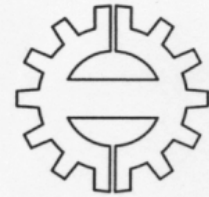Tampere University of Technology
Institute of Software Systems

David Landau

The study of old texts with the aid of digital technology: the Gothic manuscripts

The study of old texts with the aid of
digital technology: the Gothic manuscripts

David Landau

I have become interested in the Gothic manuscripts while studying the etymology of the Finnish word *juhla* 'celebration.' Since Finnish is known to be a linguistic icebox where old Germanic words which disappeared from Germanic languages are preserved, I soon found myself poring over the Gothic manuscripts. Gothic is the oldest Germanic language of which written material has survived.

However, examining the photographic rendering of the Gothic parchments is not a straightforward task. The manuscripts have been studied for more than 400 years, however, the reading of some parts of them is unreliable. This paper is the sum of knowledge and material I collected and the software I have either assembled or created to facilitate a digital deciphering and presentation of those photos. The study of the manuscripts with the aid of digital technology is only in its beginning.

David Landau

dla@cs.tut.fi
Tampere, October 9, 2001

Table of Comtents

Abstract

The advent of digital technology has opened the way for new methods in studying and presenting old texts. A digitized photo is divided into numerous elements (pixels 'picture elements') and each of them can be separately manipulated with appropriate software. As a result, the new technology offers new possibilities for filtering and singling out valuable information. The technology also facilitates easy presentation of the material with the help multimedia methods and wide distribution via the Internet and CD ROMs.

In this paper I firstly survey several projects presently conducted in the sphere of the study of old texts in various locations. Those projects are, for the most parts, in their initial stages. The amount of material to be processed is so enormous that present research is still in its pilot stage.

The lion's share of this paper is devoted to my study of the Gothic manuscripts. The idea of using digital technology for this end was proposed by professor James Marchand in an article published in 1987, and in my work I follow his ideas. The preliminary conclusion I have reached while conducting this research is that, indeed, digital technology is a very useful tool for studying and presenting the manuscripts. I have also realized that the amount of work to be done is longer than one human life time.

Key words: The Gothic language, the Gothic manuscripts, the Codex Argenteus, palimpsests, software engineering, digital image processing, multimedia

1. Introduction

One result of the spreading use of digital technology is the effort to make cultural heritage information widely available in digital form. A manifestation of this trend is the digitization of visual arts, sciences and cultural resources as well as the creation of online networks to provide access to them. The subject of this paper is how digital technology can be employed in deciphering and studying old texts with emphasis on study of the Gothic manuscripts.

The first part of the paper is a general survey of some projects currently being conducted around the world. I divide the survey into two sections; the first deals with digital study of old texts in a word processing mode and the second with texts in image mode. The former involves mainly the creation of text corpora and the latter involves digital image processing for proper deciphering and presentation.

The second part of this paper is a study of the Gothic manuscripts from a digital prospective. I describe the Gothic manuscripts and the research conducted on them from the middle of 16th century up to the present. Every generation of scholars had used the technology of its own time for deciphering and presenting the Gothic manuscripts and the paper includes a short survey of what has been done up until now.

The idea using digital technology for handling the Gothic manuscripts was conceived by Professor James W. Marchand. I became interested in the subject through one of his students, Eugene Holman, who was a teacher of mine at the University of Helsinki. In this paper I follow closely the ideas Marchand proposed in an article he published on the subject in 1987.

The next step is a survey of the Gothic manuscripts and the problems encountered by those who endeavor to make them more transparent. Once the problems are defined, I examined the possibility of using existing character recognition algorithms for better deciphering but concluded that those methods are inapplicable for this study. The research also included experimentation with existing digital filters and, again, I found out them to be of little use for this study.

Consequently, I realized that I should develop other methods and appropriate software for advancing the study of the Gothic manuscripts, and possibly other old texts, with the aid of digital technology. In the third part of this paper, I describe the methods and software I developed for this end. Included are illustrations of the work I have conducted with the methods I developed. One task was trying to restore the images of the Codex Argenteus. In this respect I believe I have succeeded; restoring one image takes around two days, which means that since there are 374 plates, they all can be restored in three to four man-years of work. The second task was to try to sort out the chaotic Ambrosian palimpsests. In this respect only the first steps were taken. The amount of work to be done is so overwhelming that any attempt to suggest any time-frame would be futile at present. I end the paper with conclusions and suggestions for further research.

_____

2. Digitizing Cultural Heritage

The tradition of preserving society's cultural heritage and enabling access to it, is as long as there has been cultural heritage. The first method, still used nowadays, is oral transmission. In the Passover Haggadah it is written that "… it would still be our duty to recount the story of the coming forth from Egypt; and all who recount at length the story of the coming forth from Egypt are verily to be praised." Indeed, every Passover eve I recount, very briefly, the story to my children.

The next step was mounting the words on some material like stone, clay, sheepskin, or papyrus. Whenever the means of preservation became too fragile and the text it contained was considered to be worthy of transmitting to the next generations, it was copied into another medium. When printing was invented, the process had been streamlined and the amount of material, which could have been preserved, grew considerably. Another innovation, photography, still enlarged the scope of material being preserved.

One theme that comes often forward when the subject of transmitting cultural heritage is being considered, is how to preserve the authenticity of the original item; in other words, how to remain as faithful to the original as possible, when transmitting it into another medium. In the case of the Gothic text, for the first two hundreds years scholars had tried, with various degrees of success, to reproduce the texts in fonts based (not accurately) on the original letters. Starting from the beginning of the 19th century, the original text has been produced as a transliteration in Latin script.

One person who was not content with this state of events was Sydney Fairbanks. As a "fruit of experiments in teaching the scripts of the great *Codex Argenteus* of Uppsala to students," he became convinced that "the monument of Gothic should be printed in Gothic letters and not transliterated into Latin characters." Following this theme he wrote (1940: 324):

> Modern scholarship should in this matter return to the sound practice of past generations. At its best, as in the Codex Argenteus, the Gothic-Wulfilan alphabet is an important and beautiful cultural heritage of the Germanic peoples. Its structure emphasizes many points in the history of the Goths themselves: the

_____

preponderantly Greek background of the alphabet looks back to early Gothic affiliations with the Eastern Empire; the runic element associates it with the first Germanic writing of any sort; and those letters which are derived from the Latin alphabet recall graphically the victorious passage of the Goths from a Greek to a Roman sphere. Transliteration into Latin letters, however exact and unambiguous this may be phonetically, reduces printed Gothic to a colorless, raceless thing from which significant national and cultural values have been erased.

Adding (p 326):

Indeed, there is a certain grim irony in rendering the literary remains of the greatest of the early Germanic peoples in an alphabet of people and church in conflict with which Gothic culture perished.

However, to these days scholars have been studying the Gothic text mostly in its transliterated form. Although the facsimile edition, done in 1927 (see more details in section 4.6.), is indeed a masterpiece, in numerous places the original text is illegible and reading the Gothic text directly from it is, to certain extent, impractical

The idea of using digital technology for making the Gothic text more accessible was conceived by Professor James W. Marchand. In his article (1987), which is the basis of my work, he writes (p.25):

Our students unfortunately usually read Gothic in transliterated form, since it is so expensive to typeset it, given the fact that the fonts can only be used on those rare occasions when one prints Gothic. It is as if you taught Russian or Greek in Latin letters. Early scholars on Gothic tried to cut or cast their own fonts, with varying degrees of success. What Christopher J. Meyer and I decided to do was a radical departure from 20th century tradition, namely to teach students to read Gothic in the Gothic script. If we wanted to imitate early workers, we could do much better than they, but that would merely be a modernization and one font. We decided early on to put our Gothic out in the original script.

However, except one page of the Codex Argenteus, posted in the Internet by Marchand in 1991 (http://ccat.sas.upenn.edu/jod/Picts/luke1cl2.gif), to the best of my knowledge, no other related work has been published since.

The same theme is expressed also elsewhere. Since 1985, a group of IBM's employees have been working on several projects, which aim at enabling worldwide access to cultural heritage (Gladney et al. 1998). Paying special attention to quality representation and intellectual property rights, the authors maintain that "a most interesting possibility raised by the digitization of antiquity is teaching university undergraduates the value and pleasure of working with original sources."

One more reason to work with the source material is the possibility of discovering details that were lost in the process of reproduction of the text in another medium. Eventually, any transliteration of old texts involves editing, no matter how one strives to be faithful to the original. Some information is lost either because it cannot be rendered properly in the target medium, or because the transcriber(s) did not notice certain details. In antiquity, even simple copying of texts was not so simple. It is well known that almost all texts of antiquity were altered during the transmission process, because the scribes who copied them made mistakes, 'corrected' earlier scribes' mistakes, or simply altered the text for one reason or another. In the case of the Gothic manuscripts, there was at least one case of attempted forgery (see below 4.3.).

Working with a digital version is not the same as studying the original image; after all, virtual reality is never identical with the real world. However, in many cases the original piece is, for one reason or another, unavailable and compromises must be done. The key for a successful digital reproduction is faithfulness. In my work, I will try to stick to this theme as much as possible.

3. Digitizing Old Texts

The idea of using computers for linguistic studies came up quite early during the history of digital technology. One of the first programs to be used was Word Cruncher, which originally was developed at Brigham Young University as a tool for easy retrieval of words and expressions from the Bible. In those days, with the spirit of digital brotherhood still supreme, the creators donated the program to the public domain, and linguists adopted it to their use. The program, which I also used extensively, creates index of words in a given text, with links to the text itself.

With the development of faster computers with bigger storage capabilities, it became possible to digitize pictures in a high quality mode. One way to prepare a digitized photo is to film it on a microfilm and, subsequently, scan the microfilm. Another way is to lay the original directly on the scanner and let the software do the rest of the work. In my work I use both systems: I scanned the facsimile edition (1927)of the Codex Argenteus (around 800 photos) into 4 CD ROMs. Other material which I do not have direct access to, I try to have photos or slides prepared for me and then scan them.

3.1. Old Text in Image Mode

For a text in a digital picture to be legible and useful, the text itself must be very clear and the picture of very high quality, that is, high density of pixels per inch. In photo-handling software one can enlarge the picture or manipulate it to a certain degree, but if the photo is in a poor quality, then not much can be done to improve it.

In priciple, digital photos containing cultural heritage can be distribted either on hard storage devices, like discettes, CD-ROMs, etc, or on-line, via the a local intranet or the Internet.

In this section I describe several projects done in this area of research. One feature common to those projects is that each of them covers only a small fraction of the collection of which it is part of. I would imagine that the best pieces were chosen for display.

3.1.1. The IBM Team

As mentioned before, a group of IBM's employees have been working on several projects since 1985 (Gladney et al. 1998). Here are some projects undertaken by the group:

- **El Archivo General de Indias, Sevilla (AGI)**, Spain. The archive houses 43,000 bundles with 86 millions pages, the most complete documentation of the Spanish administration in the Americas, from Christopher Columbus until the end of the 19th century. Until 1992, the team collected 9 million digital-image pages. Original documents, some of them being quite fragile, were scanned to provide fast access to them.

The authors estimate that between 30% and 40% of the original documents pose legibility problems, due mainly to their great age and rough handling. Damage includes faded ink, stains, and seepage of ink from the reverse side of documents. They indicate that sometimes such damage "make it extremely difficult for a scholar to read the document." To solve the problem, the team investigated procedures involving information filtering. In addition, the end users can improve the output document. One method is zooming, which enables closer look at certain details. The second available tool enables modifying the color palette in order to reduce the effect of stain ink fading, and bleed-through. In addition, assorted nonlinear spatial filtering, which can be performed on any area of a displayed page or a selected portion, are available and enable intensification of faded ink or the removal of distracting background (Figure 3.1).
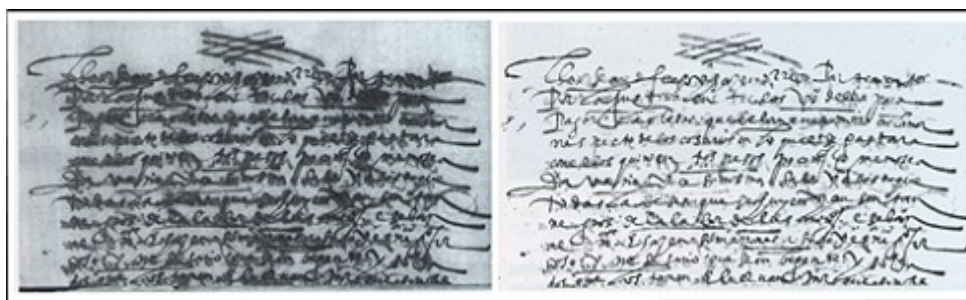

Figure 3.1: An example of ink bleed-through reduction and stain removal

A thorough discussion concerning this project: *Computerization of the Archivo General de Indias: Strategies and Results* written by Pedro González, is available at: http://www.clir.org/pubs/reports/gonzalez/contents.html.

- **The Vatican Library**. The library consists of 150,000 books and manuscripts which include early copies of Aristotle, Euclid, Homer and Virgil. The IBM team investigated the practicality of an Internet digital library service, which would allow broader access to the collection, while protecting the Vatican Library's assets. In order to find out which are the best solutions, the team digitized a significant number of manuscripts, making them available via the Internet, and collected the views of participating scholars. Among the specification provided by the library stuff was ensuring the safeguard of the material, permitting inspection of higher-resolution versions and protecting the images against unauthorized republication.

To solve the first requirement, the team developed a protective scanning environment, avoiding ultraviolet light damage. To comply with the second requirement, the images were scanned at 2600 X 3000 pixels with 24b/pixel for color pages and 8b/pixel for monochrome pages. Storage space for such high-resolution image often exceeds 20MB. In order to prepare the images for Internet service, the images were reduced to 1000 X 1000 pixels and compressed to data sizes of 150KB to 250KB. As for the third demand, the idea of offering online only low-resolution versions of little use to would-be pirates, was rejected because it did not support research in art history. Instead, the team developed visible watermarking, which indicates the owning institution, does not remove details needed by scholars, and is difficult enough to remove. A reference to a more elaborate article written by the IBM team: *Toward on-line, worldwide access to Vatican Library materials* can be found at:
http://www.research.ibm.com/journal/rd/mintz/mintzaut.html.

- **Klau Library of Hebrew Union College**. The collection consists of nearly 750,000 volumes of Hebraica and Judaica from the 10th century to present, including Biblical codices, communal records, legal documents and scientific tracts. The work on this project involves the same technology used at the Vatican library, with watermarking (see Figure 3.2) and reduced size for access on the Internet.

Figure 3.2: The First Cincinnati Haggadah
from the 15th century, with watermarking

3.1.2. The Advanced Papyrological Information System

(http://www.columbia.edu/dlc/apis/)

An integrated information system created for dealing with ancient papyri, ostraca (potsherds with writing), tablets, and similar artifacts. The project is sponsored by the American Society of Papyrologists. It is operated by a consortium of several American universities and funded in part by the National Endowment for the Humanities. The system is still under construction and it includes only small portions of the collections of the institutions involved. The idea pursued by the creators of this system is to construct a "virtual" library with digital images and a detailed database of cataloged records. The database will provide information concerning the external and the internal characteristics of each fragment, corrections to previously published papyri, and republications.

Here are few examples of collection involved in this project:

- **The Tebtunis Papyri Collection**

  (http://sunsite.berkeley.edu/APIS/)

  This collection consists of the papyrus documents that were found in the winter of 1899/1900 at the site of ancient Tebtunis, Egypt. The number of fragments contained in it exceeds 21,000. This web site provides electronic access to the images of some of the Tebtunis Papyri (Figure 3.3). Among the sites that I visited while preparing this paper, this site was probably the best.

Figure 3.3: A fragment from the Tebtunis Papyri collection

- **The Princeton University Library Papyrus Home Page**

  (http://www.princeton.edu/papyrus/)

  The Princeton papyri collection includes Greek documentary papyri such as census and tax registers, military lists, land conveyances, business records, petitions, private letters, and other sources of historical and paleographic interest from Ptolemaic (332 to 30 B.C.), Roman (30 B.C. to 300 A.D.), and Byzantine Egypt (300 to 650 A.D.). Nearly all were discovered from the 1890s to the 1920s.

3.1.3. Cambridge University Library - The Taylor Schechter Genizah Unit

(http://www.lib.cam.ac.uk/Taylor-Schechter)

The collection includes 140,000 fragments of Jewish literature and documents, rescued from the Ben Ezra Synagogue in Cairo at the end of the 19th century. The material covers aspects of Jewish religious, communal and personal life in the Mediterranean area, as well as documents which reveal relations with Muslims and Christians from as early as the ninth and tenth centuries.

