Ludwig-Maximilians-Universität München
Centrum für Informations- und Sprachverarbeitung (CIS)

Schriftliche Hausarbeit im Fach Computerlinguistik zur Erlangung des akademischen Grades Magister Artium (M.A.)

OCR Postcorrection of Historical Texts

ANDREAS W. HAUSER
August-Kühn-Str. 2
80339 München
E-Mail: andy@splashground.de

Abgabetermin: 4. Oktober 2007
1. Korrektor: Prof. Dr. Klaus U. Schulz
2. Korrektor: Prof. Dr. Franz Günthner

Contents

List of Tables

List of Figures

Abstract

More and more OCR software is available for recognizing historical texts, which are often written in Blackletter fonts. Recognition rates are still far lower, then on current fonts. Postcorrection of the OCR result is therefore an essential part of the process of getting better results.
This work concentrates on the postcorrection of historical German texts, printed since the 14^{th} century. This covers the Early New High German and New High German period of the German language.

Chapter 1

Introduction

We are approaching times where information is foremost accessed digitally. It is quickly getting our preferred way because of the ease of access, the possibilities to quickly find, search and preview digital documents.
Libraries, Companies and private organizations are trying to fulfill the need for digital preservation of and access to books and documents printed before there were digital master copies. Examples range from scientific books to pleadings, treaties and other law documents, from prose to clerical works.
Funding and coordination for this task is besides others provided by the European Union within the Lisbon Strategy by several eEurope action plans and resulting projects like MINERVA[1] (Ministerial Network for Valorising Activities in Digitisation) and DigiCULT[2].
The first step to preserve the works, is the digitization of the paper[3] copies. Usually the copies are scanned to obtain a digital image, which can then be used for archival and access. The access an image allows is rather limited though. The text, which usually is the focus of interest, is not accessible in a way, allowing for further electronic processing. To give full access, the text would need to be made available in a text encoding and the structure of the layout should also be represented in the best case. The image is still useful for verification and further information it contains. An example of how this can look like is given by the

[1] URL: http://www.minervaeurope.org/
[2] URL: http://www.digicult.info/
[3] Or similar materials like vellum, papyrus, leather or palm leafs.

ECHO project[4].

One approach to gain complete access to the texts is keying and while it has very successfully been deployed, it does not scale fast, as people have to be trained, and projects need careful planning, as the work usually is done in lower wage countries, especially China. Chinese typists have some advantages besides the low costs, they have already been trained on complex scripts, as their own script is one of the most complex, and if they don't know the documents language or a similar, they won't subconsciously "correct" the text. The error rates are very low though. A big projects like Grimm and Grimm (2004) gives an accuracy of 99.998 using **double keying** [5] and some postcorrection rules. The postcorrection rules were used to highlight errors which were then manually corrected. A rule might have e.g. highlighted all words ending in *cn*, which probably was *en* in the original text and only a part of the character was missing. The double keying of the ca. 300 Million words cost about 170.000 Euro[6].

Another approach to extract the text is to use an **Optical Character Recognition (OCR)** software. OCR software is rather fast scaling, only limited by the computer power available, and also cheaper than keying and needs less planning. But while the recognition rates of OCR engines on high quality documents printed with todays common Antiqua fonts reaches 99% and more, lacking document quality and Gothic fonts, like we find in older texts, especially in older than 1950, leads to purer recognition rates. This low quality output is known as "dirty OCR" and is sometimes a first step to give access to the text. To improve the quality the OCR output can then be postcorrected manually, which tedious and error prone.

Automatic linguistic post processing can be a valuable step and take some of the burden of the postcorrection task from the humans to the machines. The software can automatically correct a part of the errors and sometimes more importantly highlight the text parts it can't verify. Information one step above the level that OCR makes use of is applied, it operates on words instead of characters. Thus it

[4]European Cultural Heritage Online, URL: http://echo.mpiwg-berlin.mpg.de/home

[5]Double keying means the text was independently keyed in twice and differences manually postcorrected.

