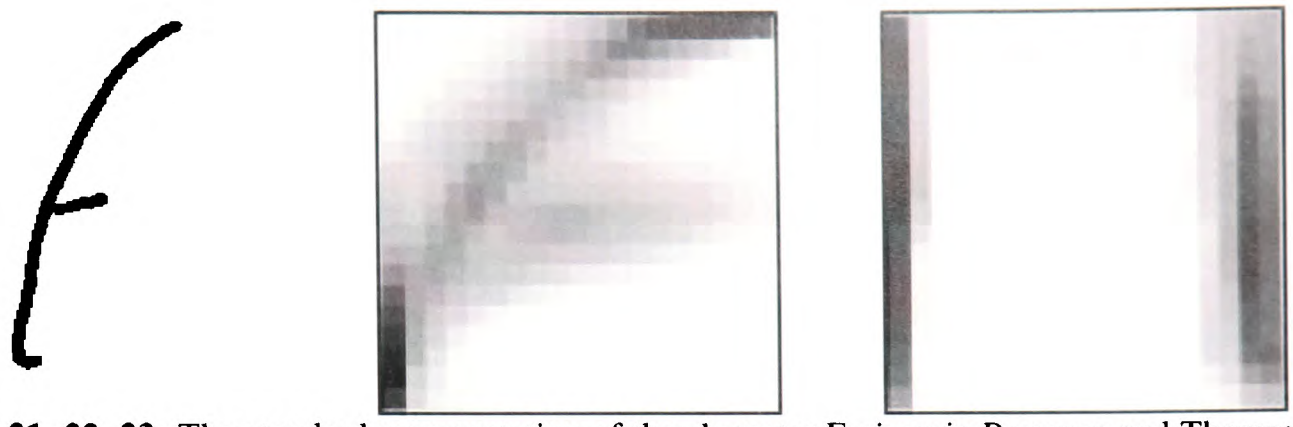


**Figure B.21, 22, 23:** The standard representation of the character E given in Bowman and Thomas (1983). The character model of the character E generated from all the ink characters in the corpus, and the character model of the character E generated from all the stylus characters in the corpus.

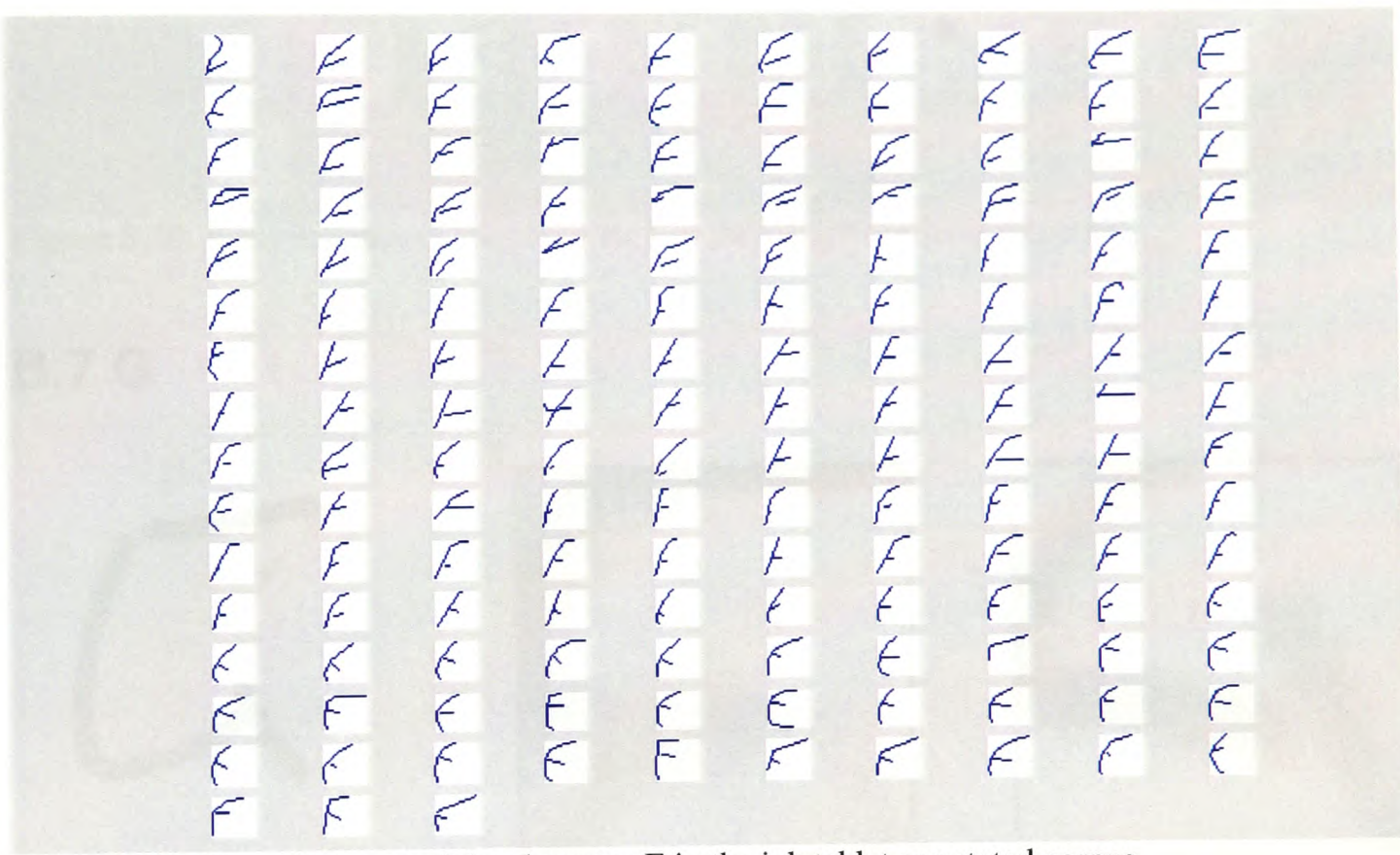


**Figure B.24:** Every instance of the character E in the ink tablet annotated corpus.

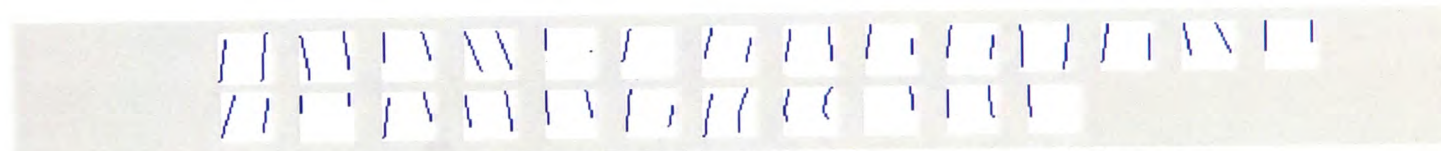


**Figure B.25:** Every instance of the character E in the stylus tablet annotated corpus.

## B.6 F

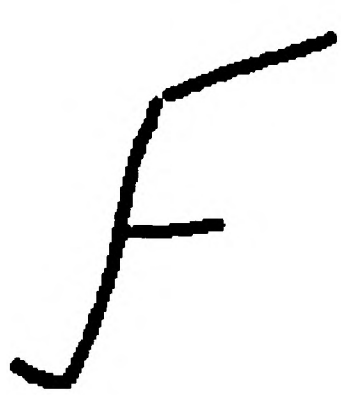


**Figure B.26, 27:** The standard representation of the character F given in Bowman and Thomas (1983). The character model of the character F generated from all the ink characters in the corpus. There were no instances of the character F in the annotated stylus tablets.