This database contains 100 images of fragments (http://www.lib.cam.ac.uk/Taylor-Schechter/GOLD/princeton_images.html) with very interesting software which enables enlargements of a section of the fragments chosen by the distant viewer (Figure 3.4):

Figure 3.4: A fragment from the Cairo Genizah collection

3.1.4. The Archimedes Palimpsest

In 1906, A Danish scholar, J. C. Heiberg found in Constantinople (now Istanbul) a palimpsest of 185 leaves of parchment containing a tenth-century copy of Archimedes' *Methods.* (A palimpsest is a parchment on which some of the writings had been washed off with new writing appearing in its place.) In September 1998, the palimpsest was bought for $2,000,000 in an auctioned arranged by Christie's. In a press release, Christie's announced that "With the help of digital imaging, scholars have been able to bring out the lower script while suppressing the upper, offering an unprecedented opportunity to study of one of the greatest scientists of all time."

3.1.5. The Petra Scrolls

(http://www.cs.hut.fi/papyrus/)

The Petra scrolls were found in December 1993 in Jordan. They are carbonized papyri, where the lampblack text is almost indistinguishable from the carbon black background. The conservation work was led by professor Jaakko Frösén who provided the researchers with samples of similarly carbonized papyri fragments from another finding.

The homepage includes report of the work done by a group of researchers at Helsinki University of Technology. The first part describes various manners the authors employed for recording the sample they received. One method was to photograph the papyrus and scan the photo. The second method involved direct digitizing, for example, with a digitized camera. After having a digitized version, they tried to improve its reading with various digital filters. (I will elaborate on this project while discussing digital image processing methods in 5.2.)

3.1.6. The Gothic Manuscripts

_____

(www.cs.tut.fi/~dla/gothic.html)

For some years I have been endeavoring to advance the study of the Gothic manuscripts with the help of digital technology. Up to now I have prepared a CD-ROM version of the facsimile edition of the Codex Argenteus (1927) and compiled a preliminary diplomatic line for line transliteration of it. I also constructed an archetype for a multimedia Gothic dictionary and a multimedia format (in PC/Windows mode) for presenting the text with examples in image mode (http://www.students.tut.fi/~dla/john.exe). The study of the gothic manuscripts occupies the lion's share of this paper.

3.1.7. The Question of Copyrights

Since we are dealing here with photos of centuries old items, nobody can really claim copyrights either for the originals or the photos. However, several of the institutes who preserve them, are keen of recovering at least part of their expenses occurred while making the material available in digital mode. One method to accomplish it is by adding watermarks to the digitized photos, so nobody will be able to reproduce them without permission. Another method is to display the photos in such quality that it would not be beneficial to publish them in other mediums, like in printing. In my opinion, the whole issue is a red herring; there is very little interest or market for this material and since those texts are part of the cultural heritage of humanity, one can only hope that as many people as possible will be exposed to those old masterpieces.

3.1.8. Hardware and Software constraints

While papyrus and parchment, not to mention stones and clay, may remain in certain conditions readable for thousands of years, as technology moves ahead the time span of material used for recording cultural heritage has been becoming shorter and shorter. While a paper may survive for a few hundred years in a controlled environment, photos fade after few decades. If fact, there is already discussions of restoring by digital image processing methods faded photographic transparencies done 30 years ago (Braudaway 1993). In digital technology, the life span is even shorter.

The amount of cultural heritage to be recorded and preserved is enormous and is growing all the time. In the current state of digital technology, it would be unrealistic to endeavor to digitize every piece of heritage with cultural value and, therefore, some criteria should

be employed in order to establish priorities. The team which digitized El Archivo General de Indias, Sevilla (see above) followed the following:

- Only complete series would be digitized - never selected individual documents. This was the first and mandatory criterion.

- Documents found to be of greatest use for consultation would be selected. A statistical analysis would locate the documentary series that researchers had most often used.

- The documents selected would cover all territories relating to Spanish colonization in the New World. This would draw on the Archivo's strength as the basic archive for history of the Americas.

- As a practical criterion, the status of document description was also considered, with a view to the work of preliminary preparation.

(http://www.clir.org/pubs/reports/gonzalez/digital.html)

In my work, I digitize every photo of the original manuscripts I can lay my hands on; the number of leaves which remained extant is less than one thousand. However, I do not scan the photos with the best possible resolution. In fact, I usually use low to medium resolution, 100 –300 dpi, since I have in my disposal only a limited amount of storage facilities. Even mounting all the material on CD ROMs with the best possible resolution is impractical.

The notion of uniform access to existing and emerging digital collections and cultural heritage information resources is being advanced with the approval of a standard (Z39.50) as ISO 23950 (Moen 1989: 45). These resources include physical artifacts and digital renderings of them, works of art, bibliographic records, full-text documents, etc. The application of the approved standard aims at creating an integrated access to textual and non-textual digital collections.

3.2. Old Texts in the Form of Corpora

A corpus, in this context, is a collection of writings or works of a particular kind or on a particular subject e.g. a collection of recorded utterances used as a basis for the descriptive analysis of a language. Apparently, the first major collection and the most widely known is the *Brown corpus*, which was put together at Brown University in the

1960 and 1970s. It composes of selected written contemporary American English and includes press reportage, fiction, scientific text, legal text, etc. Its size is around one million words. In some cases, corpora are marked up according to a certain terminology and, in this manner, one can conduct data mining procedures.

3.2.1. Data Mining on Old Texts

Data mining is the process of identifying and retrieving useful information in a large collection of data. The process involves specification of goals, data pre-processing, choosing the appropriate algorithms and functions, the search itself and the interpretation of the results. Traditionally, analysis was strictly a manual process where one or more analysts would become intimately familiar with the data and, with the help of statistical techniques provide analysis of the data. However, as the quantity of data expands so fast that manual analysis simply cannot keep pace (Fayyad & Uthurusamy 1996).

Data mining techniques are used in the areas of marketing, manufacturing, health care, etc. However, text exploration requires different methods since it is a different form of data. A text is not a relational or automatically well-defined structure, that is, its components are not a priory sorted, e.g., according to customers, products, price, size etc. The first task in the process of text data mining is to define exactly what kind of information is sought after and tag the text so that distinguishable patterns can be discerned and extracted.

In fact, there are few low-level classification and processing tasks which can be executed with the raw text. A text is, after all, a list of words and one simple undertaking would be to count how many times each word appears. In English, the list is generally dominated by the little words such 'as', 'the', 'and', 'a', 'to', etc. These are usually referred to as function words, such as determiners, prepositions, which have important grammatical roles (Manning & Schütze 1999: 20). However, even such a 'simple' task as indexing is not so simple because the same word may appear in different modes, like plural, past tense, genitive, etc. and sheer counting would not produce especially useful information.

Another procedure of extracting information from raw text is known as *Key Word In Context* (*KWIC*) where the program displays all occurrences of the word of interest lined

up beneath one another, with surrounding context shown on both sides. This kind of sorting is of limited use, since it allows examination of only particular words, and the rest of the study is done manually. Here is an example produced by the program LMC with material I collected for a paper I wrote concerning Yiddish-origin words in English:

     101 d Reagan. "An act of remarkable *chutzpah*," reflects columnist

     485 irresistible combo of charm and *chutzpah* that makes all her

     534 p in 1985 turned out to be plain *chutzpah* in 1990.

     749 ment to the raw courage, if not *chutzpah*, of those same people

     762 that does work. Half a ounce of *chutzpah* can go a long way.

    1563 calibration beyond the Yiddish *chutzpah* for "sheer effronter

    2531 23-24, 1991: 6\ Champagne and *Chutzpah* in Cologne

    2533 - Take equal parts of champagne, *chutzpah* and fashionable fri

The numbers in the left side indicate the number of the line in the corpus.

3.2.2. Creating a Corpus

The first step in executing data mining method on texts is compiling a collection of raw material for a corpus. The idea behind the selection of material is to create a balanced corpus.

Once the material has been selected, the next step is to turn it into a machine-readable form. A description of this process is given by Hickey (1993) and in the next paragraphs I will follow his foot-steps.

Texts are either scanned with optical character recognition (OCR) software or keyed in directly. As Hickey writes "In either case it is more the exception than the rule to find out that a text turns up error-free in the computer. This banal fact increases the status of the individual who is responsible for text correction."

During the 1980s, the English philology department at Helsinki University assembled the so-called Helsinki Corpus. The material included a selection of early English prose fiction. In order to compile the Corpus, the department employed students who either typed the material into the computer or scanned it with Optical Character Recognition

(OCR) software. This Corpus was followed by the Helsinki Corpus of British English Dialects, Helsinki Corpus of Older Scots, The Corpus of Early English Correspondence, and some others.

Being a student at that department at that time, I also prepared my own corpus composed of sentences having Yiddish-origin words in them. I collected the material from books and newspapers, made copy of the relevant pages and scanned the copies with the OCR program OmniPage. The next step was proof-reading and applying the available software (Word Cruncher) for linguistic research. On the whole, I collected around 50 pages of the size A4. I did not keep record of the time it took me to scan and proofread the material but the work surely consumed many days. I would dare to suggest that for a good native-speaker typist, it would have taken only a fraction of that time to type the entries directly from the original printed material.

A Finnish company created a program which enables browsing different translations of the Bible. In order to digitize texts of older Finnish translations, they used different methods; one of them was scanning with OmniPage. The experience they encountered was similar to mine, that is, the program did not recognize certain fonts. As a result, they had had to spend a great deal of time and efforts in proofreading the digitized text, which made the benefit of using this software questionable. (Miika-Markus Järvelä, personal communication.)

The creators of OmniPage promise that the program "quickly, accurately and easily converts paper documents and images into editable text for use in your favorite applications" and that "better than 99% accuracy means you'll spend far less time proofreading and editing…." However, from my experience, the quality of the output depends heavily on the characteristics of the original texts and, with due respect, "better than 99% accuracy" might be achieved only in laboratory condition.

Optical Character Recognition software may be more useful in scanning old texts, which were transcribed in Latin fonts. The reason for this is that since there are no native-speakers of extinct languages, anybody who digitizes such texts must check each letter. Some years ago, while browsing the net for material related to the Gothic language, I came across a homepage of two Belgian students in which they included a small part of

Streitberg's version of the Gothic Bible (1908) digitized. In the marginal notes, those students made it clear that preparing this portion of the text involved a great deal of work which included typing and proofreading, and they asked for volunteers to help them in accomplishing the project. I sent them e-mail, suggesting that I scan for them a sample of the text with OCR software, while they would do the proofreading. They agreed to give it a try. As it turned out, the scanning with the OCR software was a great leap forward and soon another student joined in. Eventually all of Streitberg's text became available on-line, and the participants were publicly credited for their contribution. They also created a search engine using Microsoft VBScript Regular Expressions for finding words in the text. I use this facility in my study of the palimpsests (see below 6.4). As for the text itself, I modified it following my reading of the facsimile edition of the Codex Argenteus for creating a diplomatic (unedited) transliteration of the text:

(http://www.cs.tut.fi/~dla/Codex_Argenteus/Cod_Arg.html).

(I will discuss the subject of OCR more extensively in chapter 5.)

Back to Hickey's procedures for creating a corpus. According to him, the next step is normalization of the text. This process consists of replacing variants of a grammatical form by a single form by external consensus, reached by the corpus compilers. However, there is 'almost ideological dislike' of normalization, particularly on the part of medieval scholars. The advantage of normalization is in producing texts without undue linguistic difficulties. Normalization can be achieved by creating a database of all the occurrences of invariant forms vs. normal forms and with proper software to substitute the first with the latter. Normalization is an optional operation and the users of the corpus can perform it at wish, while leaving the original text unimpaired.

The next optional step is tagging. The Brown corpus which was mentioned earlier, is a tagged one. Tagging a corpus consumes a considerable amount of time. One way to do it is by adding labels to words, identifying them grammatically. There are different system of classification and my impression is that every team that creates a corpus, composes its own terminology and symbols, for example see Figure 3.5.

Category	Examples	Claws c5	Brown	Penn
Verb, base present form (not infinitive)	take, live	VVB	VB	VBP
Verb, infinitive	take, live	VVI	VB	VB
Verb, past tense	took, lived	VVD	VBD	VBD
Verb, present participle	taking, living	VVG	VBG	VBG
Verb, past/passive participle	taken, lived	VVN	VBN	VBN
Verb, present 3SG -s form	takes, lives	VVZ	VBZ	VBZ
Verb, auxiliary *do*, base	do	VDB	DO	VBP

Figure 3.5: Examples of different terminology for the same categories

Moreover, since tagging must be done 'by hand' even different members of the same team may, at least theoretically, mark the same text elements in slightly different manners. Hickey seems to have developed a program designed to tag text in automatic, semi-automatic or manual modes. I will have to use the program myself in order to believe that it is indeed useful. Hickey maintains that one either tags completely or not at all. In fact, tagging a text is such an enormous undertaking that not all corpora are actually being tagged.

Compilers of corpora may include text-relevant information, placed at the top of a file and includes such data as author's name, name of the sample, text type, dialect, date of the original, etc. One can generate a database with this header information and with proper software be able to look for certain grammatical elements.

When dealing with old texts of Germanic languages, e.g. Old and Middle, English, Gothic, High or Low German, one is encountered with the problem of special characters, for example *ash*, *eth* and *thorn* which are not included in the old 7-bit ASCII set. The compilers of the Helsinki corpus presented them as a+, d+ and t+, respectively. The disadvantage of this method is the awkward readability. In the 8-bit set, these letters appear as æ, ð, and þ.

One useful tool for executing data mining in a corpus is a *Corpus manager* which includes table of contents, searching facility both for single words or tagged strings, and an editor for correcting mistakes or making changes.

3.2.3. An Example: The Helsinki Corpus

As an example I will give some details of the Helsinki Corpus. The Corpus, which was completed in 1991, covers a time span of several centuries and stretches over three periods of English. It contains c. 1.5 million words and consists of c. 400 samples of continuous text, dating from the eighth to the beginning of the eighteenth century. It is divided into the Old, Middle and Early Modern English. As for tagging, the founder of this corpus writes:

> One of the most important future developments of the Helsinki Corpus will be the grammatical tagging of the texts. It seems difficult and time-consuming to develop automatic tagging or parsing programs for Old and Middle English texts, but software recently developed to do interactive 'semi-automatic' tagging has made the prospects of tagged versions much more realistic than earlier (Rissanen 1993: 74).

Another problem to be solved in the future is the preparation of normalized, that is 'corrected' or lemmatized ('sorted according to words as they appear in the original text') versions of the corpus. As Rissanen writes, "scribal creativity does not seem to know any limits in producing variants..." These variant spellings cause problems in the use of the corpus for the study of syntax or lexis and their normalization would necessitate a large number of awkward compromises.

3.2.4. Problems with Old Text Normalization

'Normalization' of text would involve creating some kind of a unified text, that is, a text where the same words are spelled in the same manner and grammatical structures are identical. The idea of normalizing a text is not new and was practiced by scribes for centuries. One result of this habit is that books of antiquity came to us in different versions.

Another problem is reconstruction of ancient text. In many cases when manuscripts remained extant, the handwritten script often became blurred or simply disappeared. One way to solve the problem has been to reconstruct the missing portions. However, there is no way to prove the correctness of the reconstruction. What makes the situation more complicated is that without proper marking, one cannot distinguish between a genuine text and a reconstructed one (which may be completely wrong).

_____

When it comes to 'corrections' of old texts, the field is widely open. For example, in the case of the Gothic manuscripts, even when the text is clear, there are places where the text includes obvious mistakes done by the original scribes. In the manuscripts there are also corrections but it is not clear who did them, the Goths themselves or latter-days 'editors.' Consequently, no one can always be sure whether the corrections are indeed 'correct.'

I maintain that the term 'normalization' of old texts is simply throwing out the baby with the bath water. In fact, philologists are greatly interested in spelling and grammatical variations and changes. From those details they actually deduce their theories.

3.2.5. Copyrights

Creating a corpus involves a great deal of work and, eventually, a big sum of money. Preparing a corpus involves not only digitizing the text but, initially, deciding which text should be included. In many cases, the text is tagged according to certain specifications. For those reasons, corpora are usually mounted on CD-ROMs and many organizations charge moderate sums of money for the right to use them. In this way, they endeavor to protect their work and receive some return for their investment. In the image processing field, the method usually employed for protecting copyrights is adding watermark to the photo (see above 3.1.1.).