[6]According to a presentation in Mai 2007 by Prof. Kurt Gärtner, titled "Digitale Wörterbücher als Grundlage für historische Corpora" at the CIS (University of Munich).

can make use of the redundancy written language offers and recover the original text by looking up the OCR output in electronic **lexica**[7].

The difference between postcorrection of current texts and historical texts are mostly caused by the differences in the language used and differences in the graphic representation of the text on paper. The language differences are a result of language change, or evolution, which has influence on the electronic lexica. The graphical differences are due to change in printing technology, fashion in fonts and the age of the original documents.

In this work we will concentrate on historical German texts printed in modern times. This period begins in the 15^{th} century where the movable type printing press was invented in Europe [8] and *paper* became readily available and replaced vellum, see (Mehring, 1931, p. 9). The end of the period is around the adoption of authoritative spelling rules in 1902, which drastically slowed down the development of the German language and hence is the end of what we consider historical language.

1.1 Historical German Language

Languages are in constant change and German is no exception. What currently is understood as German, the language as spoken and written in Germany today, is called New High German by linguists. And as the name implies, before that there were older forms of High German. Linguists, see (Klu, 2002, p. XL) e.g., divide the High German language into four periods : Old High German (OHG), from about 8^{th} to 11^{th} century [9] , Middle High German (MHG), from 11^{th} to 14^{th} century, Early New High German (ENHG), from 14^{th} century to 17^{th} century, and New High German (NHG), since 17^{th} century.
The use of current OCR software depends on printed documents to obtain high character recognition rates. Periods before ENHG did not produce printed Ger-

[7]We use the lexicon and lexica alone in the sense of lists of words, to distinguish between dictionaries which contain articles for each entry.

[8]Although movable type printing had already been invented in China and Korea it has not become a likewise success.

[9]The beginning of the OHG period has not yet finally been defined. Most theories locate it in the 8^{th} century but some others between 6^{th} to 9^{th} century.

Period	*centuries*
Old High German	*8. - 11.*
Middle High German	*11. - 14.*
Early New High German	*14. - 17.*
New High German	*17. - 21.*

Table 1.1: Overview of the periods of the High German language.

man texts, as movable type printing and enough paper were not available. Therefore also OCR postcorrection is probably only sensible for ENHG and NHG.

The division into the these two periods by the linguists is based on language features like vocabulary, grammar, orthography and pronunciation. For OCR post correction at the current state the important features are vocabulary and orthography. Therefore it is possible that the periods are more sensible for OCR postcorrection.

1.1.1 Early New High German Orthography

Early New High German has no strict orthography[10] and there are no official or authoritative rules. In reading and writing spread to nearly all classes in contrast to the previous periods, where only clerics and higher nobility could read and write. The literary language shifted from Latin to German, which opened up the way for the other parts of society.
The writers and printers had three basic resources to base their writing on: their pronunciation, Latin and other writers and printers. While the Latin influence, of the written language most educated writers knew, and the other writers had a harmonizing influence, pronunciation, because of the many dialects and the different possibilities to realize them graphically, lead to variations, especially in spelling. Spelling variations are one of the most prominent features of ENHG texts.
Orthography followed in part the rule “Schreibe, wie du sprichst” (“Write as you speak”), which means, that the writers tried to write after their phonetic imagination, Müller (1990) devoted a book to this problem. This lead to a heavy influence of the dialect of the writer on the produced text.

[10]Orthography rules over the allowed sequences of characters including whitespace and punctuation marks, see (Stetter, 1994, p. 567).

In contrast to this, authors and printers, at least the more advanced, tried to write in a more abstract language, that was not bound to any of the dialects and which was pronounced as it was written, see Moser (1987). Some already argued for "Speak as you write".
The first teaching book for writing Early New High German that survived is dating back to 1525 and has 12 pages, see Doede (1950). It is not the only one from the 16^{th} century and might very well not be the first. Even those early works began to address the problem of dialects and argued for a higher level of literary language, see (Götz, 1992, p. 101). But their main problem was to teach writing at all. Because of the general low quality of the writing ability good boilerplates were also missing. Even most printers reprinted texts in lower quality than the masters. Learning by example influenced the written language. Early Linguists, so called "Schreibmeister" ('Masters of writing'), recommended examples to follow-up. They often named the original texts by Martin Luther, the Imperial Chancery of Maximilian I and the best printers.