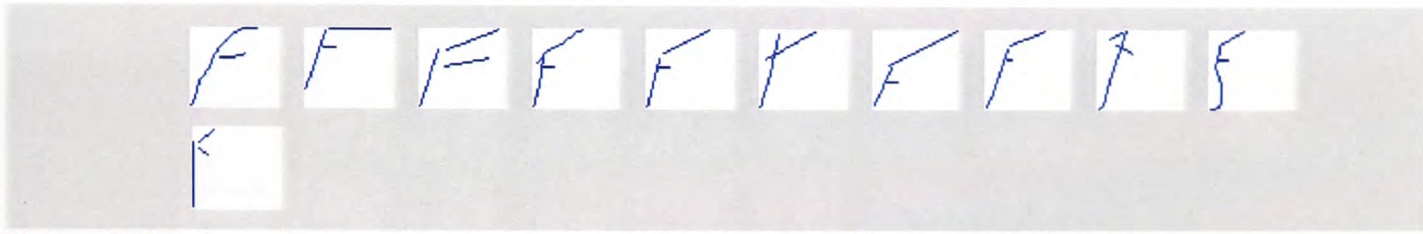
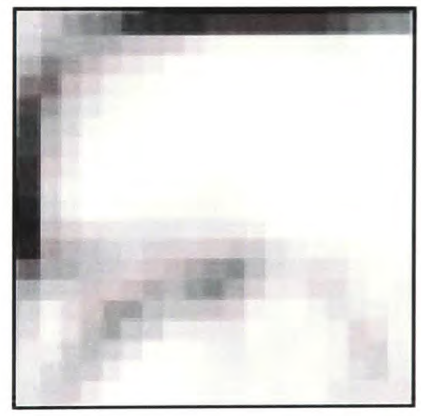


**Figure B.28:** Every instance of the character F in the ink tablet annotated corpus.

## B.7 G



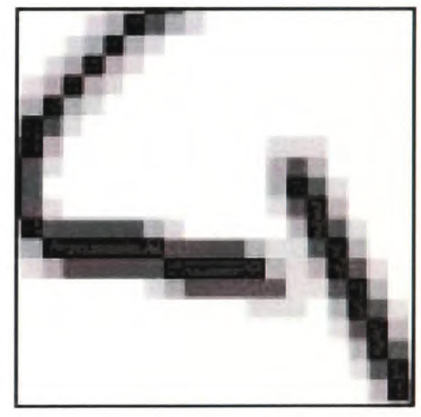


**Figure B.32:** Every instance of the character G in the ink tablet annotated corpus.

**Figure B.33:** Every instance of the character G in the stylus tablet annotated corpus.

## B.8 H

**Figure B.34, 35:** The standard representation of the character H given in Bowman and Thomas (1983). The character model of the character H generated from all the ink characters in the corpus. There were no instances of the character H in the annotated stylus tablets.

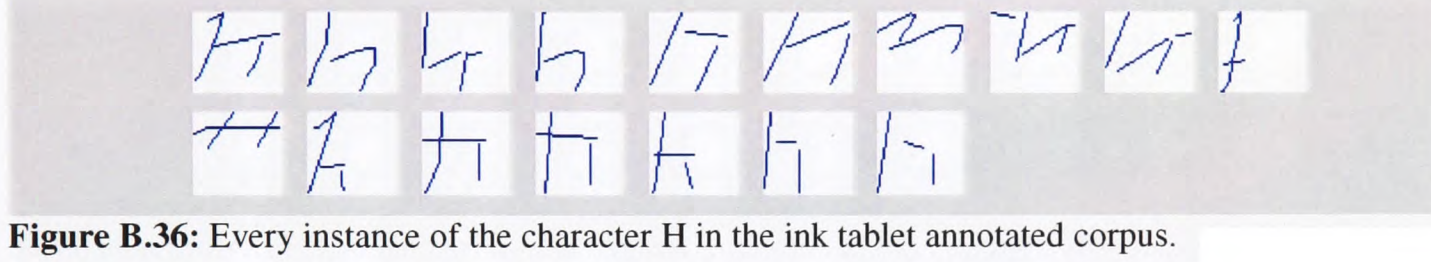


**Figure B.36:** Every instance of the character H in the ink tablet annotated corpus.

## B.9 I

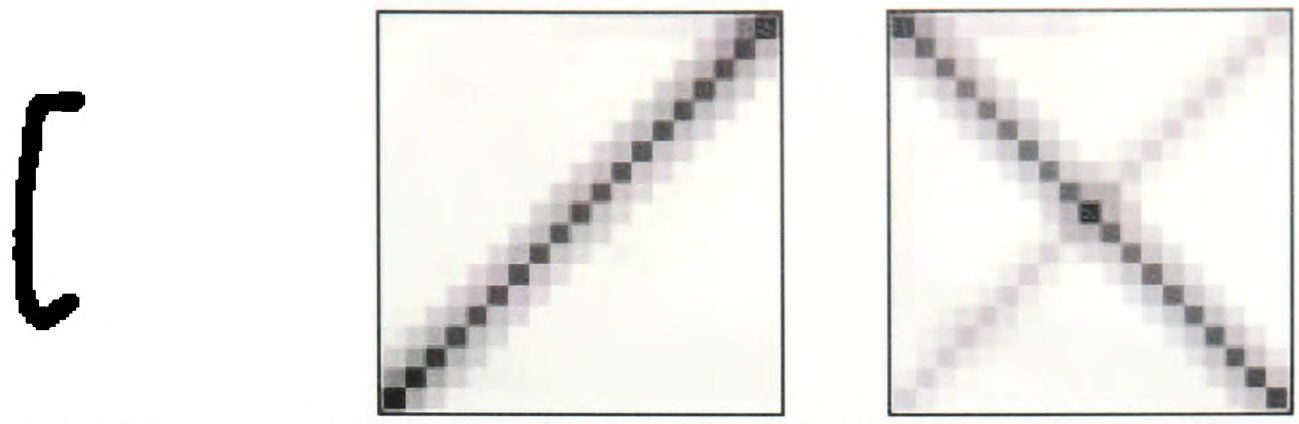


**Figure B.37, 38, 39:** The standard representation of the character I given in Bowman and Thomas (1983). The character model of the character I generated from all the ink characters in the corpus, and the character model of the character I generated from all the stylus characters in the corpus.