A case which involves reconstruction of an ancient text and the question of copyrights, was resolved recently. The Israeli Supreme Court ruled that an Israeli scholar, Elisha Qimron, has a copyright over his reconstruction of a certain Dead Sea Scroll. This is, apparently, the first time in which copyright law had been applied in court to a reconstruction of an ancient document. The judges decided that by filling in gaps of missing text, deciphering and putting together scroll fragments, the scholar had shown "originality and Creativity" that earned him a copyright. (International Herald Tribune, August 31, 2000, p. 6). For a discussion of the case and other related entries on the topic of electronic copyright collection of public domain material see: http://www.cni.org/Hforums/cni-copyright/1993-03/0045.html.

In my opinion the case belongs to a category Jews call 'Jewish Wars,' in which two or more Jews battle oven a trifle and unimportant detail. I would imagine that with a proper referencing and a proper apology, the case would not have gone so far.

3.2.6. Digital Constraints

While I can read with little difficulty photos of text which was written at Qumran two thousand years ago, or 1500 years-old Gothic manuscripts, on most computers I cannot read text I wrote ten years ago. As mentioned before, the Helsinki Corpus was prepared with 7-bits ASCII code and, as a result, some Old English letters were rendered with an extra sign, e.g. t+. Consequently, the text looks awkward and requires a certain amount of time for becoming used to it. Since standards, hardware, and software are constantly evolving, there must be a way to keep the digitized material constantly retrievable. Eventually there will be a need to create some kind of digital 'museums' where one will be able to decipher 'old' digital texts.

_____

4. The Gothic Manuscripts

In the great debate concerning the nature of God, which had flared up in the early Church, the Goths happened to be Arians. As it turned out, they were on the loosing side and as such, were destined to history's dustbin. In practice, it meant that nobody studied or copied their writings and either by intentional destroying or simply by neglect, very little of their heritage has survived.

The main reason why anybody would be interested in what was left extant from their legacy is its linguistic value; Gothic is the oldest Germanic language of which we have written evidence. As a result, the people who actually study Gothic are mostly linguists. I became interested in the subject while studying English philology, that is, the history of the English language, at Helsinki University.

4.1. The History of the Manuscripts

Apparently, for a period of 40 years during the fourth century a Gothic Bishop, Wulfila, prepared a translation of almost all the Bible into the Gothic language. In order to accomplish this task he had had to invent letters. None of the original manuscripts has survived and the lion's share of those texts we have nowadays is apparently from the sixth century.

The best preserved manuscript is the so-called Codex Argenteus - the 'Silver Book.' The manuscript contained originally at least 336 leaves of which 188 remained extant; 187 are preserved at the Library of Uppsala University and one leaf, which was found in 1970, in Speyer, Germany. The text is written on parchment on both sides; some letters, like the first line of an Eusebian canon, are written in gold and the rest are in silver, hence the name – *argenteum*, 'silver' in Latin.

The other major source of Gothic texts is the Ambrosian codices, which are located, for the most part, in the Ambrosian library at Milan. Those manuscripts are palimpsests. After the Goths were defeated, the parchments they left behind were reused for other writings; the monks washed or scraped off the old Gothic sheepskins clean and wrote over them with Latin texts. In Greek palimpsest means "rescraping." Fortunately, one can

still see, with various degree of difficulty, the Gothic text underneath. The Ambrosian codices contain altogether 346 pages and 10 leaves, divided into five groups.

A 192 pages containing parts of the Epistles

B 154 pages containing parts of the Epistles

C 2 leaves containing fragments of St. Matthew 25-27

D 3 leaves containing parts of Nehemias 5-7

E 5 leaves containing part of the Skeireins.

In addition, there are few more remains of Gothic, which are not palimpsests: the Veronese marginal notes, the Deeds, and the Salzburg-Vienna Manuscript.

4.2. The Gothic Alphabet

It is generally assumed that Wulfila himself devised the Gothic alphabet; however, there is no consensus concerning the sources and models the Goth used while designing the letters. The alphabet contains 27 symbols, of which 25 are used both as numbers and letters and two letters which serve only as numbers. Consonant gradation, that is letters deleted in compliance with grammatical rules, is marked with a bar over the place of the deleted letter.

As it became a custom, the Gothic text has been rendered with Latin fonts with few changes (see below). With the advent of digital technology, Gothic fonts were created by Boudewijn Rempt of Yamada and are freely available.[1] The font can be easily installed on a PC, although not on a Sun terminal with a Unix operating system. Indeed, this is one of the reasons why I have to write this paper on a PC.

For the transliteration of the Gothic text, one uses the ISO-8859-1 (ISO Latin1) character set, which includes such characters as Þ, þ, Ï, ï. For rendering the Gothic x one uses *hv*. As with the Gothic original, the end of a sentence is marked with a point in the middle height of the letters: ·, which is also a character in the ISO-8859-1 standard.

4.3. The Gothic Text

The text Gothic is written in scriptio continua, that is, there are no white spaces between words. To illustrate the problem here is an example from English: GODISNOWHERE can be read as GOD IS NOW HERE or GOD IS NOWHERE. The interpretation depends on the context, a fact which adds the element of semantics into the process of deciphering the text. The ends of a sentences, parts of a sentence, phrases, etc. are marked with points. The division of the text follows the Eusebian canons and an end of such canon is most often marked with a colon. New chapters are numbered with numerical letters and if a new chapter happens to start on a new line, the first letter is capitalized. Other than this, there are no capital letters in the text, not even to mark proper nouns. There are differences in style and form among the various codices.

Wulfila followed an ancient practice and abbreviated terms which were considered to be holy (nomina sacra). This phenomenon is known from Greek and Latin texts which, by all probability, were in front of him. As a result, the Gothic text is abundant with abbreviations: *guþ* 'God' is abbreviated as *gþ*, *iesus* is shortened to *is*, *iesu* becomes *iu*, *xristus* is *xs* or *xaus*, *xristu* is rendered as *xu*, *frauja* 'Lord' is written *fa*. The same practice is employed whenever these terms appear in declination modes: *guþs* become *gþs*, *guþa* is *gþa*, *fraujan* is rendered *fan*, *fraujins* is shortened to *fins*, *iesuis* reads *iuis*.

For the most parts, there are established rules as for how the text is divided into words; however, the transliteration of the text being used are 'normalised', that is, they do not necessarily follow exactly the manuscripts. For example, as mentioned above, the original text uses abbreviations for holy terminology; nevertheless the various editors filled in the complete words.

There is a very limited vocabulary of known Gothic words, no more than few thousand. To compare, a medium size English-English dictionary contains 70,000-100,000 entries; the behemoth Oxford English Dictionary contains around 450,000 and its on-line edition over a million, including slang, neologisms (new words), etc. In order to fill some of the lexical gaps, philologists reconstruct words, marking them with the asterisk sign (*). Here arise two problems: firstly, different scholars come up with different suggestions; however, where do you find a native speaker of Gothic to tell us which version is correct?

And secondly, during the centuries of the study of Gothic, not all the editors indicated that the readings of blurred spots were actually their reconstruction, and not a genuine reading of the text. As a result, one cannot really trust the existing transliterations.

And there is still another problem: Kleberg (1984: 20) writes:

> Once, in the 1670s, a falsifier was at work on the Codex. By scraping out letters and painting over them with silver paint, he made some textual alternations, apparently with the aim of providing support for some of Olof Rudbeck's theories about "Great Sweden." It is perhaps impossible to prove who the culprit was. Olof Rudbeck himself would seem to be above suspicion.

Johansson (1955: 16) suggested that the letter w 'w' was altered to a 'a' and a 'a' to l 'l' (Figure 4.1).
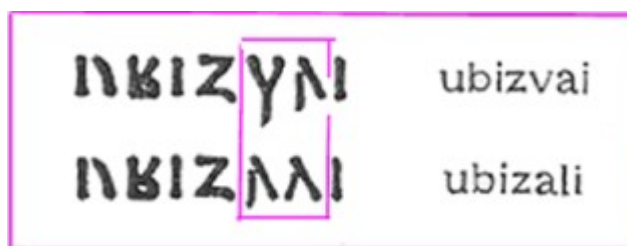
Figure 4.1: The suggested alternations

The apparent goal was to change the word 'ubizwai' which mean "porch" into 'ubizali, - "Uppsala." Examining the text in the facsimile edition (Figure 4.2), one can observe that at least the bottom left side of the letter a 'a' was deliberately cut and made l:
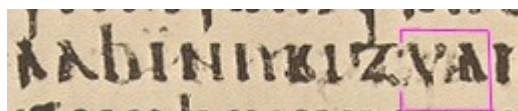

Figure 4.2: A photo of the original manuscript

which make the word 'ubizwli.' In the Biblical text (St. John 19:23) it is written that "And Jesus walked in the temple in Solomon's porch." The falsifier apparently intended to place Jesus also in Uppsala.

From the digital point of view, applying character recognition algorithms and methods as a tool for studying the text is next to impossible (see below 5.1).

4.4. Studies and Presentations of the Manuscripts

The first publication concerning a Gothic text, presently known as the Codex Argenteus, appeared in 1569. It was written by Johannes Goropius Becanus (Origines Antwerpianae), who probably obtained his knowledge from Georg Cassander and Cornelius Wouters. In 1597 a Dutchman, Bonaventura Vulcanius, published the text, bearing for the first time the title 'Codex Argenteus.' The publication was appended with text in Gothic fonts, prepared in woodcuts, followed by transliteration in Latin fonts and a Latin translation.

In the Junius' edition from 1665, the Gothic script is accompanied with a text in Latin fonts. For the Gothic script, Junius used special fonts which, in many details of design, were quite divergent from the corresponding letters in the Codex Argenteus. In 1677, matrices of those fonts had been presented to Oxford University and were used later in several publications (see below).

In Georg Stiernhielm's edition (1671) the transliteration of the text is in Latin fonts, the Icelandic and Swedish translations appear in the so-called 'Gothic' letters and the Latin translation is, naturally, in Latin fonts. In 1737, Lars Roberg, an Uppsala physician drew and made a woodcut of one page of the manuscript and prepared several impressions of it. The woodcut is still extant at the Linköping Diocesan and Regional Library (Klebrg1984: 23). The page was included in Benzelius' edition (1750). The few lines below (Figure 4.3) are a mirror rendering I prepared from a photo (Munkhammar 1988: 178).

Figure 4.3: A woodcut prepared by Lars Roberg in 1737

In Benzelius' edition the text is rendered in Gothic script and it is accompanied with a text in Latin fonts.

Elaborate copperplate facsimiles were included in Knittel's *editio princeps* (1762) of the Wolfenbüttel fragments (Codex Carolinus) of Paul's epistle to the Romans. Starting from

_____

Ihre_Zahn edition (1805), the tradition of printing the Gothic text with types based on the Gothic writing had gradually vanished.

One page of the Codex Argenteus, apparently prepared by an artist, appears in Uppström's edition of the text (1854-7). Below (Figure 4.4), is a digital rendering of a small portion of the page, which barely conveys the beauty of the drawing in Uppströn's book.

Figure 4.4: A rendering from the middle of the 19th century

In 1903 leaf 7, and in 1906 leaves 3, 4, and 8 of *Skeireins* were made available in a photo facsimile, followed in 1910 by the Giessen fragment and in 1914 the Wolfenbüttel fragments.

As time went by, there were plans for a reproduction of the Codex Argenteus by woodcut or copperplates engraving, but none of them was materialized. In 1927, a facsimile edition of the Codex was prepared and published (see below).

In 1991, Marchand posted in the Internet one page prepared by digital methods (Figure 4.5). Apparently the rendering is based on a photo from the facsimile edition of 1927.
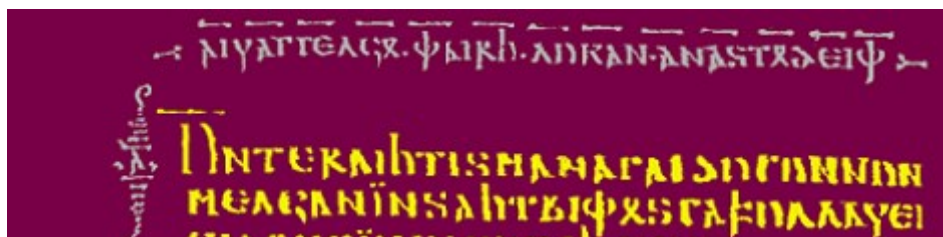

Figure 4.5: Marchand's digital restoration

After restoring the letters Marchand smoothed and filled contour and eliminated flyspecks and thumbprints with a "paint" program (Time, September 9, 1991 p. 9).

In 1998, according to an agreement signed between Tampere University of Technology and the Library of Uppsala University, I scanned the facsimile edition of the Codex Argenteus into 4 CD-ROMs, to be used in my research. In addition, I also received two slides containing photos prepared recently with advance photocopying technology. I converted these slides into a digital form. As part of my work, I have been developing software and making digital restoration of several plates from the facsimile edition. (I elaborated on this subject in 6.3.)

The first publication of Ambrosian codices appeared in 1809, prepared by Monsignor (later Cardinal) Angelo Mai and Count Carlo Ottavio Castiglione. The text is in the original Gothic script and is followed by transliteration in Latin fonts. For the Gothic script, Mai & Castiglione used matrices of Junian type, the same type which was given to Oxford University. In Figure 4.6 few lines of from the Ambrosian text featuring Junian's innovation are displayed. The letters are indeed neat, however, they are inaccurate.
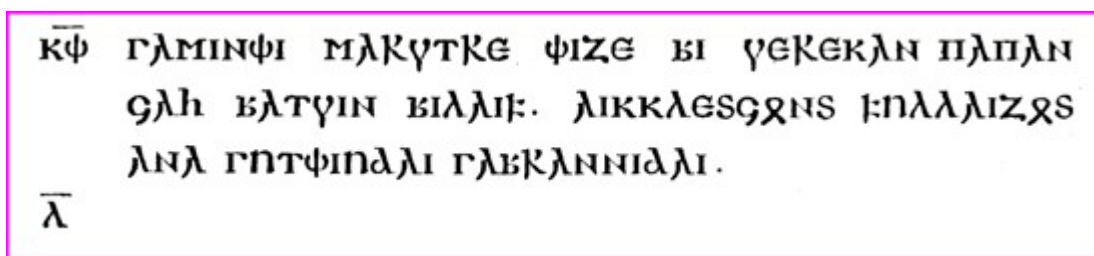

Figure 4.6: The Junian type

In Massmann's edition of the Skeireins (1833), the Gothic fonts are no longer used and the Gothic text is rendered in Latin fonts. The style that eventually had emerged is a combination of Latin and Scandinavian fonts; for example the Gothic letter ᚡ is being render with the Scandinavian Thorn *þ* (the Latin equivalent of *th*).

In 1936, a facsimile addition of the Ambrosian codices was published by Galbiati and de Vries. In my research I use scanned renderings of some of those photos. In early 1960s, the Ambrosiana collection was microfilmed. Of this collection I have scanned one black and white photo of a very good quality, published by Gabriel (1965). It is a rendering of palimpsest, which contains St. Jerome's Commentary on Isaias, written over Wulfila's

Gothic version of St. Paul Epistle to the Galatians. It appears that more recently the Codex was photographed again. I have seen some of these photos; however, I have not used them in this study.

4.5. Re-examining the Work of Earlier Scholars

One challenge facing contemporary researchers is to re-examine the work of earlier generations of scholars concerning the exact text of the Gothic manuscripts. Ebbinghaus (1985: 30) wrote:

> The question of the text is the most distressing one we are facing today in Gothic studies. In the discussion of Streitberg's 'Wörterbuch' above I have pointed out how uncertain we still are about the Italian mss. The editions of codd. Ambross. which were made on the basic of direct inspection (Castiglione, Massmann, Uppström, von der Gabelentz & Loebe [with the help from Castiglione], Bernhardt, and Streitberg [with Braun]) all contain an unknown number of erroneous readings. Only Bennett's edition of Skeireins is satisfactory.