1.1.2 New High German Orthography

The beginning of the subsequent New High German period is marked by a strong heading to an authoritative orthographical set of rules, which consequently continues the harmonisation trend in the previous period.
A big step in this period came with the publication of an influential dictionary, Adelung (1774), and its later editions. Some books that were published shortly before and their editions after the publishing of this dictionary document its strong influence, as many have been edited to follow Adelung's work in their later editions.
This was followed up by two orthographic conferences in 1876 and 1901 that tried to find a set of explicit rules to represent the workings behind the common, rather harmonized language of these days. Based on the rules devised by these conferences Konrad Duden (1829 - 1911) published a dictionary that became the authoritative source. The dictionary became known as the "Duden" and is still maintained by the Duden Redaktion, which publishes revisions about every five years to reflect the development of the German language.

Since then, some unsuccessful reforms were tried but none of them had much impact. The most recent reform from 1996 is the first that could indeed succeed especially after it was reworked from 2004 to 2006. But since some bigger publishers begann to reuse the previous rules, this might not have been finally decided.

1.2 Sources for Historical Scans

Many libraries have started to produce digital images of the books and documents in their possession, that are unrestricted by copyright, which in Germany means, books older than 70 years, as German law ends copyrights 70 years after the death of the author.
Often each title or collection of titles, under a certain topic, like law texts from the reign of some king, is digitized by a different project. The projects often maintain their own websites and present their contents in different ways, usually in reduced quality to keep image sizes comfortably for web browsing.
Links to these projects can, e.g., be found on the websites of the institutions, taking part in the projects, or on special maintaining link collections. Some example are:

Digital Collection of the Bavarian State Library http://mdz1.bib-bvb.de/~mdz/, with currently 22538 titles.

VD16 http://www.vd16.de, a directory of prints in the German speaking parts of the 16^{th} century with links to digitizations if available.

VD17 http://www.vd17.de, a directory of prints in the German speaking parts of the 17^{th} century, with 676.300 sample pages from the listed titles.

FREIMORE FREIburger Multimedia Object REpository, http://freimore.uni-freiburg.de/, by the University of Freiburg, which also contains other media content, besides scans of historical texts.

ZVDD Zentrales Verzeichnis Digitalisierter Drucke, http://www.zvdd.de/, a central directory listing site, offering digital historical prints.

Most libraries also offer digitizations on request. But like paper copies these must be payed individually. At the Bavarian State Library, for example, color scans cost

0.80 Euro per page. For OCR application color scans of at least 300DPI should be requested.

1.3 Editions

For a long time historical texts were accessed via **editions**. In general editions are books compiled from historical documents. Therefore they are published much later than the original. This usually has some advantages. They usually are printed, not handwritten, and the font might already be Antiqua, which is better for OCR, instead of Blackletter. The material is younger and more of the original prints are available, so that it is easier to find a well preserved.
Editions are often based on handwritings and prints which are hardly accessible in the archives of the libraries, until some scholar finds them and thinks of them as interesting. Finding can be difficult, as most handwritings have not yet been deciphered and thus not much relevant information is in the catalogues of the libraries, or sometimes they are only included with wrong information. Even more adventurous to uncover are texts on reused material, where e.g. an older text is hidden under a newer one, or leafs that where packed into other books.
To make the often hard to decipher handwritings accessible to a wider audience editions are reconstructed and published in print. Often more than one, sometimes many versions, in different conditions, are used. A very productive century for editions was the 18^{th} century were many editions were published.
Many editors tried to translate the text to more up to date German and diverged from the sources. Sometimes they translated the text to the current standards at their time or sometimes they "fixed" only the worst offences or sometimes even invented a special language, e.g. resembling what they imagined a standard language at the time of the author could have been.
Therefore, when postcorrecting editions one has to consider whether it is written in the language of time of the original by the author, the language of the editor or a mixture of both, which can be difficult. On the other hand they make OCR viable as they usually provide far better conditions, like font and document quality, than the original documents.