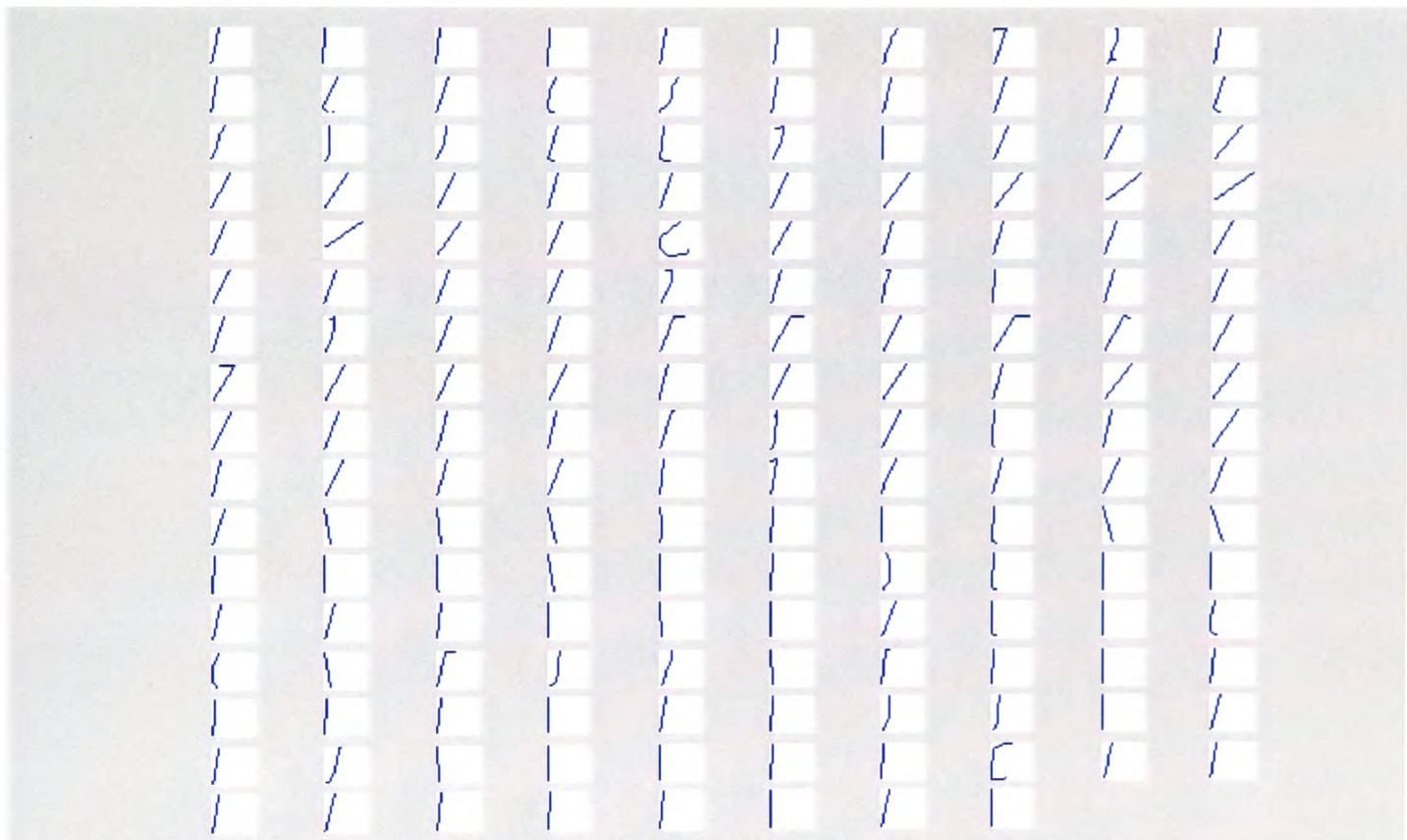


**Figure B.40:** Every instance of the character I in the ink tablet annotated corpus.

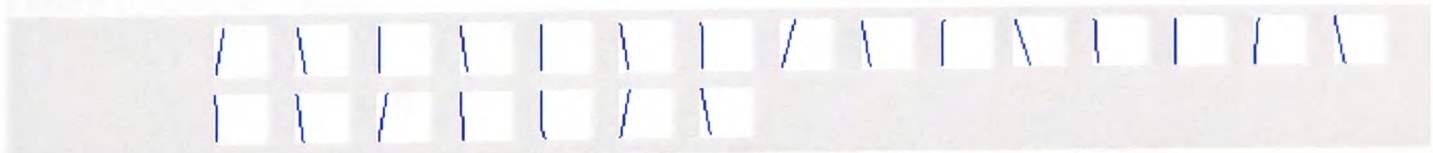


**Figure B.41:** Every instance of the character I in the stylus tablet annotated corpus.

## B.10 L

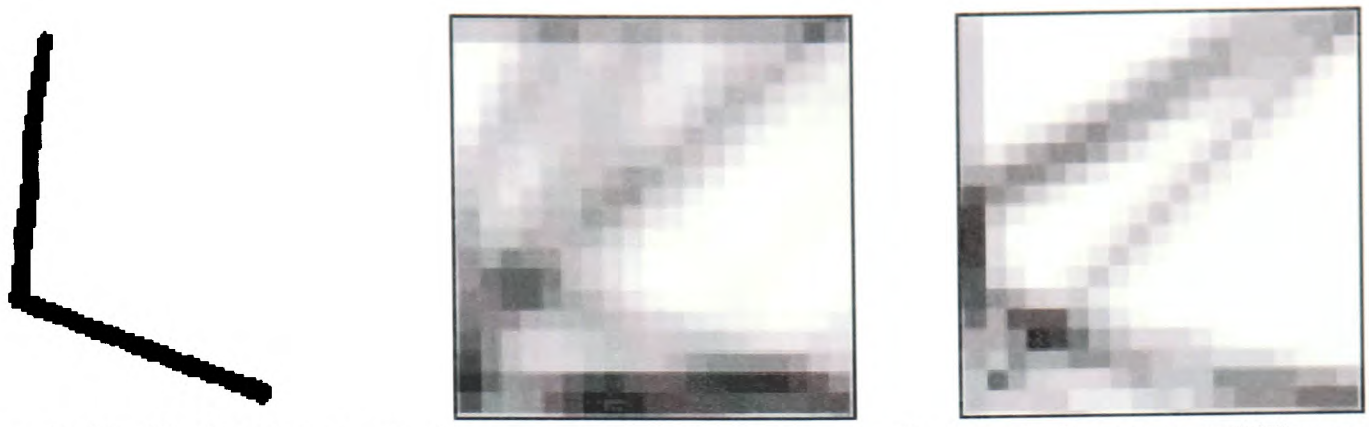


**Figure B.42, 43, 44:** The standard representation of the character L given in Bowman and Thomas (1983). The character model of the character L generated from all the ink characters in the corpus, and the character model of the character L generated from all the stylus characters in the corpus.

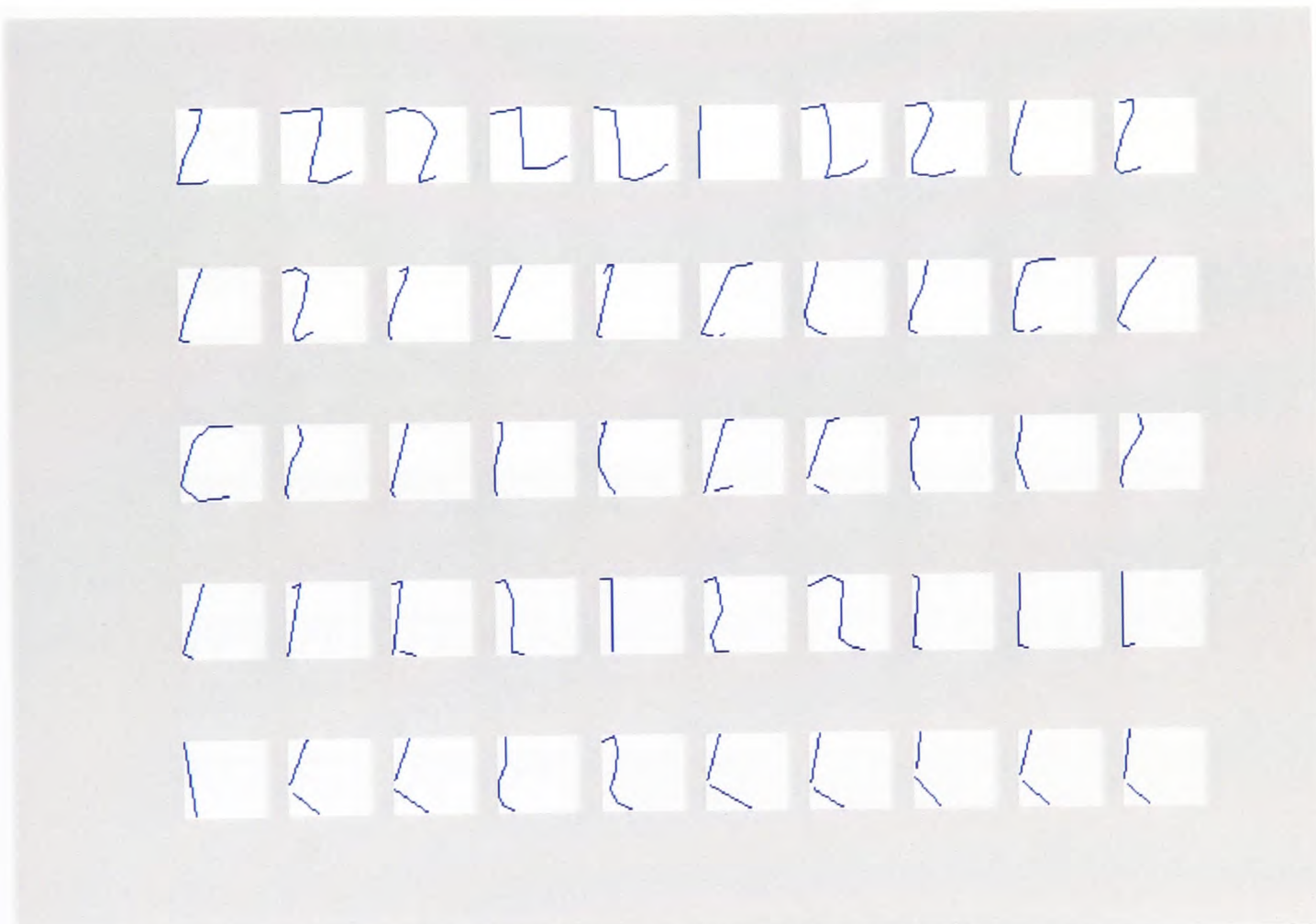


**Figure B.45:** Every instance of the character L in the ink tablet annotated corpus.

**Figure B.46:** Every instance of the character L in the stylus tablet annotated corpus.

## B.11 M

**Figure B.47, 48, 49:** The standard representation of the character M given in Bowman and Thomas (1983). The character model of the character M generated from all the ink characters in the corpus, and the character model of the character M generated from all the stylus characters in the corpus.