Skeireins, which Ebbinghaus singled out, is the Gothic commentary of the Gospel of John, composes of eight leaves of palimpsests, each written on both sides and each side has two columns. Under the title *The scribal form of the text and its Modern "improvements"* Bennett, who did the most recent deciphering, wrote (1960, 1):

> Among works that have been metamorphosed by overzealous editing, few compare with the Gothic commentary on the gospel of John. The known leaves of this treatise comprise only 800 lines averaging about 13 letters each. Yet, if every word that scholars have added, deleted, replaced, transposed, or otherwise altered were to be counted separately, the total number of emendations would be approximately 1500. Proportionally, there is at least one modification for every seven letters in the manuscript. The extent to which the commentary has been transformed by emendations, ranging from changes in individual forms to rephrasing of entire passages, can be appreciated only through examining the text of one edition after another, but even the gross statistical evidence is significant.

Bennett's work with the manuscript had started in 1948 and continued, with pauses, for ten years. He had had access to the original manuscripts and for the deciphering process he used photographic methods, which he found to be dependable (1960, 25). As it turned

out, since that time no similar project was undertaken, which means that the rest of the palimpsests, more than 300, need to be thoroughly checked. This state of affairs is the source of Ebbinghaus' comment that "the question of the text is the most distressing one we are facing today in Gothic studies."

4.6. Features of the Material

The Codex Argenteus was written in silver ink on leaves of parchment and dyed purple. The Swedish linguist Johan Ihre suggested that the Codex was printed with hot stamps. He introduced this theory in a preface to Erik Sotberg's (his pupil) first dissertation (Ulphilas illustratus Uppsala 1752). However, this theory did not find support among other scholars and although it is mentioned later in less serious literature on the Codex, it apparently died with Ihre himself (Munkhammar 1998: 159, 171). Observing the beauty of the Codex and comparing it to other works of antiquity, one indeed wonders about the way the text was produced.

While preparing the facsimile edition (1927), the researchers released the pages from their binding, so each leaf might have more conveniently and accurately been photographed. As a result, it became possible to compare the different pages of the manuscript side by side. Systematic comparison revealed that actually two scribes were engaged in the production of the Codex, one being responsible for the Gospels Matthew-John (manus I) and the other for the Gospels Luke-Mark (manus II).

As it turned out, not only there were differences in the style of the writing, but also in the quality of the original writing which resulted in different times of exposures required for the photographing process for the different Gospels. Curiously, the distinction in exposure time did not follow the division between the scribes; the times of exposures were much shorter for the texts of the Gospels of St Matthew and St Luke than for St John and St Mark (Codex Argenteus Upsaliensis 1927: 122, as corrected by Friedrichsen 1930: 190).

Friedrichsen suggested this chain of events. Assuming that both scribes were working side by side, one of them started working on the Gospel of Matthew and the other on Luke. The second scribe seems to be less careful than the first. Once both were done with their respective initial assignments, it was decided to employ an ink containing a higher

proportion of silver on the text of the remaining Gospels of John and Mark, while still continuing to use the thinner ink for the decorative columns and arcs at the foot of the text. One possible result of this theory is that during illumination, in case of John and Mark the contrast between the text and the background is greater than the other books. Friedrichsen speculated (p.192) that "this silver was troublesome and costly to prepare, and it may needed more than one application, possibly by the artist-scribes, zealous for the perfection of their work, before a reluctant treasury granted the increase expenditure on silver." One result of the change of quality of the silver is that in the flourescence photography (see below) the writing in St. Matthew and St. Luke stands out but feebly against the background, whereas in St. John and St. Mark the contrast is much greater.

In general, the leaves exhibit different kind of damages, like discoloration and falling off of the silver, the fall off of the gold and the penetration of the ink from one side of the leaf to the other. After investigating and experimenting with different methods, the scholars, Theodor Svedberg and Ivar Nordlund, concluded that none of them would satisfactorily solve all the difficulties. As it turned out, two methods have proved to be decidedly superior to the rest (Codex Argenteus Upsaliensis, 119). The first method consisted of photographing with reflected ultra-violet light of the wave-length 366 μμ and the second was fluorescence photographing, in which the fluorescence is excited with the same wave-length, 366 μμ. Since these two methods complement one another in certain respects, the compilers of the facsimile edition decided to include in it a reproduction of each page by both these methods. The scholars rejected the thought of retouching of the plates since it necessarily introduced a factor which tended to make the photograph no longer a faithful reproduction of the original manuscript.

It was also evident that those methods did not successfully reproduce the gold in the text. Therefore, it was decided to add a supplementary collection of photographs on a smaller scale done in different methods, which reproduced better the lines inscribed in gold. For this collection, three methods of photography have been used: photography with a yellow filter, with secondary X-rays and with oblique illumination. The second method, using X-rays, turned out to be the most fruitful in the reproduction of traces of gold. This part of the work was done at the X-rays department of the University Hospital in Uppsala.

In addition, to avoid disturbances on the pictures caused by the penetration of the silver from the reverse side, when the flourescence method was used, the time of exposure was increased. However, as a result the lines written in gold, paragraph signs and the decorative columns with their parallel figures, were often exposed too much. This disadvantage was remedied in some degree by covering the text with one or more layer of transparent paper during the printing from the plates.

One more feature of the manuscripts that emerged during the work was that of the two sides of the parchment, the flesh-side (that is, of the skin of the animal) was always faded more than the hair-side. As a result, the average time of exposure for the latter was somewhat longer than the former. Another feature was that in the ultra-violet method, the structure of the parchment was revealed in the photograph.

The facsimile edition is indeed a masterpiece. One of the scholars who took part in the project, Theodor Svedberg (1884-1971) a chemist, was honored in 1926 with the Nobel Prize for Chemistry for developing the ultracentrifuge to facilitate separation of colloids and large molecules.

In recent years, several photos were taken with the best available equipment (Lars Munkhammar, personal communication). As I mentioned before, I received for my study two photos from this new set.

The Ambrosian manuscripts are palimpsests, that is, the sheepskins on which the original Gothic script was written, were washed away and written over in Latin. Figure 4.7 displays an example of a portion of a palimpsest.

Figure 4.7: A palimpsest

_____

In this case the Gothic text is relatively clear. Often it is quite hard to distinguish between the original Gothic text and the Latin letters written over it (see 6.4).

A facsimile edition of the Ambrosian Codices was published in 1936 by Galbiati and de Vries. As far as I can see, the introduction, written in Latin, does not give any technical details concerning the photographing methods. Each page is reproduced in one form. For this research, I use digital renderings scanned from photographic slides made from this edition. Apparently, a few years ago the Ambrosiana museum had prepared a new collection of photos of the manuscripts. They guard it zealously and although I have seen some of the photos, I have not used them. In any case, unlike the facsimile edition of the Codex Argenteus, it seems that no special techniques were employed while photographing, and I doubt whether those photos reveal more information than those in Galbiati and de Vries edition.

The Ambrosian codices were written by several scribes. As a result, there are several types of letters. Figure 4.8 displays some variations of letters in the various manuscripts, as rendered in the introduction to the facsimile edition of the Codex Argenteus.

Figure 4.8: Variations of letters in the various manuscripts

According to Marchand (1987: 26) there are 22 hands in the Gothic manuscripts.
In addition, the text is accompanied by several notations. The chapter numbers are marked with symbols, which include in the middle letters carrying numerical value (Figure 4.9).
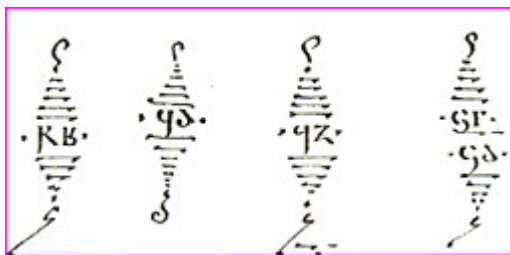
_____

Figure 4.9: Chapter numbers

As mentioned before, sacred terminology is abbreviated. In the Codex Argenteus the abbreviation is marked with a bar above it (Figure 4.10).
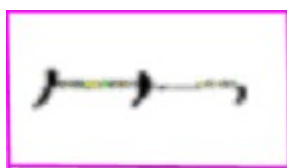

Figure 4.10: An abbreviation mark

In the manuscripts one find ligatures, that is, two or more letters joined together forming one character or type, a ligature. Figure 4.11 displays the word mahts, St Matthew VI: 13, plate 9, line 11 (the Codex Argenteus).

Figure 4.11: A ligature

The text was corrected in several ways by different hands and ,apparently, in different eras. Figure 4.12 demonstrates one such correction.
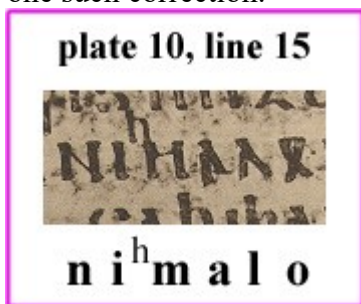
Figure 4.12: A correction

During the centuries some of the silver and gold letters of the Codex Argenteus faded or were contaminated with other material, see for example Figure 4.13.

_____

Figure 4.13: A faded and contaminated text

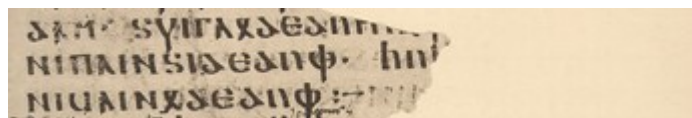In several places some of the text is simply missing (Figure 4.14).

Figure 4.14: A torn parchment

From the digital image processing point of view, the manuscripts present a plentitude of different problems, as a result, different approaches must be sought after and experimented with – for each problem its own solution. It is very possible that present technology does not offer solutions to some of the problems.

5. Examining Various Digital Image Processing Methods

In his article (1987), Marchand articulates the idea that following the personal computer revolution, the individual humanist "now has available to him computing power far beyond that of mainframes of the sixties" and as a result, the humanist is liberated from the mainframe, from the keepers of the mainframe and the "tyranny of the programmer." In this section I will examine different digital image processing methods which might be relevant to my work on the Gothic manuscripts.

5.1. Optical Character Recognition (OCR) Methods

In his survey of future applications the computer offer to the study of the humanities, Marchand (1987: 20) writes that "we are able to input texts in many different ways, and with the continuing work on the OCR, we may be able soon to begin reading manuscripts such as the Codex Argenteus, the major source of our knowledge of Gothic, with its regular handwriting, without benefit of the human finger."

As far as I know, up to now no one has created appropriate software for this kind undertaking, at least not as far as the Codex Argenteus is concerned. In the next few sub-sections I will examine this topic.

5.1.1. A General Survey of the OCR Field

Optical character recognition is the process of converting scanned images into a format the computer is able to process. The less an OCR system need human intervention, the better it is. The field is divided into two major areas of research and development: printed and handwritten text. [3]

The first step in any character recognition process is noise removal. If the image includes some background image, interference of foreign elements, dirt, etc. they should all be removed since noise can deceive any algorithm.

The next step is segmentation. The aim is to create recognition blocks or clusters which include characters or words to be recognized. The features of each segment are extracted.

The third step is recognition. Once a set of features of each character has been established, the recognition model should be able to identify it.

The last step is post-processing. In this stage, characters are classified, clustered into words and are compared with an appropriate so-called learned set or lexicon. This word-based recognition model works relatively well when the lexicon is small, for example when comparing a postal code number and the actual address or when verifying the amount of money written on a check in numbers vs. in words. It does not work very well when a large lexicon, e.g. the complete vocabulary of a language, is used. For example, one can apply a spelling checker to verify word spelling; however, while the spelling may be correct, the word itself might be a wrong choice.

Reliable character segmentation and recognition depend upon both the original document quality and the scanning process. To improve a poor quality original, one can use image enhancement and noise removal methods. One major problem in this process is how to remove such noise that touches characters without erasing information needed for successful recognition. One major step in this process is isolation of individual characters from the text image. In case of handwritten text, distinguishing among the letters can be very difficult because characters cannot be reliably isolated, especially when the text is a cursive handwriting.

Character recognition algorithms include two essential components: the feature extractors and the classifier. A feature extractor derives the features that a certain character possesses. The derived features are then used as input to the character classifier.

One common classification method is template matching, also known as matrix matching. A letter is, after all, a set of pixels in a matrix and therefore can be compared to another set of pixels. Comparing an input character image with a set of templates (or prototypes) from each character type may reveal various degrees of similarity between the input image and the templates. The template which has the highest degree of similarity, wins the contest and its identity is assigned to the input character.

Another method is structural classification. Characters are defined in terms of strokes, holes, concavities, etc. For example, the letter 'P' may be described as a vertical stroke with a hole attached on the upper right side. Another example: the lowercase 'k' has one vertical line joined in the middle to two diagonal lines. The structural features of the input character are extracted and are compared to a list of predefined lists of features and, again, the best wins. It is claimed that construction of a good feature set and a good rule-base of features can be time-consuming.

One direction of research concerning feature extraction is the search for invariants. These mathematical entities capture the characteristics of the character by filtering out all attributes which make the same character assume different appearances. A character can be deform because of geometrical transformation, have width variation or various font shapes etc. Applying normalizing transformations may produce well define variations and result in a single prototype per character.

To reduce the measure of misclassification, different mathematical routines are employed. For example, one may apply discriminant function classifiers which are used in order to reduce the mean-squared error, Bayesian classifiers which seek to minimize the loss function through the use of probability theory, and artificial neural networks that employ mathematical minimization techniques.

Any classification method needs a great deal of training for a proper response on characters that cause errors. All classifying methods may have limitations; as it is the rule in software engineering, one can never test against all possible errors. While recognition of machine-printed characters can reach in ideal laboratory condition, over 99%, in practice the achievement is lower (see also above 3.2.2. and below 5.1.2.). For results to be more credible, it is generally agreed that recognizers should be tested in realistic conditions of utilizations and on realistic test data. One team (Isabelle Guyon & Colin Warwick) suggests that public competitions be organized. When it comes to handwritten character recognition, rates descent even more, since every person writes differently. In fact, only a few OCR systems can read hand-printed alphanumerics. It seems that work on methods for recognition of degraded printed text and of running handwriting is still in the sphere of the research and development arena. Bow (1992: 13) writes:

> Applications of hand-printed character recognition are mainly for mail sorting. This problem has been studied for a long time. Due to the wide variations that exist in handwriting, the correct recognition rate is still not high enough for practical use.

5.1.2 Optical Character Recognition in Microfilmed Archives

In the beginning of the 1990s, two employees of the Technical Research Center of Finland (VTT) embarked on a research aiming at determining whether OCR is accurate enough for generating automatically full-text indexes for newspaper collections, either from the original newspaper pages, or 35 mm microfilm frames. In a publication from

_____

1994, the researchers Riitta Alkula & Kari Pieskä, describe the research and its results and conclusions.

A major tool of preserving newspaper collections and documents is mounting them on microfilms. The New York Public Library started experimenting with microfilming newspapers in 1933. Since then, microfilms have been a medium of choice for preserving large newspaper collections. It carries low cost tag, is durable, and is easily managed. However, reading a microfilm is uncomfortable, processing and distributing microfilms is burdensome, and retrieving information can be conducted only in the old-fashion method, that is, reading and searching.

With the advent of digital technology, experiments have been conducted in order to investigate the feasibility of digitizing both printed library material and microfilms. In this research, the test material was scanned with a microfilm scanner and an A4 size page scanner. The resulting image files were processed by OCR software to produce editable text files. The best recognition results were obtained with high contrast documents that contain sharply defined dark characters on white background. Stains and dirt were likely to cause troubles in scanning, since it was difficult to separate the text from the noise. The scanning resolution was 300. Increasing scanning resolution improved accuracy, however the recognition speed significantly decreased. There was direct correlation between image quality and OCR accuracy. Poor quality of microfilms, that is, poor contrast, twisted images, "show-through" of print from the reverse side of the page, etc. produce poor results.

The paper includes (p. 19) calculations made in connection to another project. One researcher (Saffady) had calculated that an A4 page with 2000 marks recognized with 90 percent accuracy would contain 200 errors. At five second per error, correction would require approximately 17 minutes. A professional typist will write the entire page in less than 12 minutes. With problematic material, it is cheaper to retype it than correcting a text file full of errors.

In order to improve the quality of the scanned material, image enhancement method were used to clean up the digital images and in this way give better prospects for OCR. However, this process, which included increasing contrast and sharpness of the image, occasionally resulted in loss of information when pixels were removed from degraded characters. As it turned out, enhancement could not be applied automatically. In addition,

OCR works best with simple type fonts and layouts; however, the library material is diverse and, as a result, recognition of characters appeared to be slow and unpredictable.