Figure 1.1: Picture of the Manuscript B. of the Zimmerische Chronik scanned by wikisource from Gunter Haug und Heinrich Günter: Burg Wildenstein über dem Tal der jungen Donau, Leinfelden-Echterdingen 2001, p. 38.

An example for such handwritings can be seen in Figure 1.1, which shows the Manuscript B that is available of the Zimmern Chronicle. A sample page of the 1881 edition by Karl August Barack can be seen in Figure A.1 in the appendix. More information about the Zimmern Chronicle can be found in Section 5.3, where it used for evaluations.

1.4 OCR Postcorrection

Optical Character Recognition (OCR) software tries to extract from images the text a person could read on them and outputs the text in a text encoding, e.g. ISO-8859-1 encoding or UTF-8, a Unicode encoding.
In a first step the image is divided in boxes that might contain characters. In the second step the character in the box is recognized. A simple approach might take the part of the image in the box and compute a distance to samples of characters in a data base. E.g. it could recognize it as the character in the data base which has the most overlapping black pixels.

OCR software works far from perfect and many characters are misrecognized as different characters, or not at all. The OCR output has to be validated and corrected if necessary. Postcorrection can be done manually by a person or semi-automatic, that is when software helps the person with the correction, or automatic, that is when software runs unsupervised without a person.
Each of these modes requires a different amount of time, persons need to spend for the task. The most time is spent when the task is done manually and it's also error prone when the corrector is native in a similar language - his mind will "correct" similar spellings to what the one is used to.

1.4.1 Semi-automatic Postcorrection

Semi-automatic postcorrection improves over manual by adding the human correctors e.g. by underlining or otherwise visually highlighting of the questionable words. The software that realizes that is usually running the encountered words against a lexicon, a list of correct words, to validate them. Words not in the lexicon are marked and eventually a correct one or more, similar to the unknown, are suggested as correction candidates. The person can then choose from the list, or override it with different word. This is very similar to the spell checking software built into editors and office suites. And indeed in Section 6.2, an integration of historical lexica, explored in Section 3, into the OpenOffice suite's spell checking backend is presented.

1.4.2 Automatic Postcorrection

While semi-automatic postcorrection can be very helpful, this work is about automatic linguistic postcorrection. Automatic postcorrection needs no human interaction and runs stand alone. The big advantage is that it can be run on large data and as long as it improves the correctness of the digitized texts, manual or semi-automatic postcorrection done after it, will benefit.
The automatic postcorrection process can be divided into the following phases:

- Tokenization, where the OCR output is split into words.
- Lookup, where the correction candidates are chosen from the lexica.

- Correction, where a correction candidate is possibly chosen from the list

The central data base for postcorrection are therefore the lexica. Because of this a main task of this work will be devising lexica that are useful for postcorrection of historical tests.

1.4.3 Errors

W$^{\mathrm{Orig}}$ / text$^{\mathrm{Orig}}$ The original word / text on paper.

W$^{\mathrm{Orig}}$ / text$^{\mathrm{OCR}}$ The word / text as (mis)recognized by the OCR software

W$^{\mathrm{Corr}}$ / text$^{\mathrm{Corr}}$ Postcorrected word / text.

Given these definitions three types of postcorrection errors can be defined. If WˆCorr = WˆOCR $\neq$ WˆOrig, postcorrection leaves an OCR misrecognition uncorrected. This can happen under two conditions. If W^{OCR} is contained in the lexica, it is assumed to be correctly recognized and therefore not corrected. This is called *false friend* error. The second condition is the missing of $W^{O}rig$ in the lexica and that a plausible correction can not be found, so the word is left uncorrected.

WˆCorr $\neq$ WˆOCR = WˆOrig Postcorrection mistakenly corrects for the worse.

WˆCorr $\neq$ WˆOrig, WˆCorr $\neq$ WˆOCR, WˆOCR $\neq$ WˆOrig Postcorrection gets correction wrong.