**Figure B.50:** Every instance of the character M in the ink tablet annotated corpus.

**Figure B.51:** Every instance of the character M in the stylus tablet annotated corpus.

## B.12 N

**Figure B.52, 53, 54:** The standard representation of the character N given in Bowman and Thomas (1983). The character model of the character N generated from all the ink characters in the corpus, and the character model of the character N generated from all the stylus characters in the corpus.

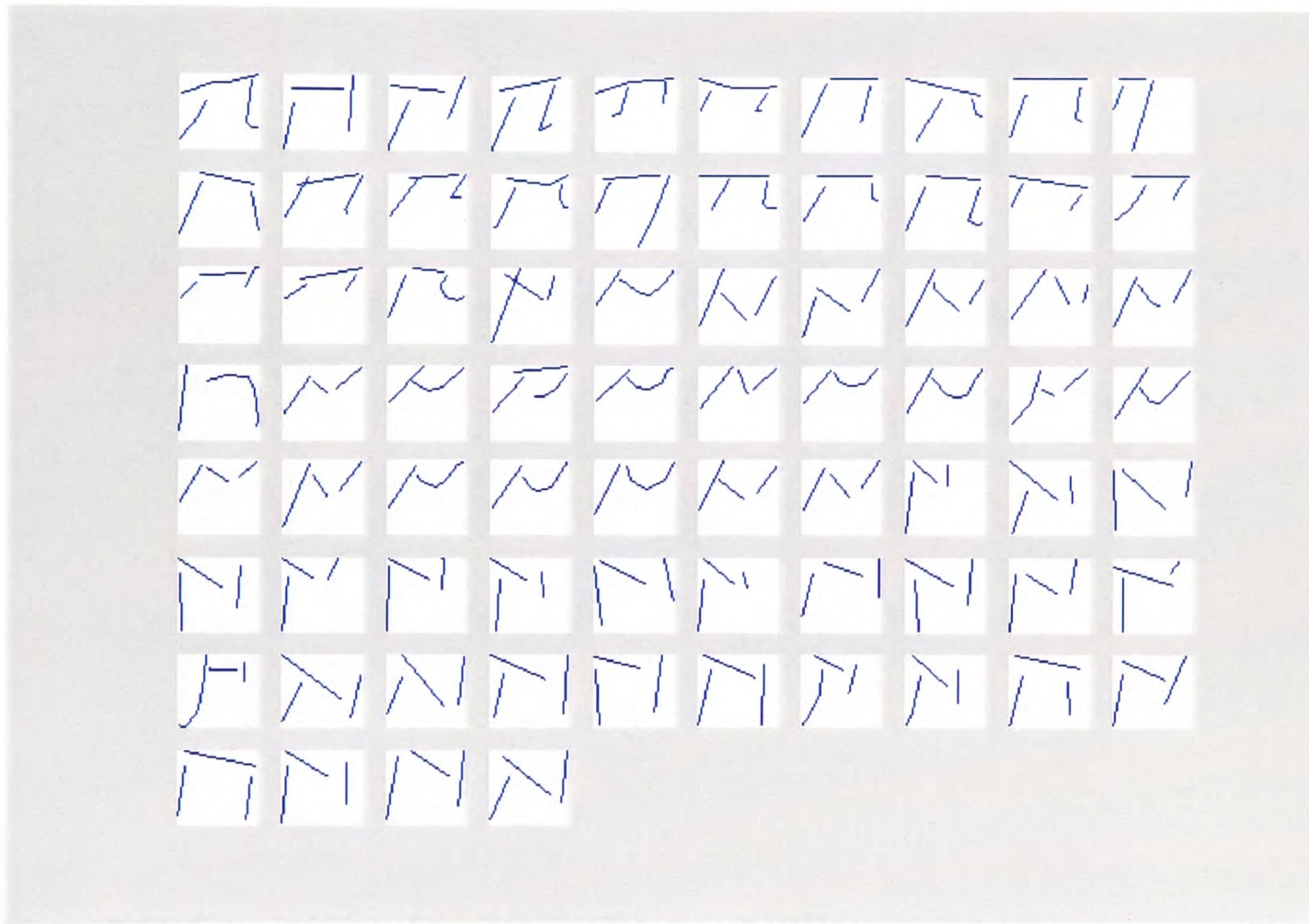


**Figure B.55:** Every instance of the character N in the ink tablet annotated corpus.

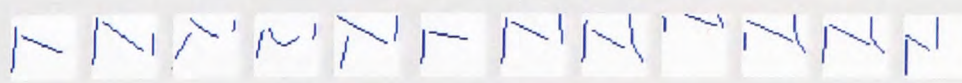


**Figure B.56:** Every instance of the character N in the stylus tablet annotated corpus.

## B.13 O

**Figure B.57, 58, 59:** The standard representation of the character O given in Bowman and Thomas (1983). The character model of the character O generated from all the ink characters in the corpus, and the character model of the character O generated from all the stylus characters in the corpus.

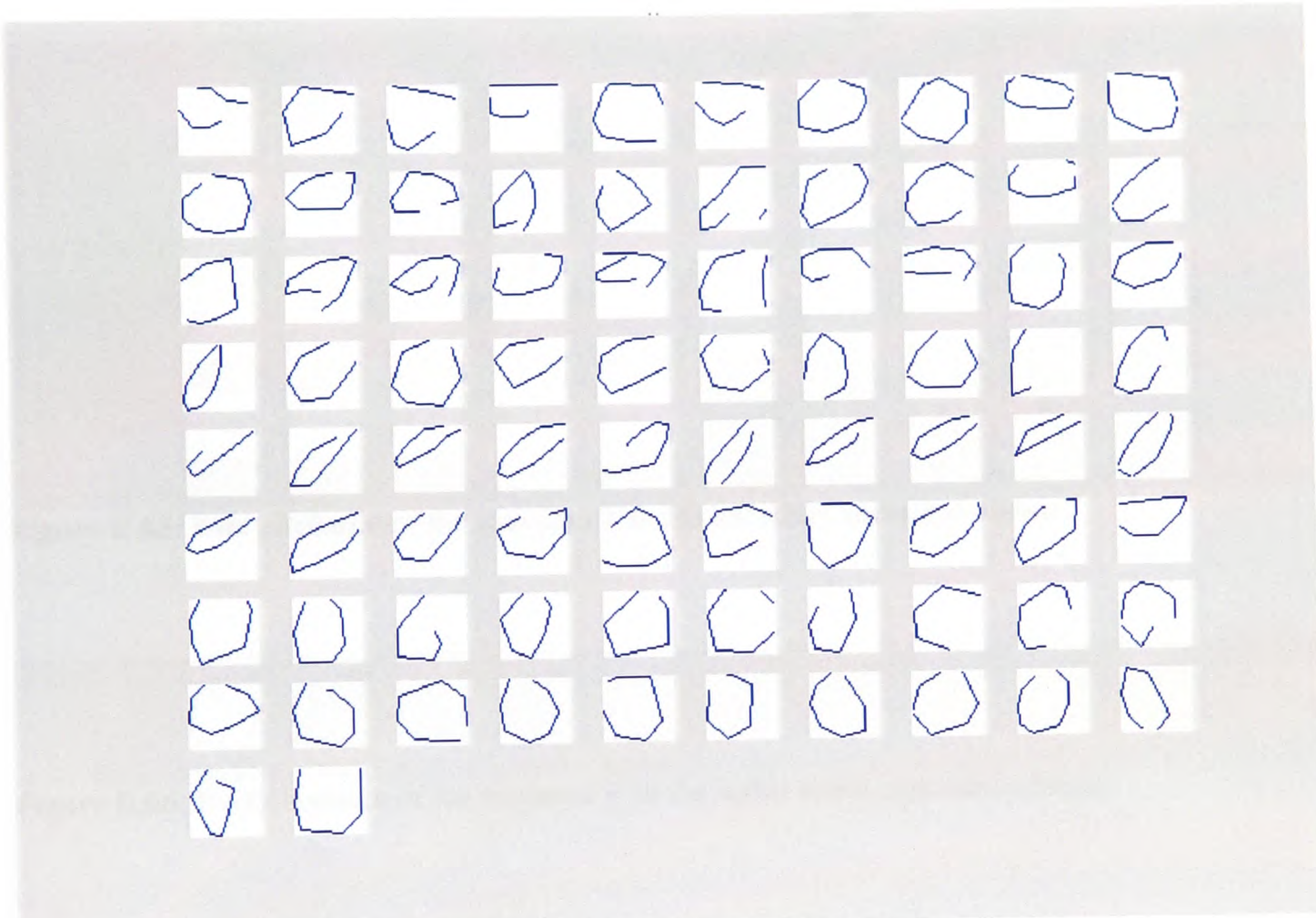


**Figure B.60:** Every instance of the character O in the ink tablet annotated corpus.

**Figure B.61:** Every instance of the character O in the stylus tablet annotated corpus.

## B.14 P

**Figure B.62, 63, 64:** The standard representation of the character P given in Bowman and Thomas (1983). The character model of the character P generated from all the ink characters in the corpus, and the character model of the character P generated from all the stylus characters in the corpus.