Typical mistakes found were substitution of one letter for another, two adjacent characters recognized as on character, e.g. "in" as "m" or "rn" as "m" and characters which were not recognized at all. Other problems were recognizing dirt or other stain as characters and deleting altogether too obscure letters or spaces. Obviously, accuracy in recognizing characters affects the accuracy of recognizing words.

To get relatively good results, the characters accuracy must be at least 98 percent in order to produce good recognition of words; after all, one spelling mistake cause a misreading of the whole word. As it turned out, the accuracy of recognized words was only 44 percent with scanned microfilmed images. Therefore, the authors concluded that "Microfilm scanning is a promising technology, but currently not adequately established for this project." They recommend that since more and more newspapers produce their texts in electronic form, indexes can be complied directly from these systems. As for digitizing newspaper archives on large scale, there is no readily designed solution.

5.1.3. The Gothic Text as a Candidate for OCR Methods

The first step in any OCR process is removing noise from the text, otherwise the optical device will mistake noise for characters or will not be able to distinguish between the letters and the noise. There are two kinds of Gothic text, the Codex Argenteus and the palimpsests. As far as the latter is concerned, the image is in most part 'noise.' I cannot imagine any way one can separate between the Gothic letters and the other elements in the page to such an extent that OCR method can be tried (see chapter 7). As for the Codex Argenteus, it takes me around two days to clean one plate (see chapter 6). However, by the time the text is clear no more character recognition is needed. Moreover, the amount of text to be cleaned is finite and final. The best tool for character recognition is still the human eye and not the computer. As a human being, I see no reason to distress over this state of affairs.

5.2. Digital Filtering: the Petra Scrolls

As mentioned before (3.1.5.) the Petra scrolls, which are carbonized papyri, were found in December 1993 in Jordan. A group of researchers at Helsinki University of Technology either photographed the papyrus and scanned the photo or used digitized camera. After having a digitized version, they tried to improve its reading with various digital filters.

They found out that the various ways to produce digital images led to different results. The measurements and tests have shown that the best contrast can be achieved in the near-infrared band of spectrum. Still, better spatial resolution in normal light may lead to better overall impression. A variety of digital filters and other image processing methods had been examined to enhance the images. Below are few examples. Figure 5.1 presents a photo of one papyrus from the collection.

Figure 5.1: A papyrus

The photo resulted from thresholding all pixels below 80 were set to 0 (black) and all those above to 255 (white) is shown in Figure 5.2.
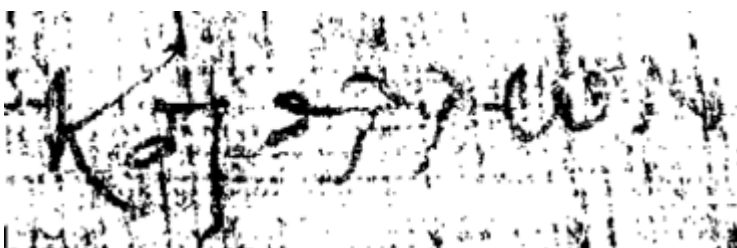
Figure 5.2: Thresholding the original photo

Figure 5.3 displays the image after histogram equalization.

Figure 5.3: The image after histogram equalization

Finding the horizontal derivative of the sample image (histogram equalized) produced these results (Figure 5.4).
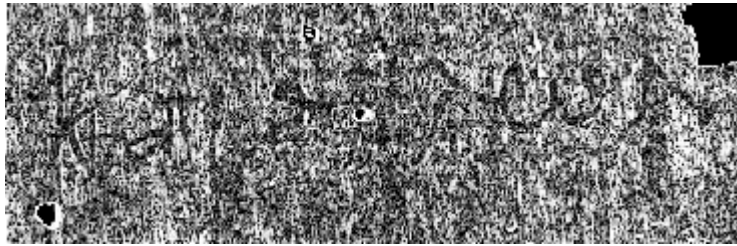
Figure 5.4: The horizontal derivative of the sample image

5.3. Exploring Frequency Domain Methods

One of the methods of extracting distinguishing features of a character is to transform it into the frequency domain. In this manner, one may be able to achieve an invariant representation of the spatial domain image. In his master's thesis, Tomi Lindholm (1998) studied and compared three feature extraction techniques. The first method was conducted in the space domain, and the other two include transformation into the frequency domain, using the Fourier and Hadamar transforms. The classification itself was conducted by utilizing the Intel 80170 ETANN (Electrically Trainable Analog Neural Network) chip.

The initial task was to reduce the measurement space in order to facilitate classification. The first method involved thresholding. All pixels below a threshold level were reduced to black and above it to white. The result was a bitmap image. The second method used the Fourier Transform. Applying this method on a discrete spatial domain image results in a two-dimensional frequency domain representation. The third method, the Hadamard transform, also transforms the spatial domain image into the frequency domain; however it uses the original image in a binary mode. The objective of the research was to train the Artificial Neural Network to carry out handwritten character classification and recognition tasks and, according to the author, the tasks were successfully completed and the objectives of this project were therefore fulfilled (p. 80).

From the more general point of view, using the frequency domain for feature extraction is problematic and seems not to be widely explored. One reason might be the need to have

the acquired data appropriately represented. In this research, the characters were originally presented each as a bit map of 640x480. Three steps in the feature extraction process were the same for all three representation types and consisted, among other things, of scaling the size of the images from 640x480 first to 128x128 and further to 8x8 which, obviously, causes a remarkable loss of information.

The objective of the feature selection and extraction process was to reduce the dimensionality of the measurement space in order to fit the data into a space suitable for classification. Only the features necessary for the recognition process are retained so that the classification can be implemented on a vastly reduced space. Consequently, the original bit map should have been of a very high quality, which is not a very common feature found in characters outside the laboratory. This kind of study would be impractical if applied to a running text, not just for one character or a small set of characters at a time.

_____

6. Studying the Gothic Manuscripts

In his article (1987: 18), Marchand describes one project where digital image enhancement techniques were employed in order to read manuscripts which cannot be read by the naked eye. The first part of the process is obtaining a picture of the manuscript, digitizing it and downloading it into the computer's memory. Marchand writes that

> the software assigns to the picture 256 levels of gray, each of them addressable. One can then ask the machine to do whatever one wants with each level of gray, give it any color, ignore it, etc. If, for example, the original manuscript is a palimpsest, in which one text is written over another, it is possible to separate the scripts. We are frequently able to read letters and words in Gothic manuscripts, for example, which have resisted all previous efforts. This takes, of course, enormous computing power,… The recent advent of FFT (fast Fourier transform) programs for the personal computer will lighten our burden tremendously.

In this chapter I examine existing filters as well as trying to follow Marchand's ideas concerning digital handling of the Gothic manuscripts.

6.1. Examining Existing Filters

It seems that the path to proceed in this research is through the spatial domain, which is the composite of pixels (picture elements) constituting the image. Spatial domain methods are procedures that manipulate directly on those pixels (Gonzales & Woods, 1992: 162). In general, the fundamental function is:

$$g(x,y) = T[f(x,y)]$$

where $f(x,y)$ is the output image, $g(x,y)$ is the processed image and T is an operator on $f(x,y)$. In most cases, the operator is a mathematical calculation, such as convolution or correlation, where the multiplier is a square matrix of the size of 3x3 or 5x5 whose center is moved from pixel to pixel, starting at the upper left corner.

One such method, known as mask processing, allows the values of f in a predefined neighborhood determine the value of g at (x,y). The value of this predefined neighborhood determines the nature of the process, such as image sharpening.

Another approach involves transforming the values of pixels above and below certain threshold or range of pixels in order to produce higher contrast in the vicinity of that certain threshold or range. Since the transformation at each point depends only on the gray level at that point, this method belongs to the category of point processing.

Eventually, the aim of this study is to highlight a specific range of gray levels, which constitutes the Gothic writing and eliminate everything else. One approach would be the enhancement of the edges of the letters. There exist various sharpening spatial filters which were designed for enhancing edges that have been blurred. One such category is derivative filters (Gonzalez & Woods, 1992, 197). With the help of Matlab, I will examine the applicability of three such filters: Roberts, Prewitt and Sobel, to the decipherment of the Gothic Manuscripts.

Figure 6.1 displays in grayscale mode an image taken from St. Mark, chapter 5.
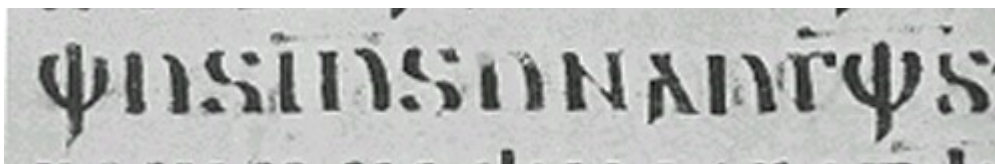
Figure 6.1: One line from the Codex Argenteus

Initiatiating the image into matlab with the command:

    I = imread('path/file_name.file_mode');

and the appropriate Matlab command:

    I = edge(I, 'sobel');

Figure 6.2 renders the results obtained in this process.

Figure 6.2: The line after being processed by the sobel filter (Matlab)

Matlab requires that before the filter is applied, the photo should be turned into a black and white mode. The result of the filtering is a binary picture where 1 is white and 0

black. Using the Roberts and Prewitt filters on Matlab produces identical results to those of Sobel.

The results produced by Matlab are inadequate since important information; the bars over the abbreviations which mark 'Jesus'(i-s-) and 'God,'(g-v)- which are quite clear in the original photo, have not been reproduced in the filtered version.

The program Gimp (an acronym for GNU Image Manipulation Program), which works on the Unix platform, enables the use of Sobel filter and the results are better (Figure 6.3).

Figure 6.3: The line after being processed by the sobel filter (Gimp)

For examining a palimpsest, I used a portion of the Gothic calendar (Figure 6.4).
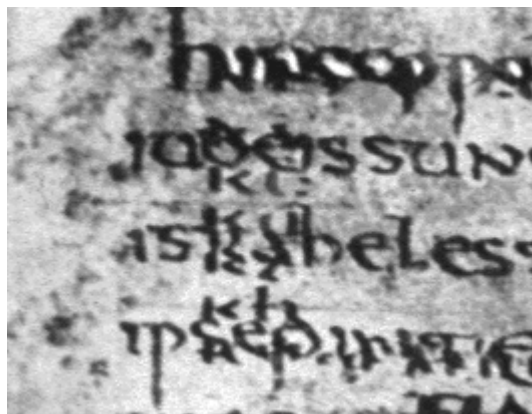
Figure 6.4: A photo of a portion of the Gothic calendar before filtering

The results obtained by Matlab's Sobel (Figure 6.5) are, for all practical purposes, useless).
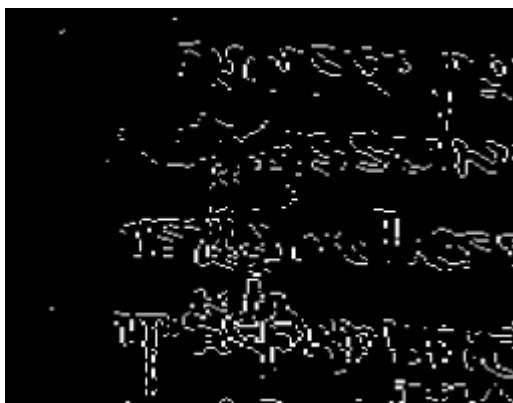
Figure 6.5: Matlab's rendering of a portion of the Gothic calendar with sobel filter

Gimp's Sobel produces a better image (Figure 6.6) than the one created by Matlab, however not adequately useful
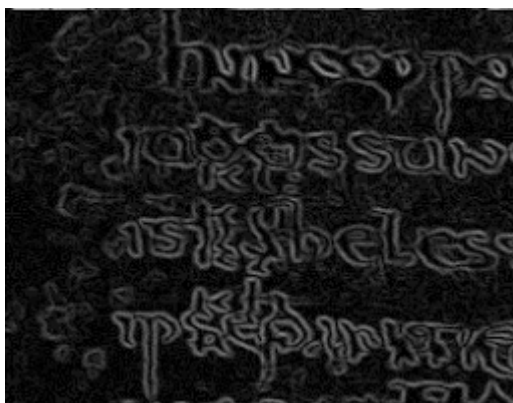
Figure 6.6: Gimp rendering of a portion of the Gothic calendar with sobel filter

Whith Laplacian filtering Gimp produces an image (Figure 6.7) which is of no practical use.

Figure 6.7: Laplacian filtering

Those filters are too crude for dealing with the Gothic manuscripts. The main problem is the abundance of noise, which the filter is unable to properly separate from the crucial data. One reason for the filter inadequacy is the fact that the difference between important

_____

and irrelevant pixels may constitute merely one or very few integers. In other words, there is a need for a filter which enables fine tuning. Since such filter does not seem to exist, I figure I should device my own software for better handling of the problems.

6.2. Developing Software

A digital picture is a two-dimensional array of integers. In a black and white photo each number represents the grayscale level of the pixel. The first task is to change the mode of picture into such that I can have direct access to it. With the help of the program xv I change the photo into an ascii mode with the type ending *.pgm* ('portable graymap file format').

Here is an example of a very small photo in ascii mode:

```
P2
# CREATOR: XV Version 3.10a  Rev: 12/29/94 (PNG patch 1.2)
8 6
255
117 135 145 135 135 141 130 148 110 116 121  87  99 115  90 117 145
149 134 135 144 128 129 146 152 135 100 118 127  97  99 133 141 144
117 127 117 112  88 171  99 159 128 118 134 131 160 142
```

The first line, p2, identifies the file type; the second line is a comment; the third line gives the width and the height of the matrix, starting at the top-left corner of the graymap; the fourth line indicates the maximum value of the grayscale, starting from 0 (black) and up to the highest number (white). In this case, the number is 255, which means a black and white photo of 8 bits. Programs that read this format should be as lenient as possible, accepting anything that looks remotely like a graymap.

In a colored photo, each pixel composes of three numbers, the first represents red, the second blue and the third green (RGB). Any other color is a combination of these three basic colors. Here is an example of the above picture, this time in colors:

```
P3
# CREATOR: XV Version 3.10a  Rev: 12/29/94 (PNG patch 1.2)
8 6
255
138 116  79 155 131 109 166 143 109 157 133  98 149 134 111
155 140 116 150 126 101 163 149 115 125 111  77 135 112  92
139 116  98 104  86  59 118  97  68 131 116  82 108  88  63
136 114  91 161 148 105 176 140 120 156 133  89 157 132  97
157 143 121 149 123  98 151 126  92 165 142 118 175 149 116
155 131 109 122  97  64 137 112  97 150 125  88 114  95  68
109 102  71 154 128 107 155 143 110 165 141 108 137 115  80
132 136  89 137 110  96 128 113  78 106  86  60 186 171 138
```

```
123  95  64 180 156 123 150 125  92 138 112  99 153 138  81
152 126 105 182 158 124 157 142 110
```

Unlike typical other programs that manipulate the same photo until an explicit 'save' command is given, each run of the function in the software I develop creates a new file. In addition, with the parameters which manipulate the image, I feed in also the name of a text file which records the parameters given each time. In this manner, I can follow the command I used and study them for detecting different alternatives. Obviously I also have to use the 'delete' command extensively in order to avoid packing the computer with big files.

6.2.1. Manipulating Grayscale Images

Following Marchand's idea, I wrote in C++ a program which does whatever one wants with each level of gray, give it any color, ignore it, etc." The first step in manipulating an image is turning it into an ascii mode. Next, the program reads the input file, retrieves the first four lines and transfers them, unaltered, to the output file. After that, the program reads each number from the input file, turns it into an integer and checks it against the arguments given to it. If the condition is fulfilled, the pixel receives a new value given as a parameter:

```
if ((Pixel <= Upper_Pixel) && (Pixel >= Lower_Pixel)){

        Pixel = Target_Pixel;
}
```

In other words, if the input pixel is equal to the upper or lower given values or between them, it receives a new value given by the variable Target_Pixel. The basic idea is to threshold the photo. The program enables thresholding any combination of pixels.

The key for being able to use the program is knowledge of Gothic. One has to know what to look for, what to emphasize and what to erase. The process of handling the manuscripts is slow and tedious, recursively examining each step, that is, if the results are insufficient, there is a need to return to the previous step and modify the arguments.

The biggest advantage of this method is that it allows the examination and the alteration of even one single pixel. Once overlapping Gothic and Latin letters are identified, one can

search for the borders among them by carefully calibrating the amount of change. That method of progressing does not guarantee a successful decipherment, but it may give a better grasp of the text.