Chapter 2

Special Problems of Historical Texts

OCR on and postcorrection of historical documents and texts shows a number of problems that do not show at all, or only alleviated, when dealing with contemporary documents and texts. Some problems can only be dealt with in the OCR software, e.g. adding recognition capabilities for historical special characters, and some can or have to be dealt in postcorrection.
The problems can be divided into two main categories, graphic problems and graphemic and lexical problems. From the two kinds of problem categories, graphical problems tend to have the bigger influence on the recognition rates of the OCR and graphemic and lexical have the bigger influence on postcorrection.

2.1 Graphical problems

The category of graphical problems deals with problems and particularities of historical documents and texts on the document and character level. That means, the quality of the documents, the fonts, the graphical representation of the characters, the character set and its realization.

2.1.1 Document Quality

Document quality often has a very strong effect on the OCR recognition rates. Darker parts of the document make it hard to recognize the fonts and damages can even erase characters or make them hard to read. Documents from the Early

New High German period have often kept a remarkable quality, but hundreds of years still show. The good paper used at the time is usually unbleached and a lot darker, more brown than white. Therefore prints from that period generally have a lower contrast than todays papers, see Figure A.4 in the appendix for a typical document, showing the darker paper.
Historical documents might have survived adventurous situations and even if they were well preserved, missing parts, incomplete pages, unrecognizable and fading characters must be expected. Information that is not there can only be guessed, if reconstructed at all. But if only some characters of a word are damaged or missing, postcorrection might very well be able to guess a correct reconstruction, because of the redundancy written language features. A damaged character in a trigraph like *sch*, as in *s˜hwert* where originally *schwert* was printed, for example, e.g. can often be reconstructed correctly.
The more regular damages could either be handled with handcrafted rules or statistically derived weights. Such handcrafted rules were e.g. used after the DWB, see Section 3.1.2, was double keyed.
Many documents show *broken characters*, that means the parts of the character are not fully connected, which is due to the age of the documents. Some documents show this in nearly every character. OCR is known to deal suboptimal in this case and recognizes e.g. *iiber* instead of *über*. Another similar common damage is a missing horizontal strike in an *e* , leading the OCR to recognize it as a *c*. This not only a problem of OCR though, humans can often not recognize such damaged characters either.
More severe damages are outside the intent of OCR postcorrection. More specialized techniques for reconstruction of some of the damages though, has outperformed humans, e.g. in the reconstruction of the *Epic of Gilgamesh*.

2.1.2 Fonts

While today the most wildly used fonts are Antiqua based, which was invented as early as the 15^{th}, it only became dominating in the recent 200 years and even later in Germany. Before that Blackletter font families with many subfamilies were more common. The earlier types were handmade and it became a profession

and art to make them. Therefore the prints can be much more varying even when printed with the same font. This makes it harder for OCR in general and if the OCR is trained on some text, it might help less than expected on another text.

2.1.2.1 Font Families

The German DIN 16518 for classification of types specifies these families for Blackletter type:

a. Gotisch (Gothic)

   Example:

   𝔄 𝔅 ℭ 𝔇 𝔈 𝔉 𝔊 ℌ ℑ 𝔍 𝔎 𝔏 𝔐 𝔑 𝔒 𝔓 𝔔 ℜ 𝔖 𝔗 𝔘 𝔙 𝔚 𝔛 𝔜 ℨ 𝔞 𝔟
   𝔠 𝔡 𝔢 𝔣 𝔤 𝔥 𝔦 𝔧 𝔨 𝔩 𝔪 𝔫 𝔬 𝔭 𝔮 𝔯 𝔰 𝔱 𝔲 𝔳 𝔴 𝔵 𝔶 𝔷

b. Rundgotisch (Rotunda)

   Example:

   𝕬 𝕭 𝕮 𝕯 𝕰 𝕱 𝕲 𝕳 𝕴 𝕵 𝕶 𝕷 𝕸 𝕹 𝕺 𝕻 𝕼 𝕽 𝕾 𝕿 𝖀 𝖁 𝖂 𝖃 𝖄 𝖅 𝖆 𝖇
   𝖈 𝖉 𝖊 𝖋 𝖌 𝖍 𝖎 𝖏 𝖐 𝖑 𝖒 𝖓 𝖔 𝖕 𝖖 𝖗 𝖘 𝖙 𝖚 𝖛 𝖜 𝖝 𝖞 𝖟

c. Schwabacher

   Example:

   𝔄 𝔅 ℭ 𝔇 𝔈 𝔉 𝔊 ℌ ℑ 𝔍 𝔎 𝔏 𝔐 𝔑 𝔒 𝔓 𝔔 ℜ 𝔖 𝔗 𝔘 𝔙 𝔚 𝔛 𝔜 ℨ 𝔞
   𝔟 𝔠 𝔡 𝔢 𝔣 𝔤 𝔥 𝔦 𝔧 𝔨 𝔩 𝔪 𝔫 𝔬 𝔭 𝔮 𝔯 𝔰 𝔱 𝔲 𝔳 𝔴 𝔵 𝔶 𝔷

d. Fraktur

   Example:

   𝔄 𝔅 ℭ 𝔇 𝔈 𝔉 𝔊 ℌ ℑ 𝔍 𝔎 𝔏 𝔐 𝔑 𝔒 𝔓 𝔔 ℜ 𝔖 𝔗 𝔘 𝔙 𝔚 𝔛 𝔜 ℨ 𝔞 𝔟 𝔠 𝔡 𝔢
   𝔣 𝔤 𝔥 𝔦 𝔧 𝔨 𝔩 𝔪 𝔫 𝔬 𝔭 𝔮 𝔯 𝔰 𝔱 𝔲 𝔳 𝔴 𝔵 𝔶 𝔷

e. Fraktur-Varianten

The **prototypographs**, invented around the year 1450, printed in Gotica, Textura, Rotunda, Bastarda, Gotico-Humanistica and also in the Antiqua, which most of todays dominating font family are based on.
Early on types became associated with text categories:

a. Rotunda for latin

b. Bastarda for german text

c. Antiqua for humanistic literature

d. Fracture for Martin Luthers reformation texts

2.1.2.2 Decorated Initials

Another peculiarity is the very frequent use of picture like decorated glyphs for the initial letter of a chapter sometimes even of paragraphs. Here is an Example of such an initial:

The picture resembles the letter basically but can even be colorized and be embedded in drawings of high complexity. The character after such an initial can be capitalized. In the OCR output these initials often leads to one character missing, where decorated initials are used in the text. Postcorrection must be careful not to mistakenly correct the following capital character, allowing for words with two consecutive capitals, like *DEmnach*.

2.1.3 Capitalisation

In contemporary German the rules for capitalisation, that is, when to use a Capital letter at the beginning of a word, are rather strict. All nouns are capitalised and all words after a sentence introducing punctuation mark. Capitalisation in historical German is varying strongly.
According to (Bergmann and Nerius, 1998, p. 779ff) the first letter of the following parts were usually capitalized in texts from period from 1500 to 1710:

- heading

- paragraph
- running title
- text
- sentence, following a punctuation mark

Capitalisations was also used but less strictly:

- after a decorated initial[1]
- at the beginning of sentences, not following a punctuation mark
- at the beginning of subordinate clauses

For nouns the period shows a strict development in capitalization, that might have started a little earlier and is not completely finished at the end. At the beginning about 60% of proper names are capitalized and only about 5% of the appellatives. While capitalization in proper names already reaches 97% in 1560, capitalization in appellatives reaches 92% in 1680, see (Bergmann and Nerius, 1998, p. 832ff.). Besides nouns, certain adjectives, e.g. referring to highly recognized entities such as *Königlich* ('royal') or places, could also be capitalized. Capitalization of adjectives shows very high variations, see (Bergmann and Nerius, 1998, p. 875ff.).