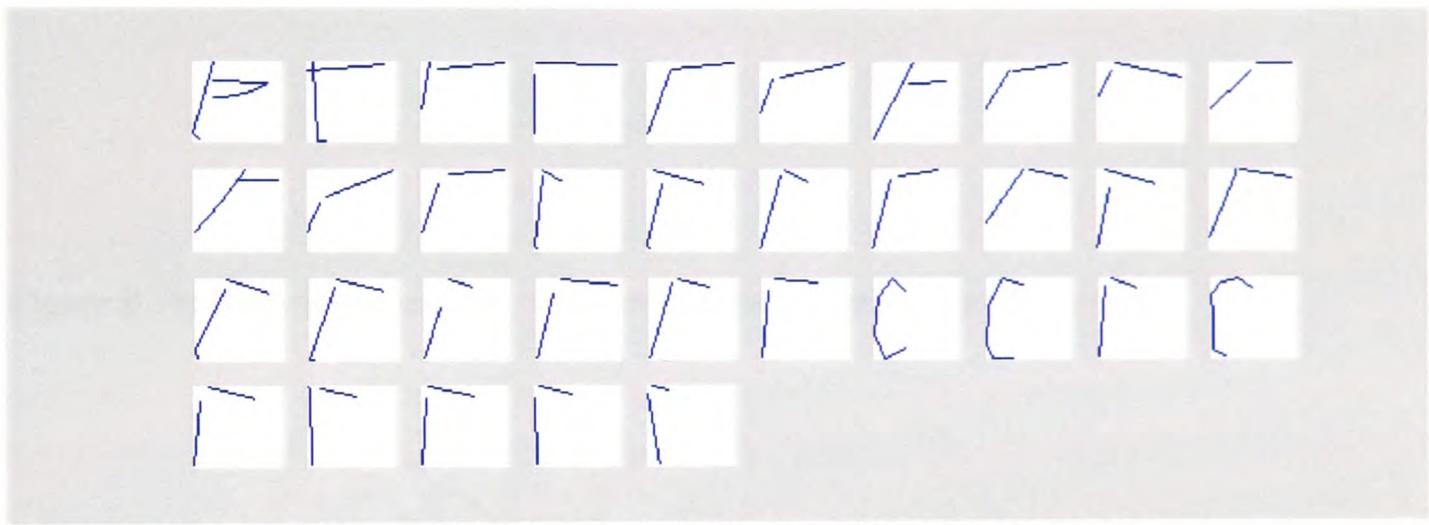


**Figure B.65:** Every instance of the character P in the ink tablet annotated corpus.

**Figure B.66:** Every instance of the character P in the stylus tablet annotated corpus.

# B.15 Q

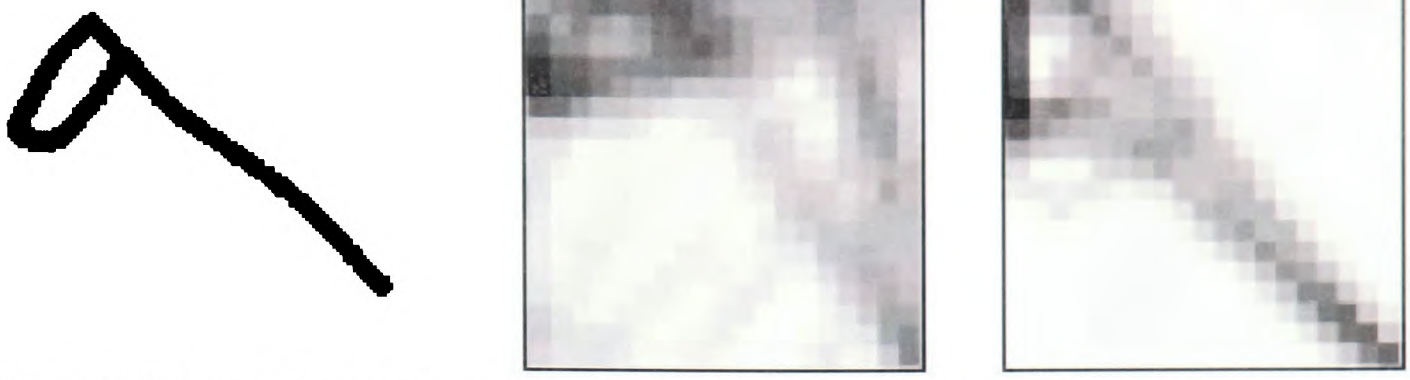


**Figure B.67, 68, 69:** The standard representation of the character Q given in Bowman and Thomas (1983). The character model of the character Q generated from all the ink characters in the corpus, and the character model of the character Q generated from all the stylus characters in the corpus.

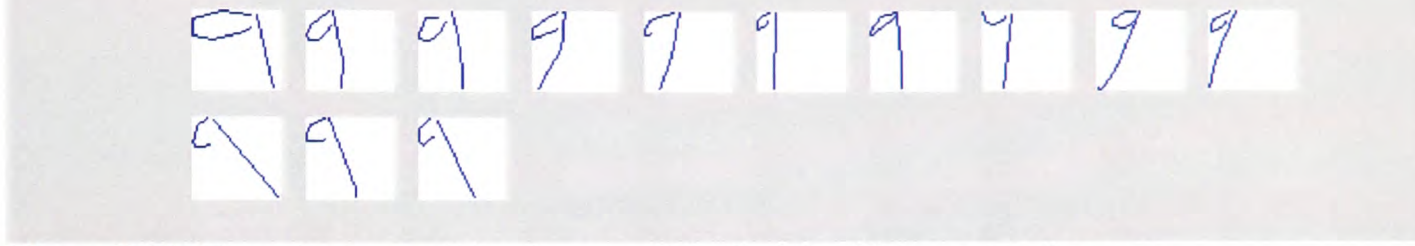


**Figure B.70:** Every instance of the character Q in the ink tablet annotated corpus.

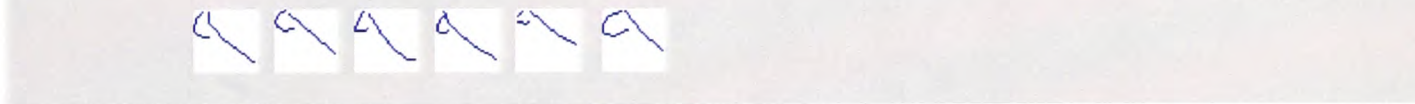


**Figure B.71:** Every instance of the character Q in the stylus tablet annotated corpus.

## B.16 R

**Figure B.72, 73, 74:** The standard representation of the character R given in Bowman and Thomas (1983). The character model of the character R generated from all the ink characters in the corpus, and the character model of the character R generated from all the stylus characters in the corpus.