The Gothic text was preserved in different form and quality, and each one of them needs its own process. Here are few examples. To start with, Figure 6.8 displays a line from the Codex Argenteus:
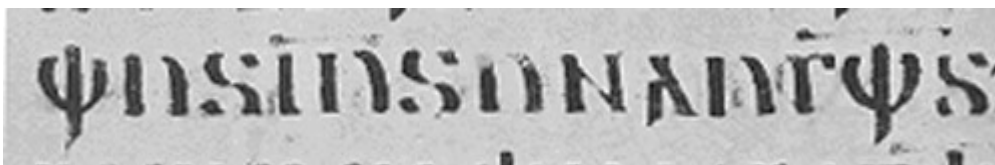

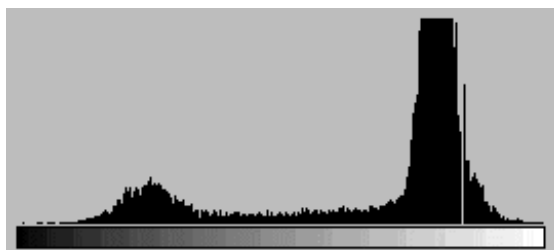Figure 6.8: A line from the Codex Argenteus

Figure 6.9 presents the histogram of the line above.

Figure 6.9: The histogram of the line in Figure 6.8

After a trial and error process, I gave all the pixels between 40 and 90 the value of 0 which, in practice, means that the letters were darkened (figure 6.10).
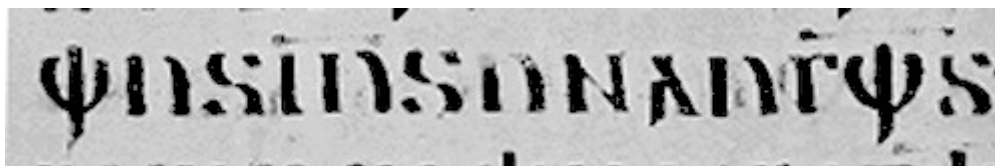
Figure 6.10: The line of Figure 6.8 after pixel manipulation

The new histogram is displayed in Figure 6.11.
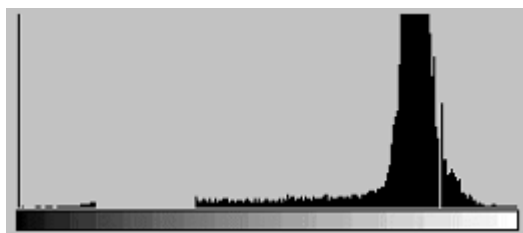
Figure 6.11: The histogram of the photo in Figure 6.10

Next, I gave all the pixels between 205 and 254 the value of 255 (Figure 6.12).

Figure 6.12

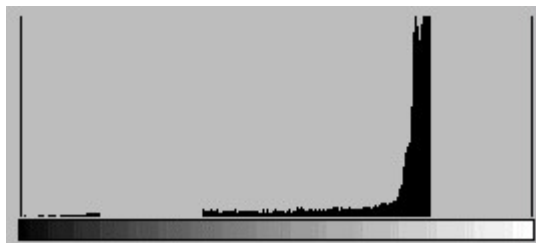and its histogram is presented in Figure 6.13.

Figure 6.13: The histogram of the photo in Figure 6.12

The tool I prepared has enabled me a full control of the values of the pixels. One may argue whether the processed image is better than the original; in any case, during the process, no significant information was lost.

In the next example (Figure 6.14), the original photo was partially distorted by stains and possibly seepage from the other side of the parchment.

Figure 6.14: A contaminated line from the Codex Argenteus

Figure 6.15 displays the same line after darkening the pixels between 0 and 80 and giving all those between 150 and 255 the value of 255 (white).
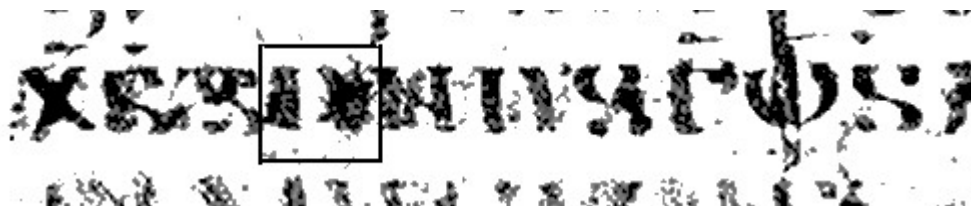

Figure 6.15: The line in Figure 6.14 after pixel manipulation

To clarify one letter, I handled it separately. Using the original photo, I gave the pixels between 64 and 69, 53 and 57, 106 and 112 the value of 255 (see Figure 6.16).
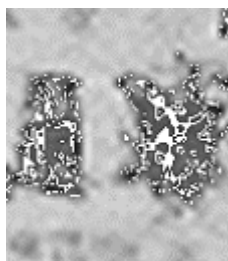

Figure 6.16: Handling one letter

In the 'negative' form one can distinguish the letter u . The text reads, as expected, "xs sunus gþs" ('Xristus son of God').

The next example (Figure 6.17) is from a palimpsest (Ephesians 1: 21). The Gothic text is between two lines written in Latin.
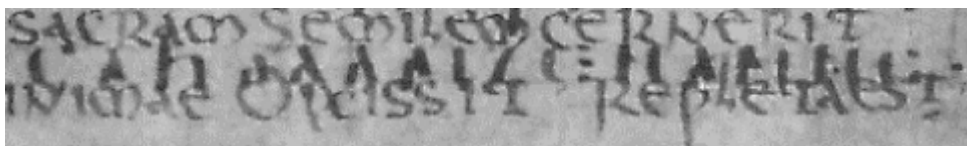

Figure 6.17: A line from a palimpsest

Figure 6.18 displays the same line after blackening the pixels between 0 and 100 and whitening all those above 130.
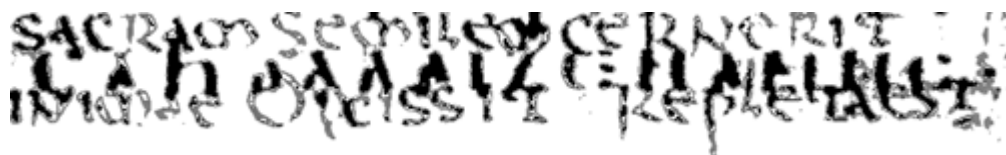

Figure 6.18: The line of Figure 6.17 after blackening the letters

With the helps of tools provided by Photoshop, I cropped those Latin letters which could clearly be separated from the Gothic ones (Figure 6.19). The Gothic text reads: "ja allaize namne·" ('and every name').

Figure 6.19: The line in figure 6.18 after cleaning

In the next example I endeavor to show that, despite an accepted reading, a certain word does not appear in the text. Figure 6.20 displays the months-line of the Gothic calendar (frames added).
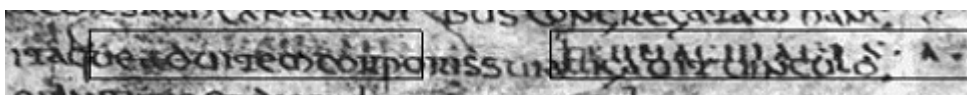

Figure 6.20: The months-line of the Gothic calendar

In the right frame, one can read 'frumajiuleis ·l·'. In the left frame, following a reading from 1833, it is generally accepted that the word 'naubaimbair' exists. After whitening all the pixels above 100, almost all the pixels in the right frame remain extant, however, the left frame is void of intelligible Gothic letters (Figure 6.21).
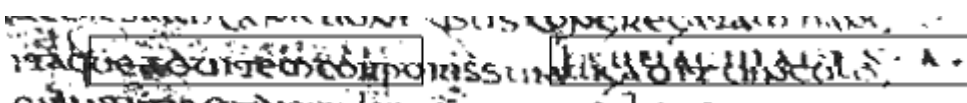

Figure 6.21: The months-line of the Gothic calendar after whitening the pixel above 100

The most difficult task is to clarify those Gothic lines which are just under the Latin text. Below (Figure 6.22) is an example taken from the calendar.
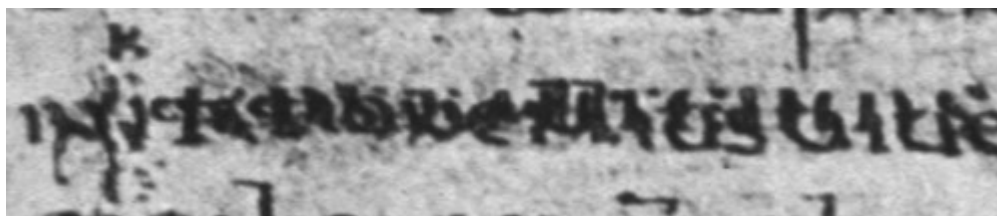

Figure 6.22: A line from the Gothic calendar

One way to separate the different fonts is to take a small portion, one letter or two at a time, and manipulate the pixels until the borders of the Gothic letters are distinguishable from the Latin ones. The first two letters are displayed in figure 6.23.
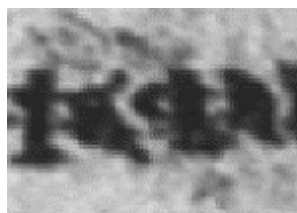

Figure 6.23: Two letters before 'clearing'

After whitening the pixels between 110 and 254, 44, 45, 52-55, 62-65, some borders have emerged (figure 6.24), which are the Gothic ku.

Figure 6.24: The letters after 'clearing'

Another way would be to detect the edge by giving the pixels between 60 and 80 the value of 0 (Figure 6.25).

Figure 6.25: Edge enhancement

Obviously, this is not much improvement on the original, however it assists in reading the text. One should also take into account the possibility that in some places the Gothic and the Latin texts had merged so thoroughly that there is no way to separate them.

6.2.2. Manipulating Color Photos

Enlarging the scope of bits from 8 (grayscale – 256 variations) to 24 (RGB – *truecolor*) enables a more delicate manipulation of the pixels. At first I simply enlarged the program to handle three successive numbers the same way as it did with one number. However, I soon found out that instead of automatically changing the value of each number, a far more practical solution would be to insert the three numbers into a vector and restrict the conditions for their manipulation. The resulting C++ code is:

```
If ((RGB[0] <= Upper_Pixel_Red) && (RGB[0]  >= Lower_Pixel_Red)&&
    (RGB[1] <= Upper_Pixel_Green)&& (RGB[1] >= Lower_Pixel_Green)&&
    (RGB[2] <= Upper_Pixel_Blue)&& (RGB[2]>= Lower_Pixel_Blue))

   {
     RGB[0] =  Target_Pixel_Red;
     RGB[1] =  Target_Pixel_Green;
     RGB[2] =  Target_Pixel_Blue ;
   }
```

In other words, only if the three colors are each between certain ranges, than the parameters given for new values are applied.

In addition, I made three more functions which enable the manipulation of one color at a time, while the other two remain unchanged. The RGB colors must fulfill the 'if' condition but the change of value apply only to one color. In this manner, there is no need to change all the colors if only one color need to be changed.

In the ultra-violet and the X-Ray photos the background is dark and the letters are a bit less dark. I found it more convenient to negate the colors by the code:

```
Pixel = 255 -Pixel
```

As with the grayscale process, the first step is to change the photo into ascii mode, which, in practice, causes more than doubling the size of the file. In this way, a typical scanned plate with the size of 3MB becomes an almost 7MB one. Processing the photo with one of the more elaborated functions I have developed lasts almost two minutes.

6.3. Digital Restoration of the Codex Argenteus

In 1940 Fairbanks wrote (314) that "since modern photography has brought the chief manuscripts to almost everyone's door, these monuments of the artistic and cultural past of one of the greatest of the Germanic peoples cannot be ignored." As it has turned out, the facsimile editions of both the Codex Argenteus (1927) and the Ambrosian Codices (1936) did not advance the study of Gothic. One reason is, no doubt, shifting interest in linguistics. Gothic was a major topic of studies in the 19th century, however, starting with the Swiss linguist F. de Sausseur (1857-1913), the founder of structural linguistics, the foci of interest moved to other linguistic fields. The second reason is that one cannot really read directly from the photos. After examining the facsimile edition of the Codex Argenteus I concluded that around 70-80% of the text is legible, around 5% is illegible and the rest can be read with varying degrees of difficulty. For reading the more difficult lines, one should compare and examine the flourescent, ultra-violet and, if there is a photo in the supplement section, of the X-rays, or yellow filter or oblique illumination photos (see 4.6).

Fairbanks also suggested the creation of Gothic fonts based on the script of the Codex Argenteus. He wrote (1940: 327):

> It is easily within the power of the capable and sensitive designer on the staff of more than one modern type-foundry to produce on the basis of the *Codex Argenteus* a showing which would make it forever unnecessary to design another Gothic font. *CA*, now happily within the reach of many in photofacsimile, provides the model of a definitive Gothic printing font, and it is to be hoped that we may not have long to wait for so important and attractive an adjunct to Germanic philology.

As it turned out, Fairbanks's hope did not materialize until a new technology - digital – has become ubiquitous. As mentioned before, a TrueType Gothic font was created by Boudewijn Rempt of Yamada Language Center at the University of Oregon. The font is free, easily obtainable[4] and installed. With it I can write Gothic on this computer:

A b g d e q z h v i k l m n j u p r t v f x o

Indeed, one can use this font for writing Gothic, although, I must admit, I could not find on the keyboard all the promised letters. If I understand correctly, they may be found on a Mactintosh.

In principle, with this set of letters one can reconstruct the Gothic text. However, although I have not tried it, I figure it would be quite hard to create the same form as the original manuscript. In order to restore the original style, there are, in my opinion, two ways. The first is simply to redraw the text. There exists such a restoration, attached to Uppström's edition (see above 4.4.). I assume that this certain page (St. Matthew. VI: 9-16) was chosen because it includes the Lord's Prayer. The second way is digital restoration, as suggested by Marchand. He also created one such page (see also above 4.4.). Using the software I prepared and assembled, I set about to follow this path. Up to now I have restored 4 plates and my expertise in this is obviously not polished yet. I figure that with more experience I will become more articulate. After all, the software accomplishes a great deal of the work; the rest is digital handicraft.

The first step in the restoration process is to choose the mode of the photo to work with, flourescent, ultraviolet or, if one exists in the supplement, a third one. From the limited experience I gathered, the flourescent image is easier to work with than the others. If one of the other alternatives seems to be of better quality, I negate the colors so the writing is in dark colors and the background bright.

With the aid of the software the threshold of the image is adjusted as much as possible. The text is rendered black and the background white. The basic idea is to threshold different segments one at the time. The main problem is that different parts of the page have different qualities. One way to solve the problem is to cut the page into different segments, however, as far as the Codex Argenteus is concerned, I have not done it yet; while dealing with the palimpsests, I actually cut each line into different photos (see below). Experience shows that after several thresholdings, no more than five, extra cleaning in one spot may cause deletion of essential information in other spots of the photo.

The next step is going through the processed photo, cleaning the noise that was left, checking that no information has disappeared, and restoring with a digital brush the letters or part of letters that disappeared from the original leaves during the centuries passed since the Codex was prepared. This part of the process takes around two days. I assume

that with better experience it might take less, but, at the same time, since for the time being I restore pages which are in relatively good condition, it is very likely that as my work proceeds, I will have to spend more time on restoring individual letters.

After the cleaning and checking is done, I change the color of the letters into silver, and those which were originally written in gold into yellow, following a tradition established by Uppström and Marchand. The reconstruction of missing letters or parts of them is marked green. The original Codex was written on purple parchment and for this reason both Uppström and Marchand painted the background in this color. Marchand wrote: (1987: 26):

> Imitating hands did not suit us, for we wanted more. We wanted to reconstitute a manuscript, to make a replica of the Codex Argenteus which looked more like the Codex than it itself did. That is, we wanted a manuscript which looked like the original ca. 520 A.D., one which people could read … The results are good enough to fool experts.