2.1.4 Punctuation marks

Up until the 17$^{\text{th}}$ century sentences and subordinated clauses could divided by:

```
. (dot)
, (comma)
: (colon)
; (semicolon)
```

[1] See also Section 2.1.2

which are common today, but also by the / (Virgule). Abbreviations could also be ended with colon, not only full stop, see also (Moser, 1929, p. 5-10). Note though, the beginning of a new sentence was often only marked by a capitalized word and no punctuation mark was used at all. Many historical texts especially the earlier ones, do not show a sentence structure and punctuation marks are more used to separate shorter syntactical units, without structuring them in more comprehensive structures.

2.2 Lexical and Graphemic problems

Problems that show on the language level are categorized und lexical and graphemic problems. Lexical problems are related to the words, that can be found in historical texts, and graphemic problems with their realization.
The lexical problems in the time since the 14^{th} century can be classified into three problem classes spelling variations, morphological change and variation, historical vocabulary, compounds and whitespace misrecognition, foreign vocabulary, chains of editions, special character set and abbreviations.

Analyzes of these problems on data and how to solve them for postcorrection lexica is given in Section 3, especially in the hypothetical lexica subsections.

2.2.1 Spelling Variations

Spelling variation in historical can refer to two kinds of variation. There are *diachronic variations*, that is spelling varies in different times in history. There are *synchronic variations*, that is spelling varies at one point in history. Diachronic variations are caused by **language change**. Synchronic variations variations are often related to dialects, which often cause *local variations*, or because the writing system allows for more than one graphic realization.
While todays standardized German literary language is not allowing for much variations in spelling, variations can still be found in e.g. family names like *Meier* [ma], *Maier* [ma], *Mayer* [ma], which is an example for different graphic realizations.

Variations are in this Section and the rest of this thesis are given as *old spelling* $\rightarrow$ *modern spelling*. To encode a historical spelling variation with a *t* instead of a *d* in modern spelling, as in *teutsch* vs. *deutsch*, we write $t \rightarrow d$.
Because of the harmonization process that nearly spans all the time from the 14^{th} to the 20^{th} century, generally the diachronic and the synchronic variations are becoming less and less as the 20^{th} century is approached. The greater the distance to it the more spelling variations are found.[2]
In the period before the standardization of German orthography at the beginning of the 20^{th} century, spelling is already very close to what was standardized. The period directly after the standardization is even very close to the standard. One reason for this that some states already had school books following the standard before the standardization and the rest introduced them after the standard was made.

2.2.2 Morphological Change and Variation

The morphological change and variation in historical texts is generally of less visible. While there are a number of changes to the syntax, the changes to the inflected vocabulary are only little. That means, while especially the genitive forms were reduced and replaced by e.g. dative or accusative forms in certain positions, the forms itself are still present in the vocabulary. The following example will illustrate this. In Early New High German one can find *ein Pfund Fleisches*, which would be *ein Pfund Fleisch* in contemporary German. So while this construction lost the genitive, the genitive form *Fleisches* itself is still retained as in *der Verkäufer des Fleisches*. So for the postcorrection lexica the changes like these have no direct influence, although it might have influence on the frequency of the word forms if those were used.
But some morphological change that removed some of the word forms from the vocabulary have been taking place. Most of them are related to the realization of Schwa e in certain derivation suffixes, like *-et*, as in *denket*, which would be

[2]This is of course a generalization because some writers were much ahead of their time, while later writers were well behind their time, or tried to follow different rules. A famous example for the latter is Jacob Grimm.

formed as *denkt* in contemporary German. The morphological rules that were obtained are presented in Section 3.2.2.

2.2.3 Historical Vocabulary

The change in vocabulary in the New High German is not very high and a contemporary spell checking lexicon, like introduced in Section 3.1.1, can often cover more than 80% and sometimes even 98% of the words in a text. The missing words are often compounds that might even be used todays, but even though the spell checking lexicon does not list them. This might be due to the general productivity of the German language when it comes to compounding.