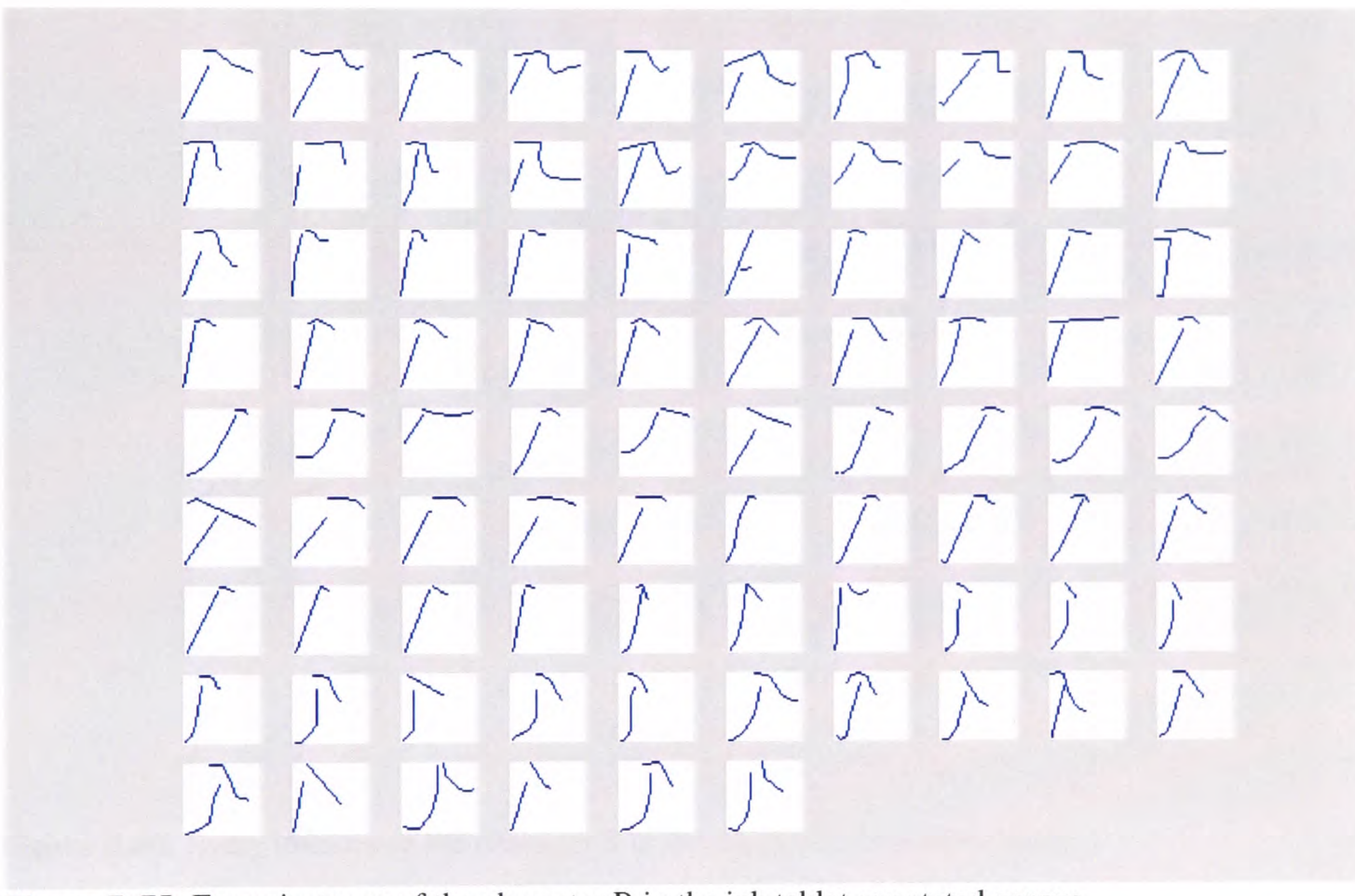


**Figure B.75:** Every instance of the character R in the ink tablet annotated corpus.

**Figure B.76:** Every instance of the character R in the stylus tablet annotated corpus.

# B.17 S

**Figure B.77, 78, 79:** The standard representation of the character S given in Bowman and Thomas (1983). The character model of the character S generated from all the ink characters in the corpus, and the character model of the character S generated from all the stylus characters in the corpus.