In this respect, I decided to deviate from the established tradition. In my view, a computer screen, with due respect, is not a parchment and what looks beautiful on one medium, does not necessarily look the same on another. Moreover, since most printers are still black and white, and, in addition, even with color printers one does not always know what comes out of the printer, instead of a purple background I mark the contour of the letters with purple pixels. In this manner I enhance the edge of the letter and, in general, sharpen the appearance of the text (Figure 6.26), even if the printing is done on a black and white printer.
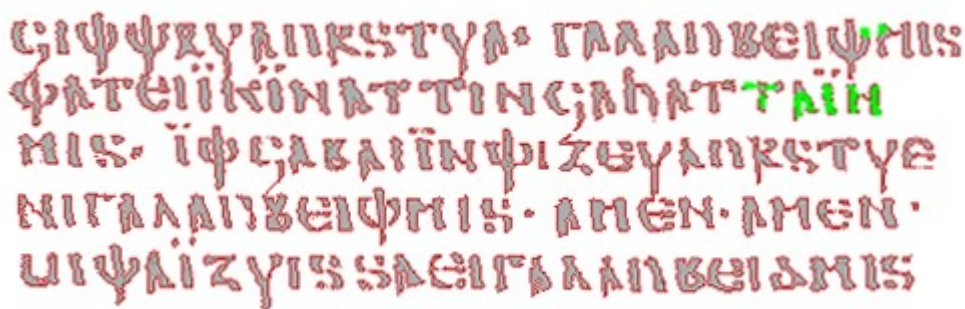

Figure 6.26: A restored portion of the Codex Argenteus

_____

In order to enhance the contour I wrote another function. The basic idea is that in the vicinity of the edge one pixel has the color of the background, in this case white, and the pixel on the edge of the letter is not white. The function retrieves two pixel at a time, push them into a vector of six places (three for each pixel) and than checks whether the first three have the same color of the background and the other three do not. If the condition is true, than the color of the edge is changed to the required color:

```
if((RGB[0] == First_Pixel_Red) && (RGB[1]==First_Pixel_Green)&&
   (RGB[2] == First_Pixel_Blue)&&

  (RGB[3] != First_Pixel_Red)&&(RGB[4] != First_Pixel_Green)&&
  (RGB[5] != First_Pixel_Blue))

{
  RGB[3] = Second_Pixel_Red;
  RGB[4] = Second_Pixel_Green;
  RGB[5] = Second_Pixel_Blue ;
}
```

The program checks a situation where the first pixel is part of the letter and the second belongs to the background. In this case, too, the edge of the letter received the color of the contour given in the parameters:

```
else if ((RGB[0] != First_Pixel_Red) &&
        (RGB[1] != First_Pixel_Green)&&
        (RGB[2] != First_Pixel_Blue)&&

        (RGB[3] == First_Pixel_Red)&&
        (RGB[4] == First_Pixel_Green)&&
        (RGB[5] == First_Pixel_Blue))
    {
      RGB[0] = Second_Pixel_Red;
      RGB[1] = Second_Pixel_Green;
      RGB[2] = Second_Pixel_Blue ;
    }
```

If the pair of pixels in the vector does not fulfill any of these alternatives, the program leaves the value of the pixel unchanged.

To cover a situation were both pixels are just before the edge are just after it, I added a function which skips the first pixel in the photo and starts retrieving two pixels at a time from the second pixel of the photo. Another problem is that the functions advance row after row, and in this manner do not check the top and bottom parts of the letters. To solve the problem I simply rotate the image so the colons become rows. Running both functions again, all edges are being enhanced. After this process is complete, all which is left to do is to rotate the photo back to its original position.

The last step is to change the mode of the photo into one readable by HTML and insert it in my homepage: http://www.cs.tut.fi/~dla/C_A_Restore/restore.html. (See also Appendix A.)

In 1970, one leaf of the Codex Argenteus was found in the Cathedral of Speyer, Germany. A good photo of it was published in Munkhammar book (1998: 98). I scanned the photo and for few hours tried to use the software I develop for cleaning it, but to no avail. My conclusion is the some kind of filtering like ultraviolet must be performed, before one attempts further manipulation. Those who studied the Petra Scroll reached the same
Conclusion; the measurements and tests they performed have shown that the best contrast can be achieved in the near-infrared band of spectrum. (see above 5.2.).

6.4. Deciphering the Palimpsests

Success in deciphering the palimpsests means the failure of someone else to accomplish his task given to him, namely, the erasing any traces of Gothic from the expensive parchments. The opposite is also true, that is, failure to decipher the Gothic text means that somebody had done his work properly, without taking a stand on the assignment given to him. What I am trying to say is that the palimpsests are a chaos. To study a photo of a palimpsest, I first cut it into lines and each line I divide into three parts. Using the software I have in my disposal, I try to find segments of letter. Unlike the handling of the Codex Argenteus, in studying the palimpsests I do not try to remove the noise, even from the simple reason that, from the point of view of the monks who washed away the original Gothic text, the noise is what I am interested in.

Here is the study of the first few lines of plate 50a of Codex Ambrosiana A (Romans XIII: 13,14 and XIV: 1-5). If I agree with the present reading I reproduce the accompanied letter in black; if I cannot confirm the reading, I reproduce it in an outline form. Places that I think are problematic, I mark with a question mark. For finding possible similar words I use the searching engine of Project Wulfila (see above 3.2.2.). For comparing to the Biblical text I use the King James Version.

In order to enable more efficient examining of the photo, I divide each line into three parts and enlarge them.
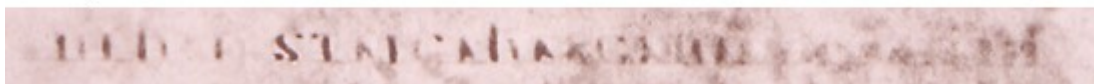
50a, LINE 1

Figure 6.27: Plate 50a line 1

50a, LINE 1, PART 1

NIhAIFSTAI

Figure 6.28: Plate 50a, line 1, part 1
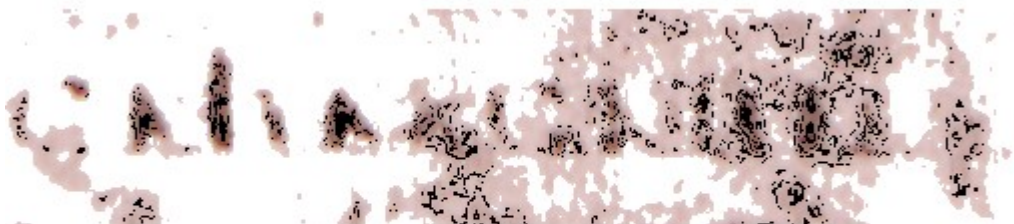
50a, LINE 1, PART 2

GAhAAGANA·

Figure 6.29: Plate 50a, line 1, part 2

50a LINE 1, PART 3

Figure 6.30: Plate 50a, line 1, part 3

It seems as if there are some letters in the third part (Figure 6.30). However, according to the Biblical text there should be none. Apparently, these are traces of letters from the other side of the parchment.

50a, LINE 2

Figure 6.31: Plate 50a line 2

50a, LINE 2, PART 1

Figure 6.32: Plate 50a, line 2, part 1

50a, LINE 2, PART 2

Figure 6.33: Plate 50a, line 2, part 2

50a, LINE 2, PART 3

Figure 6.34: Plate 50a, line 2, part 3

50a, LINE 3

Figure 6.35: Plate 50a line 3

50a LINE 3, PART 1

Figure 6.36: Plate 50a line 3, part 1

The present reading leikis 'body' (Figure 6.36) is well attested and semantically correct. However, the space between the letters seems to be too wide and, in addition, it seems that there is something else there. I raise the possibility of having the word leihtis 'flesh' there, with the combination of 'h' and 't' as one letter (ligature, see above 4.6.). In fact, the King James version uses the word 'flesh': "…and make no provision for the flesh" (Romans XIII: 14). The word leihtis appears once in the Gothic text in II Corinthians 1: 17: "… do I purpose according to the flesh…"
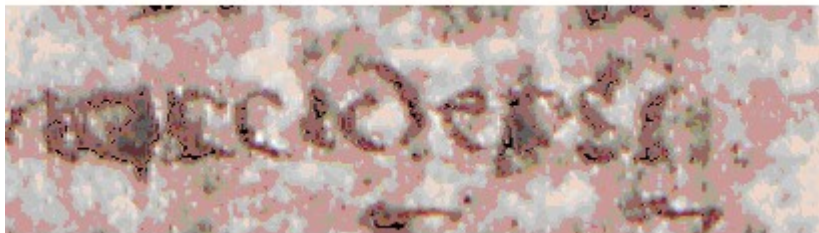
Figure 6.37: Plate 50a line 3, part 2

In this section (Figure 6.37) I could not discern any Gothic letter. I would suggest that the transliterated text "munni taujaiv is a reconstruction.

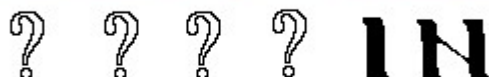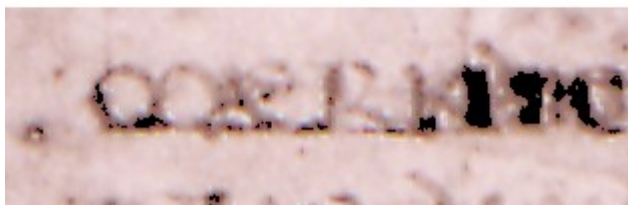Figure 6.38: Plate 50a line 4

Figure 6.39: Plate 50a line 4, part 1

Since this is the beginning of the line (Figure 6.39), some text must have been written there. One possibility is that the reconstructed word taujaiv is divided into two parts and the second part is at the beginning of this line.
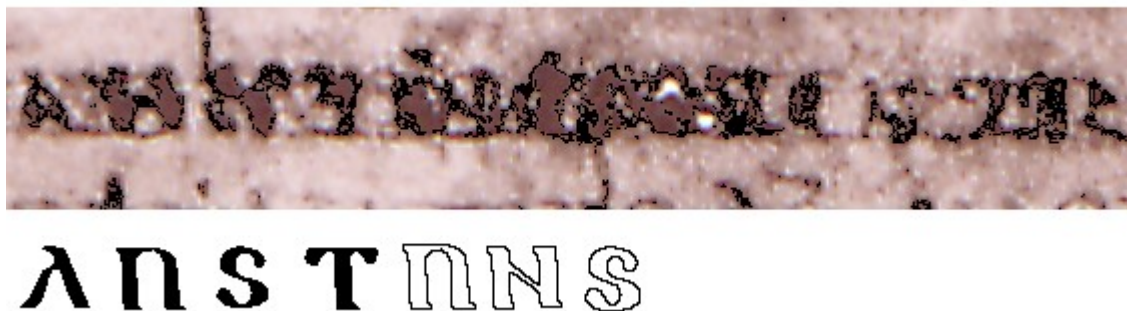
50a, LINE 4, PART 2

Figure 6.40: Plate 50a line 4, part 2

It seems as if there are still Gothic letter in this line (Figure 6.40), however, according to the Biblical text there should be none.

In the process of deciphering the Gothic manuscripts the best tools still are sharp eyes, experience and a good knowledge of Gothic. One also need a concordance or a search engine for exploring different alternatives in other sections of the text. Parts of the text are highly unintelligible and my guess is that it was unintelligible also almost two hundreds years ago when the first studies of the Ambrosian codices were conducted. Without better photos such as those prepared for the facsimile edition of the Codex Argenteus, it would be extremely hard to discern more details in those spots. However, careful digitization may help in distinguishing which part of the text was actually being read and which part is a reconstruction.

6.4. Preliminary specifications for *Wulfila*

When making the preliminary plan for this research, my original intention was to try to conduct it with Matlab. However, very soon I realized that I would need to write my own software. In addition to this software, I use three different image processing programs: Photoshop, xv and the Gimp; the first has only a PC version and the other two are designed to work on the Unix operating systems. Presently I work in front of two screens, one of a PC and the other one is a Unix terminal. To transfer data from one system to the other I use the program WS_FTP. Obviously, it is not very convenient to move from one program to another and from one operating system to the next. Eventually I will have to combine all the functions I need into one program, which works on a single operating system.

I call the designated program *Wulfila* in honor of Gothic Bishop who created such an excellent translation of the Bible into a language that before him did not have letters. In this chapter I specify the features I plan to include in this program. The format of this specification document is based on the Finnish Standard Association standard SFS-EN ISO 9001, which follows the international standard ISO 9001:1994 "Quality systems – Model for quality assurance in design, development, production, installation and servicing." The form here is an adaptation done at Tampere University of Technology. The specifications here are actually outlines. In the next stage, this section will be separated from this paper into an independent document, and enlarged.

1. Introduction

The purpose of the paper is to describe the specifications for a program designed to assist in deciphering, restoring and presenting old manuscripts. The immediate use is for solving problems encountered while handling the Gothic manuscripts.

1.1. Purpose and Scope

There are various existing digital image processing methods and algorithms, which are useful for handling manuscripts. The purpose of this design is to assemble them into one program and together with few more general tool for handling photos, like copying, cutting, pasting, saving, etc. In general, the idea is to make a program for handling a limited number of problems, not an encompassing digital behemoth.

In order to be able to handle an old text with a program, one must be able to familiar with the text itself, simply for knowing what to look for, what to keep and what to delete. For this reason, the designated user is a scholar who knows the language being studied and, at the same time, be familiar enough with digital technology in general, and also with features of digital image processing, e.g. the structure of a color photo.

1.2. Product and Environment

The name of the program is *Wulfila* after the 4th century mostly forgotten Gothic Bishop. Initially it will work with the Unix operating system and possibly Linux. In a later stage, if the need arises, it might be adapted to be used on a PC. Since the field of old text study is quite limited, it is assume that very few people will use it. Therefore, the purpose is not

to create an 'easy' program that one will be able to use after half an hour tutorial but, rather, a program that offers precision and diversity. It is not essential that the program execute its task fast. More essentially, it must produce results as accurate as possible.

1.3. Definitions, Acronyms and Abbreviations

Palimpsest: A palimpsest is a parchment on which some of the writings had been washed off with new writing appearing in its place

CA: The Codex Argenteus

ASCII: (The **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) 8-bits (ISO 8859-1).

Bitmap: A raster graphic file that consist of an arrangement of *pixels* (picture elements) that represent an image

RGB: Additive primary colors, Red, Green and Blue, used to create all other colors e.g. on a computer monitor. When pure red, green, and blue are superimposed on one other, they create white. A single byte is frequently used to store each of the color intensities, allowing the image to capture a total of 256x256x256 =16.8 million different colors.

Alpha data: Additional 8 bits that provide transparency information

1.4. References

This document is part of a study of different methods for studying old manuscripts.

2. Overview of system

Generally, the program will display a window with interfaces that enables executing certain functions, either directly, like save, cut, etc. or inserting parameters and then executing, like values of RGB colors. The command will be either marked with the mouse or typed on the keyboard. In case there is a need for other applications not specified here, the program should be compatible with other photo-handling programs.

3. Architectural design

The size of the program is small, probably a few thousand lines of code, and the general idea is to use as much as possible ready components, e.g. the container data structures and generic functions of C++ standard library (STL). The interface will take a form of a window with collapsing command columns. The architecture will be composed of two basic modules, the first incorporates the window and the interface command for the

general maintenance of the program and the second includes the interface of the functions which manipulate the image.

The program will not have a database. All the files created will be placed at the specified directory and will be opened by the program, manipulated and then closed. In this manner, the results are independent of the program and can easily be transferred to other destinations. The program will record the details of certain commands in text files specified in the commands themselves.

4. Detailed design

The program should be able to perform these tasks:

A pull-down file menu to *open*, *read*, *write*, *close*, *save*, *save as*, *delete* bitmap files formats of 8 bits (256 shades grayscale or 256 colors), 16 bits (thousands of colors) or 24 bits (millions of colors - *truecolor*). The following lists, by extension, the bitmap file format the program uses:

.TIF Tagged Image File format, used primarily in desktop publishing.

.BMP the Microsoft Windows bitmaps format

.JPG files follow the standards set by the Joint Photographic Experts Group of CCITT. These files use various *lossy* compression methods, that is, methods which cause loss of image quality as the file compression factor is increased. .jpg files are generally used for displaying photos on the Internet.

.GIF The Graphic Interchanged Format is a 8-bit format developed for portability among all PCs. Used mainly for Internet display.

.PGM Portable Graymap. A Unix native format for exchanging images in grayscale. The file composes of pixels in ASCII format.

.PPM Portable Pixelmap. A Unix native format for exchanging images in colors. The file composes of pixel in ASCII format.

The function to manipulate the bitmaps are:

The Color Picker used for choosing colors. On pressing the Color Picker button, a small information window appears. When marking any spot on the picture with the mouse, the information window displays the RGB and alpha values of the spot chosen. For grayscale image the information box displays an intensity values ranging from 0 to 255, where 0 is

black and 255 white. For a color (indexed) image, the box displays index numbers for each color from 0 to 255. The alpha value of an indexed image is either 0 or 255.