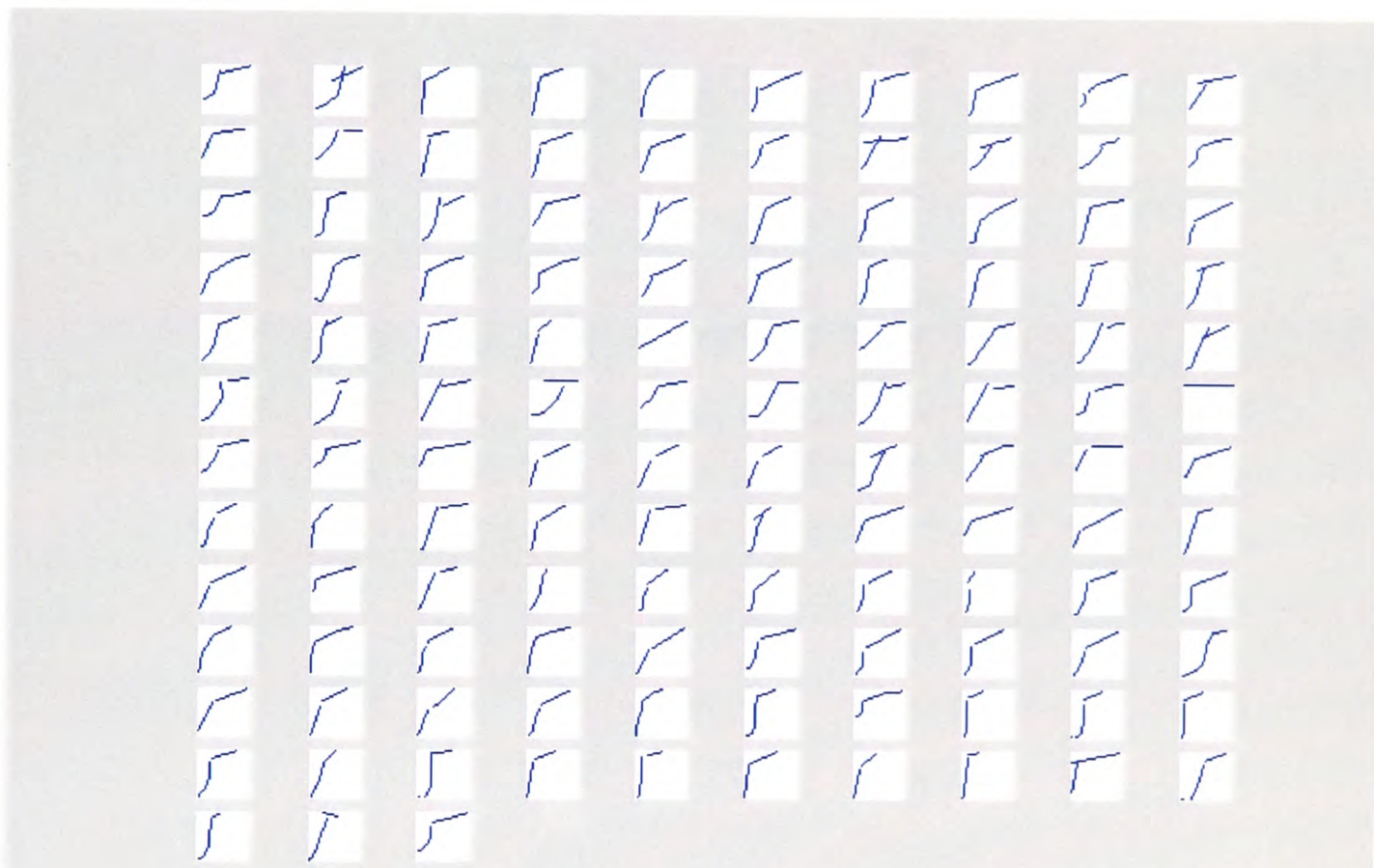


**Figure B.80:** Every instance of the character S in the ink tablet annotated corpus.

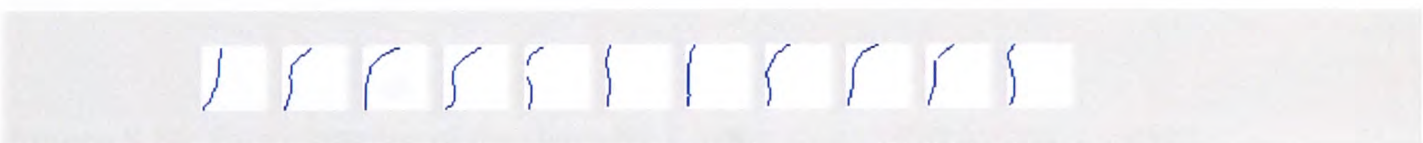


**Figure B.81:** Every instance of the character S in the stylus tablet annotated corpus.

## B.18 T

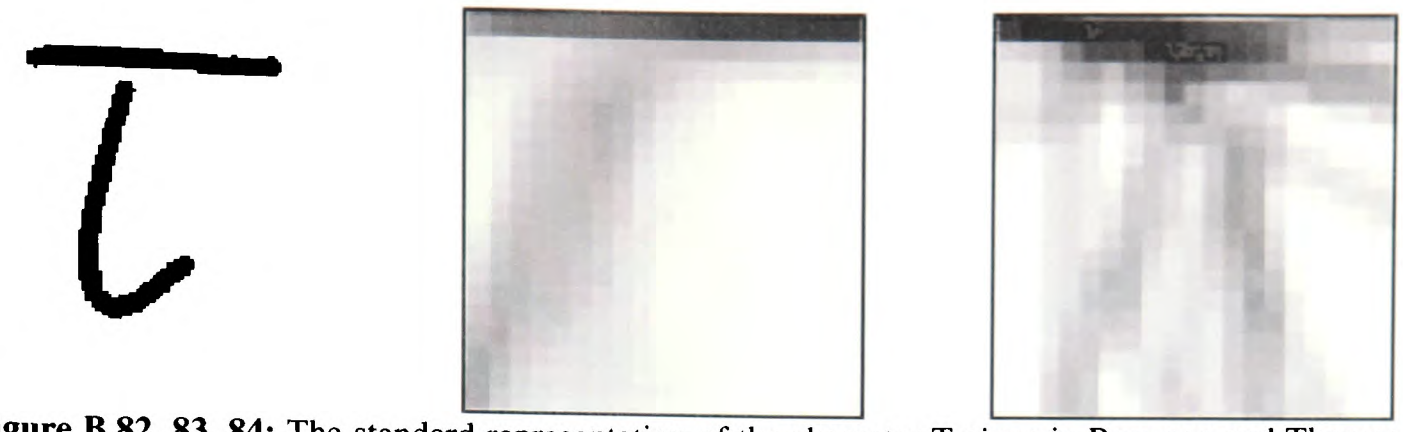


**Figure B.82, 83, 84:** The standard representation of the character T given in Bowman and Thomas (1983). The character model of the character T generated from all the ink characters in the corpus, and the character model of the character T generated from all the stylus characters in the corpus.

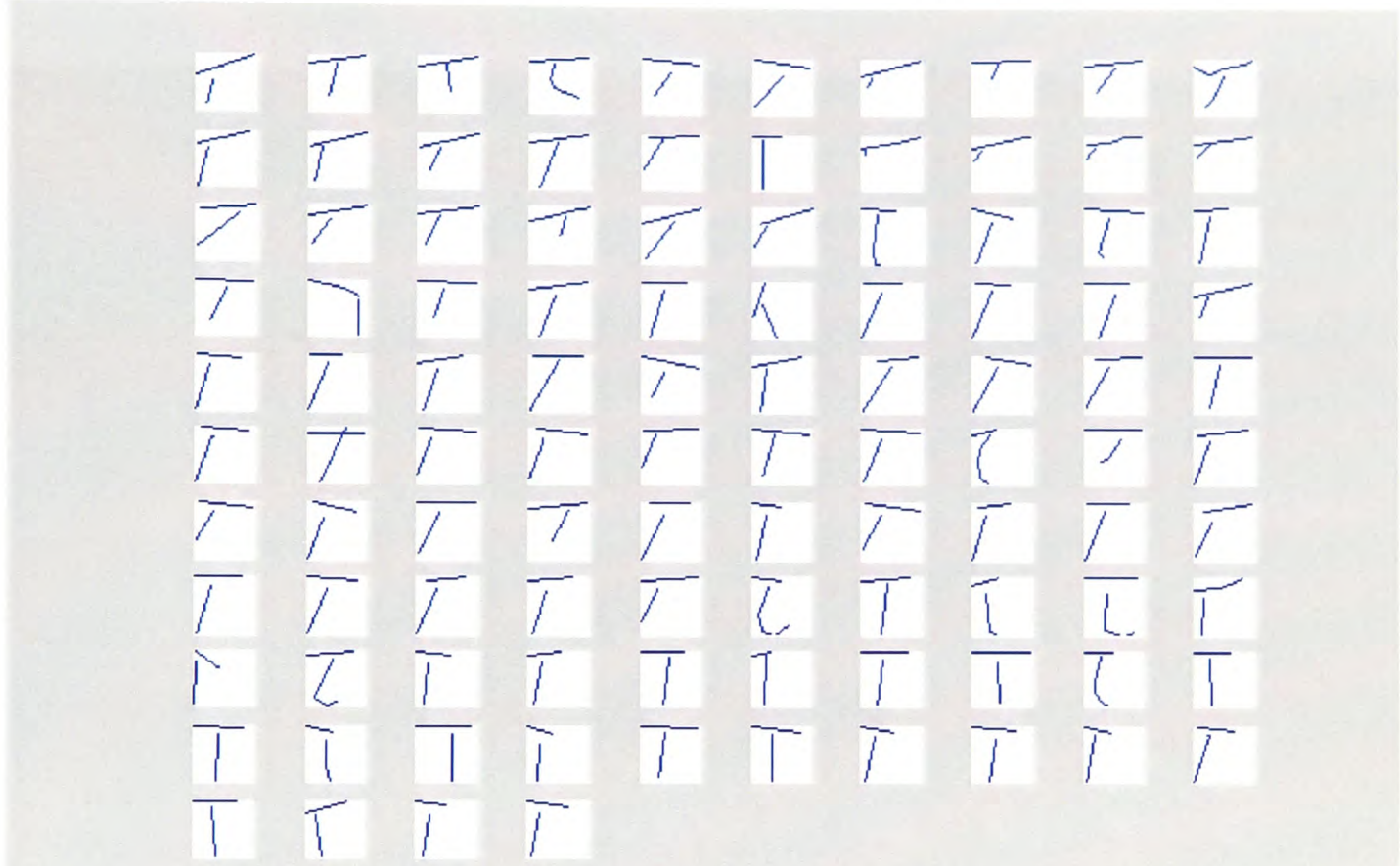


**Figure B.85:** Every instance of the character T in the ink tablet annotated corpus.

**Figure B.86:** Every instance of the character T in the stylus tablet annotated corpus.

# B.19 U[2]

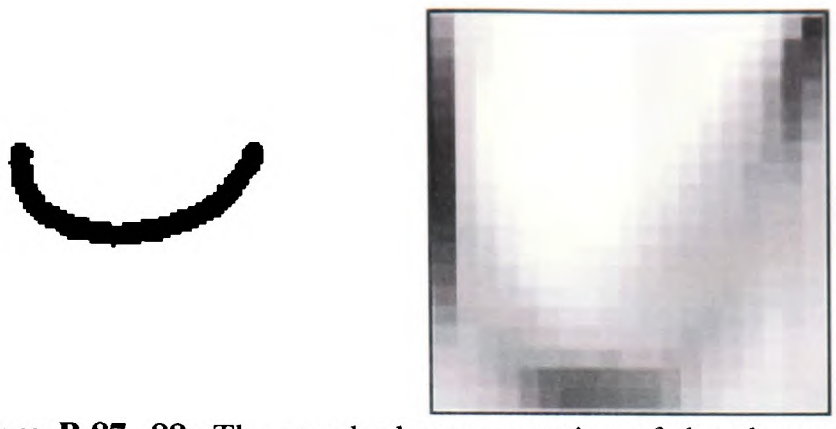


**Figure B.87, 88:** The standard representation of the character U given in Bowman and Thomas (1983). The character model of the character U generated from all the ink characters in the corpus. There were no instances of the character U in the annotated stylus tablets (that were not in the test set of data).

**Figure B.89:** Every instance of the character U in the ink tablet annotated corpus.

[2] The characters U and V are somewhat interchangeable, but are both present in the data set due to the fact that they were used by Bowman and Thomas in the Vindolanda texts, and each U or V character was annotated according to this text.

## B.20 V

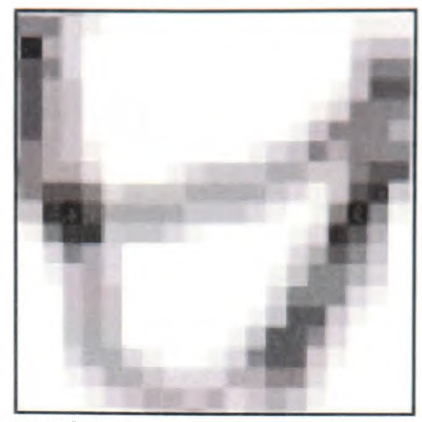


**Figure B.90, 91, 92:** The standard representation of the character V given in Bowman and Thomas (1983). The character model of the character V generated from all the ink characters in the corpus. There were no instances of the character V in the annotated stylus tablets (that were not in the test set of data).