Palettes The flat thin tablet of wood or porcelain, used by an artist to lay and mix colors appears digitally as a box with either ready-made or personal assortment of indexed colors where the user will be able to edit and delete colors.

Pencil creates lines of different width. The pencil will not produce fuzzy edges. By choosing white color, the pencil serves as an eraser.

Color Selection Dialog enables manual changing of the colors. By typing the range of values of each color and the target color, in one command the whole image is transformed into another image, that is, another file according to the parameters given (see 6.2.1, 6.2.2.)

Grayscale and Color negation will give each pixel the opposite value (see 6.2.2.).

Zoom A magnifying tool which enables zoom in and zoom out with a click and release of the mouse button. The canvas will always adapt to the amount of zoom level.

Contour enhancement enables marking the edge of a pattern with a desired color (see 6.3.)

Events recording will save in specified text files every transformation done by the program

5. Specific technical solutions

The program should be designed to be as 'decentralized' as possible. With basic frame maintenance commands, like open, save, close, copy, cut, paste, etc. the rest of the functions should be 'independent' and easily corrected and amended, without affecting or depending on other parts of the program. In this manner, adding extra tasks should be easy.

After compilation, one would be able to mount the program on a CD-ROM and use it on other computers. As mentioned before, initially the program is designed to work on the Unix operating system and possibly Linux. Since the essence of the program is creating a new file each time a transformation is executed, it is the responsibility of the operating system to examine whether free memory space is available.

7. Conclusions

Digital technology offers new possibilities in handling old texts. However, there are no magic solutions for reading and presenting old texts. One can only hope that the current hierarchy of priorities, which put high on the ladder communication and medical applications, will leave some room for studies in the domain of cultural heritage.

In this paper I try to demonstrate that digital technology assists in restoring images of old texts and may help in deciphering those hard spots in the manuscripts which escaped tools used by earlier generations. In addition, this technology offers a high level of transparency to the process; in earlier stages, very few people had had access to the original text and the concerned scientific community should have trusted their judgments. With the advent of photography, the text became more widely available. Indeed, for this work I use photographs of the originals. However, with the wide use of CD ROMs and the easy access to the Internet, there are greater possibilities for creating access to the manuscripts and their studies. In fact, each scholar can, even with existing software, execute a great deal of digital manipulation.

The Gothic manuscripts introduce two main challenges, restoring and deciphering. The Codex Argenteus has been widely and extensively studied and the transliterated text used is generally reliable. Earlier attempts to restore it for smooth reading did not go beyond a few plates. With the software I have prepared, it takes me around two days to restore one plate. I figure it would take me three to four full years to restore all the leaves. Since the last person who used Gothic in daily life died well over one thousand years ago, the restoring process will end there.

With the palimpsest, any discussion of restoring is in vain. The most pressing problem is producing filtered photos to work with. Since the holders of the manuscripts, the curators of Ambrosiana Museum in Milano, are uncooperative, one possible path of research would be to try to imitate optical filters with digital methods. To imitate the affect of such filter on a 24 bit picture would be impossible, but to flatten the picture into 8 bits (256 colors may produce some results. The first step would be to take a photo with, e.g. an ultra-violet filter, of a palette of 256 colors and compare the transformed colors with the

original. The next step would be to prepare a program that transforms a photo into a 'filtered' one. Writing a program with one 'if' and 255 'else if' is quite tedious and time consuming, and the only consolation is to know that the computer will have to work much harder every time it processes an image. The main drawback for such an attempt is that one actually needs the original manuscript. Genuine ultraviolet and infrared radiation can sometimes reveal information that cannot be seen or photographed with white light. However, since the original manuscript is not available and, moreover, ultraviolet radiation may cause damage to the parchment, one must experiment also with unconventional methods.

It is that digital technology cannot solve all the problems encountered while studying the old manuscripts; however, this technology offers new methods and possibilities which were unavailable to earlier generations of scholars

Notes

1. http://babel.uoregon.edu/yamada/fonts/germanics.html

2. ftp://ftp.lang.uiuc.edu/pub/images/luke1cl2.gif

3. The general background material in this sub-section is based on articles at:

"http://www.cs.berkeley.edu/~fateman/kathey/char_recognition.html

http://cslu.cse.ogi.edu/HLTsurvey/ch2node3.html,129 (Sargur N. Srihari & Rohini K. Srihari, Richard G. Casey, Abdel Belaïd, Claudie Faure & Eric Lecolinet, Isabelle Guyon & Colin Warwick, Rejean Palmondon)

http://www.ccs.neu.edu/home/feneric/charrec.html (Eric W. Brown)

http://www.socr.org

http://www.cedar.buffalo.edu/Publications/TechReps/OCR/ocr.html: Sargur N. Srihari & Stephen W. Lam

4. http://babel.uoregon.edu/yamada/fonts/germanic.html

References

Alkula, Riitta & Pieskä, Kari. 1994. Optical Character Recognition in Microfilmed Newspaper Library Collection: A Feasibility Study. VTT Research Notes 1592. Espoo: Technical Research centre of Finland.

Bennett, William Holmes. 1960. The Gothic commentary on the Gospel of John: skeireins aiwaggeljons þairh iohannen. A Decipherment, Edition, and Translation (Modern Language Association of America Monograph Series 21). New York (Modern Language Association), 2nd [unaltered] ed. 1966.

Benzelius, Eric. 1750. Sacrorum Evangeliorum versio Gothica ex Codice Argenteo emendata atque suppleta, cum interpretatione Latina & annotationibus Erici Benzelii non ita pridem Archiepiscopi Upsaliensis. Edidit, observationes suas adjecit, et Grammaticam Gothicam præmisit Edwardus Lye a. m. Oxford, Clarendon Press.

Bow, Sing-Tze. 1992. Pattern Recognition and Image Preprocessing. New York: Marcel Dekker, Inc.

Braudaway, Gordon W. 1993. Restoration of Faded Photographic Transparencies by Digital Image Processing. IS&T's 46th Annual Conference, p. 287-289.

Codex argenteus Upsaliensis iussu Senatus Universitatis phototypice editus Upsaliae, s. a. 1927

Ebbinghaus, Ernst A. 1985. Gothic Lexicography, part II. General Linguistics, Vol. 25, No. 1 218-231.

Ekstrom, Mihael, P. (ed.). 1984. Digital Image Processing Techniques. Orlando: Academic Press, Inc.

Fayyad, Usama & Ramasamy, Uthurusamy. 1996. Data Mining and Knowledge Discovery in Databases. Communication of the Acm. Volume 39, Number 11 (November), 24-26.

Gabelentz, H.C. de & Löbe, J. 1836-46. Ulfilas: veteris et novi testamenti versionis gothicae fragmenta quae supersunt ad fidem codd. castigata latinitate donata adnotatione critica instructa cum glossario et grammatica linguae gothicae conjunctis curis ed. H.C. de Gabelentz et J. Loebe. Leipzig.

Gabriel, A.L. & Garvin, J.N., 1965. Folia Ambrosiana. NotreDame: The Mediaeval Institute University of Notre Dame.

_____

Galbiati, G. & de Vries, J. 1936. Ulfilas: Wulfilae codices Ambrosiani rescripti epistularum evangelicarum textum goticum exhibentes / Phototyp. ed. et prooemio instr. a J. de Vries. Ed. G. Galbiati inscripta. Florentiae: Augustae Taurinorum.

Gladney, Henry M. et al. 1998. Digital Access to Aniquities. Communications. Volume 41, Number 4, pp. 49-57.

Gonzalez, Rafael C. & Woods, Richard E. 1992. Digital Image Processing. Reading, Ma: Addison-Wesley.

Grape, Anders & Friesen, Otto von. 1928. Om Codex argenteus. Uppsala. (Skrifter utgivna av Svenska litteratursällskapet; 27).

Fairbanks, Sydney. 1940. On Writing and Printing Gothic. Speculum. Vol. 15: 313-327.

Friedrichsen, George Washington Salisbury. 1930. The Silver Ink of the Codex Argenteus. Journal of Theological Studies. XXXI, pp. 189-192.

Hickey, Raymond. 1993. Application of Software in the Compilation of Corpora. In Kytö et. al.: 165-186.

Johansson, Viktor. 1955. De Rudbekianska Förfalskningarna I Codex Argenteus. Nordisk Tidskrift för Bok- och Biblioteksväsen (Årgång 42), pp. 12-27.

Junius, Franciscus. 1665. Quatuor D[omini] N[ostri] Jesu Christi Evangeliorum versiones perantiquæ duæ, Gothica scil. Et Anglo-Saxonica. Quarum illam ex celeberrimo Codice argenteo nunc primum depromsit Franciscus Junius F[rancisci] F[ilius], Hanc autem ex codicibus mss. collatis emendatius recudi curavit Thomas Mareschallus, Anglus. Cujus etiam observationes in utramque versionem subnectuntur. Accessit & glossarium Gothicum: cui præmittitur alphabetum Gothicorum, Runicum, &c.opera ejusdem Francisci Junii. Dordrecht (Typis et sumptibus Junianis; excudebant Henricus & Joannes Essæi).

Kleberg, Tönnes. 1984. Codex Argenteus, the Silver Bible at Uppsala. Uppsala: Uppsala University Library

Knittel, Franz Anton. 1762. Ulphilae versio gothica nonnullorum capitum epistolae Pauli ad Romanos. Braunschweig.

Kyttö, Merja et al. 1993. Corpora Across the Centuries. Proceedings of the First International Colloquium on English Diachronic Corpora. Amsterdam: Rodopi.

Lacy, Alan. F. 1980. Cyzicus and the Gothic calendar. General Linguistics. Vol 20 No. 2: 99-103.

Lim, Jae S. 1984. Image Enhancement. In Ekstrom, M. (ed).

Lindholm, Tomi. 1998. Handwritten Character Recognition Using Neural Network Processor. Master of Science Thesis. Department of Information Technology, Tampere University of Technology.

Maij, Angelo & Castiglione, Carlo Ottavio.1809. Vlphilae partivm ineditarvm. Mediolani. 1829(Ed. Castiglione, Carlo Ottavio) gothica versio Epistolae divi pavli ad corinthios secvndae.

1833. Scriptorum veterum nova collectio e vaticanis codicibus edita (1-10). Rome: Typis Vaticani.

Manning, Christopher D. & Schütze, Hinrich. 1999. Foundation of Statistical Natural Language Processing. Cambridge: Massachusetts: MIT Press.

Marchand, James W. 1958. The Gothic Language. Orbis. Vol. VII, No.1: 492-515.

1987. The Uses of the Personal Computer in the Humanities. Ideal (Issues and Developments in English and Applied Linguistics), Vol. 2: Computers in Language Research and Language Learning. 17- 32.

Maßmann, Hans Ferdinand. 1834. Skeireins aiwaggeljons pairh Iohannen, Auslegung des Evangelii Johannis in gothischer Sprache. Aus römischen und mayländischen Schriften, nebst lateinischer Uebersetzung, belegenden Anmerkungen, geschichtlicher Untersuchung, gothisch- lateinischem Wörterbuche und Schriftproben. Im Auftrage seiner k. H. des Kronprinzen Maximilian von Bayern erlesen, erläutert und zum ersten Male herausgegeben. München: Jaquet.

1857. Ulfilas. Stuttgart: Liesching

Moen, William E. 1998. Accessing Distributed Cultural Heritage Information. Communications. Volume 41, Number 4, pp. 45-48.

Munkhammar, Lars. 1998. Silverbibeln: Theoderiks bok. Stockholm: Carlsson Bokförlag.

Rissanen, Matti. 1993. The Helsinki Corpus of English Texts. In Kyttö et. al.: 73-81.

Schäferdiek, Knut. 1988. Das gotische liturgische Kalenderfragment - Bruchstück eines Konstantinopeler Martyrologs. Zeitschrift für die Neutestamentliche Wissenschaft und die Kunde der Älteren Kirche.Vol. 79, pp. 116-137.

Sotberg, Erik B. 1755. Ulphilas illustratus. Sockholm: Salivius.

Stiernhielm, Georg. 1671. D. N. Jesu Christi SS. Evangelia ab Ulfila Gothorum in Moesia Episcopo circa annum à nato Christo CCCLX. Ex Græco Gothicè translata, nunc cum parallelis versionibus, Sveo-Gothicâ, Norrænâ, seu Islandicâ , & vulgatâ Latinâ edita. Stockholmiæ. Typis Nicolai Wankie regij typogra.

Streitberg, Wilhelm. 1960 [1908]. Die Gotische Bibel. Heidelberg: Carl Winter.

Uppström, Anders. 1854-7. Codex argenteus sive Sacrorum Evangeliorum Gothicae fragmenta. Uppsala

1861. Fragmenta gothica selecta ad fidem codicum ambrosianorum carolini vaticani. Uppsala, . (It contains the Ambrosian C fragment of Matthew, Codex Carolinus and the Skeireins.)

1864-68. Codices gotici Ambrosiani. Upsaliae.

Appendix A: Digitally Restored Images of the Codex Argenteus

(See also: http://www.cs.tut.fi/~dla/C_A_Restore/restore.html)

St. John XIV: 6-12. Plate 106

Digital restoration made by David Landau (TUT), July 2001

St. Luke I: 6-14. Plate 136

Digital restoration, made by David Landau (TUT) July 2001

St. Mark IV: 4 10. Plate 294.

Digital restoration, made by David Landau (TUT), July 2001

St. Mark IV: 16-21.Plate 296.

Digital restoration, made by David Landau (TUT), July 2001

Tampere University of Technology
Institute of Software Systems Reports

1. R. Kurki-Suonio, and T. Kankaanpää, On the design of reactive systems, 1988.
2. R. Kurki-Suonio, Specification of serializable databases by joint actions, 1988.
3. P. Puusaari, L. Sintonen, and T. Virvalo, In-machine communication - a general solution for interconnecting independent intelligent actuators, 1988.
4. A. Valmari, Some polynomial space complete concurrency problems, 1988.
5. R. Kurki-Suonio, and H.-M. Järvinen, Action system approach to the specification and design of distributed systems, 1989.
6. H.-M. Järvinen, R. Kurki-Suonio, M. Sakkinen, and K. Systä, Object-oriented specification of reactive systems, 1989.
7. R. Kurki-Suonio, Operational specification with joint actions: serializable databases, 1989.
8. H.-M. Järvinen, and R. Kurki-Suonio, The DisCo language, 1990.
9. P. Jokela (ed.), TCP/IP protocol environment, 1990 (in Finnish).
10. H.-M. Järvinen, and R. Kurki-Suonio, DisCo specification language: marriage of actions and objects, 1990.
11. H.-M. Järvinen, and R. Kurki-Suonio, The DisCo language and temporal logic of actions, 1990.
12. N. Glibovets, The class of neighboring heuristic algorithms for graph coloring, 1991.
13. N.N. Glibovets, R. Kurki-Suonio, K.G. Zadneprovsky, An algorithm for the isomorphism problem of connected graphs without loops and folded edges, 1991.
14. K.Systä, Pertti Kellomäki, and Rauno Mäkinen (eds.), Proceedings of the Nordic Workshop on Programming Environment Research, 1992.
15. V. Kaufman, Pattern Calculus, 1993.
16. A. Valmari, and M. Tienari, Compositional Failure-Based Semantic Models for BASIC Lotos, 1993.
17. A. Valmari, and M. Setälä, Visual Verification of Safety and Liveness, 1995.
18. T. Mikkonen, A Layer-Based Formalization of an On-Board Instrument, 1998.
19. T. Aaltonen, and R. Pitkänen, Verifying Safety by Combining Joint Actions with a Process-Algebraic Approach, 1999.
20. T. Systä, Proceedings of OOPSLA 2000 Workshop: Scenario-based round-trip engineering, 2000.
21. K. Koskimies, Proceedings of OOPSLA 2000 Workshop: Methods and Tools for Object-Oriented Framework Development and Specialization, 2000.
22. M. Hakala, J. Hautamäki, K. Koskimies, J. Paakki, A. Viljamaa, A. Viljamaa, Task-Driven Specialization Support for Object-Oriented Frameworks, 2001.
23. M. Harsu, A Survey of Product-Line Architectures, 2001.
24. A. Puhakka, and A. Valmari, Liveness and Fairness in Process-Algebraic Verification, 2001.
25. M. Harsu, Ohjelmien analysoinnin ja uudelleenmuokkauksen tekniikat, 2001.
26. D. Landau, The study of old texts with the aid of digital technology: the Gothic manuscripts, 2001.