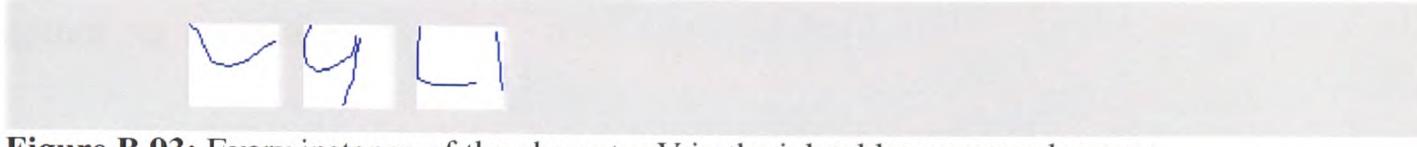
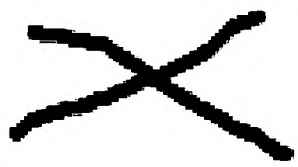


**Figure B.93:** Every instance of the character V in the ink tablet annotated corpus.

## B.21 X



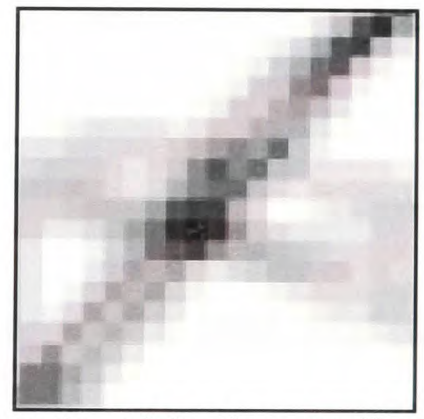


**Figure B.94, 95:** The standard representation of the character X, provided by one of the experts in the Knowledge Elicitation exercises. The character model of the character X generated from all the ink characters in the corpus. There were no instances of the character X in the annotated stylus tablets.

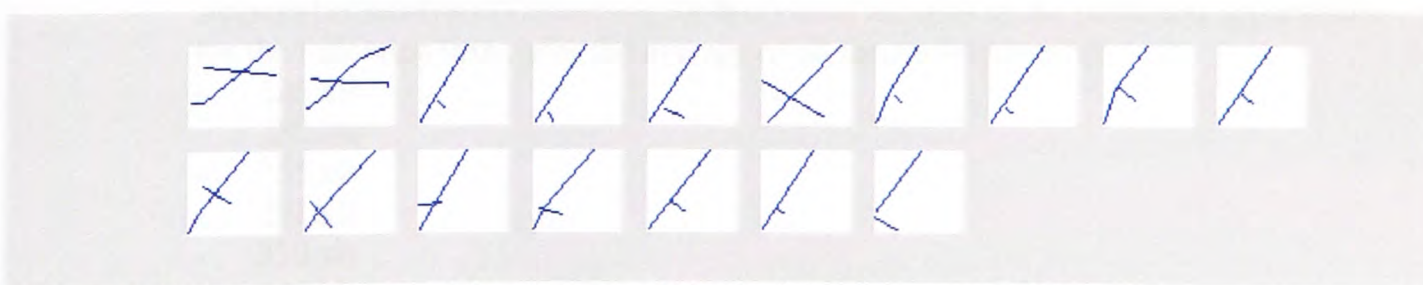


**Figure B.96:** Every instance of the character X in the ink tablet annotated corpus.

# APPENDIX C

# CDROM

## C.1 CDROM

There is an accompanying CDROM to this thesis, which contains the data sets used in the research, plus the Annotated Corpus of the Vindolanda Tablets (see Appendix A.) Should the CDROM become detached from this thesis, please contact the author for another copy.

The data on the CDROM is arranged in folders that correspond to the chapters in which the data is used or constructed. Image data is given in JPEG files. Textual and statistical data is presented in the original file formats (either Microsoft Excel '97 or Word '97).

## C.2 Contents

Folders are highlighted in bold, with a brief explanation given of the folder contents.

- **Chapter 2**

  - **Apparatus analysis** (Containing the individual analysis of the published apparatuses for the following texts, plus the average of all the analysed texts):
    - 225.xls
    - 250.xls
    - 255.xls
    - 291.xls
    - 310.xls
    - 343.xls
    - 344.xls
    - average.xls
  - **Apparatus v transcripts** (Containing a word list which compares the words present in the transcripts to those present in the published apparatuses):

- word lists.xls
- **Comparison of individuals** (Comparing the statistics regarding the individual experts from the transcripts of their conversation)
  - Comparison of use of individual words.xls
  - Speed.xls
- **Data analysis** (Collation of the statistics regarding individual experts)
  - Comparisons.xls
- **Expert B 1491** (Comparing the two readings Expert B made of 1491)
  - Expert B 1491.xls
- **Ink v Stylus** (Comparing the reading of the ink and stylus tablets)
  - Ink v stylus wordlists.xls
  - Ink v stylus reading levels.xls
- **Think Aloud Protocols** (Containing all the images and transcripts regarding the TAPs)
  - **Images** (Containing JPG images of the texts used in the TAPs)
    - 1491.jpg
    - 1543.jpg
    - 797.jpg
    - 974.jpg
  - **Ink Tablets** (Containing all the transcripts of the readings of the ink tablets)
    - **1491** (Containing all the transcripts referring to tablet 1491)
      - **Expert A** (Containing the transcript of expert A's discussion)
        - 1491A.xls
      - **Expert B** (Containing the transcript of expert B's discussions)
        - **first time** (Containing the transcript of the first discussion)
          - 1491B.xls
        - **second time** (Containing the transcript of the second discussion)
          - 1491B2.xls
      - **Expert C** (Containing the transcript of expert C's discussion)
        - 1491C.xls
    - **1543** (Containing all the transcripts referring to 1543)
      - **Expert A** (Containing the transcript of Expert A's discussion)
        - 1543A.xls
      - **Expert B** (Containing the transcript of Expert B's discussion
        - 1543B.xls
      - **Expert C** (Containing the transcript of Expert C's discussion
        - 1543C.xls
    - **Instructions to Experts** (Containing the Instructions to the Experts)
      - Instructions.doc
  - **Stylus Tablets**
    - **1593** (Containing all the transcripts referring to tablet 1593)
      - **Expert A** (Containing the transcript of expert A's discussion)
        - 1593A.xls
      - **Expert B** (Containing the transcript of expert B's discussion)
        - 1593B.xls
    - **797** (Containing all the transcripts referring to 797)
      - **Expert A** (Containing the transcript of Expert A's discussion)
        - 797A.xls
      - **Expert B** (Containing the transcript of Expert B's discussion
        - 797B.xls
    - **Vindolanda Ink Texts Corpus** (containing text of all the available ink texts)
      - Allinktexts.doc

## - Chapter 3

- **Annotated Corpus** (Containing all files regarding the corpus)
  - AnnotationViewer.jar
  - Vindolandacorpus.html
  - **Vindocorpus** (Containing all image and SGML annotations)
    - *all image files, and SGML files* (222 files)
- **images** (Containing large images of all of the annotated tablets)
  - 225back.jpg

- 225front.jpg
- 248front.jpg
- 255front.jpg
- 291front.jpg
- 309front.jpg
- 311front.jpg
- 797.jpg
- 974.jpg

- **letter forms** (Containing text regarding the letter forms derived from the sources)
  - **apparatus** (incidence of letter forms in the published commentaries)
    - *individual file for each character*
  - **ink texts transcripts** (incidence of letter forms in the ink text transcripts)
    - *individual file for each character*
  - **stylus texts transcripts** (incidence of letter forms in the stylus text transcripts)
    - *individual file for each character*
- perseus comparison (Comparing letter frequency to that of the perseus corpus)
  - perseus comparison.xls

## - Chapter 4

- **lexicostatistics** (Containing all the derived statistics from the Vindolanda text corpus)
  - bigraph.xls
  - letter frequencies.xls
  - perseus comparison.xls
  - word list.xls