Early New High German shows more vocabulary, that is unknown to a contemporary spell checking lexicon. The fraction of unknown stems is much higher, e.g. *sechen* ('to care'). The compound problem is also increased, in part because of the higher rate of unknown stems, if they are a part in the compound, but also because ancient stems can survive longer in compounds. E.g. in contemporary German *himbeere* is composed of *beere* ('berry') and *him*, which is not a stem in modern German, but can be traced back to Old High German *himi* ('hind').
Many of the compounds in older New High German and Early New High German that are not lexicalized today show similar composition of differently lexicalized parts. For example some of the forming parts are in use today but the compound itself is not. E.g. in *Daß Narrenschyff ad Narragoniam* by Sebastian Brant (1457-1521) we find the compound *winterbutz* ('scarecrow'), which is composed of *winter* ('winter'), which is still in use, and *butz* ('ghost'), which is not. The compound itself is not used any more and in modern German *Vogelscheuche* would be used instead.
Another example is *meistenteil* ('bigger parts'), where both parts, *meisten* ('most') and *teil* ('part'), are still in use, but the composition is not. Instead the contemporary compound would be *großteil*, which is composed of *groß* ('big') and *part* ('teil').

Another problem with compounding especially in Early New High German texts is, that there are no strict compounding rules. Rules that determine when two or more words are written together without separating spaces and what forms are

used, are difficult to find for Early New High German. This makes post correcting whitespace insertions and deletions extremely hard. Before post correction of compounds is possible further research is necessary, e.g. in the upcoming Wanzeck (unpublished).

2.2.4 Foreign Vocabulary

Before German became the dominating literary language, Latin was the most wildly used. Nearly all books had been written in Latin, which was also a spoken language in educated circles and throughout the church. When German began to overtake, it was only natural that many Latin words where still used. For this reason a Latin lexicon was included in out tests, see Section 3.1.3.1.
Up until the 19$^{\text{th}}$ century foreign words have often not been adjusted to German spelling conventions. Most prominently Latin and French words still retained their spelling, keeping a *c* instead of *k*, as in *redaction* vs. *redaktion*, including characters not present in German. Examples are *æ* in Latin words and *é* in French words. Few words kept their foreign characters up until today, like *café*, although derivations, like *cafeteria*, have usually lost them.

2.2.5 Chains of Editions

Many of the historical texts are not from the original author. The texts have often several times been copied (with errors), translated, extracted, summarized, expanded, adjusted to the current or local ruler/spelling/dialect etc. Some texts are the product of a long line of ancestors, or **chain of editions**. Seldom all the ancestors, especially the earlier ones, are known.
Each of the persons involved in the production of the ancestors and the text itself can have left their marks in the language of the texts. Todays we know that e.g. early printers have heavily edited the texts on all levels. The stories might have been changed and orthography adjusted to local conventions. Later when scholars, like from the BLV in the 19$^{\text{th}}$ century, compiled editions from older handwritings, they also often diverted from the originals.
The longer the chain and the more different its parts are, the more change can be expected in the final text. For postcorrection this challenging mixture of influences

Symbol	*Abbreviates*	*Example*	*Unabbreviated*
similar to	us	Christ9	Christus
similar to 8	is	verdammn8	verdammnis
similar to	*final* m	leichna	leichnam
similar to ‘	r, er	wasse’	wasser

Table 2.1: Some abbreviations used, especially in older prints and hand writings.

can have a negative effect. Which parameters, e.g. lexicon, should used? Maybe language profiles, discussed in Section 8.2.1 can help to give an answer in future research. For now, the parameters are either based on the direct ancestors, if the editors seemed to try to follow it, or on the publication date of the edition itself, if the editors tried to “translate” it.

2.2.6 Special Character Set

In historical texts many graphs that are not in use anymore are in common usage. These graphs became less and less used, so that modern German is commonly restricted to 29, not counting those in foreign words.
<abcdefghijklmnopqrstuvwxyzäöüß>
In historical German texts at least the following additional graphs were common:
$<a^e o^e u^e$åůâêîôû>
Also many ligatures were used for *ch* and *ck* e.g. Latin words were also often spelled with *æ* and *œ* ligatures.
Other languages besides German, like English, have similar increasing grapheme numbers in historical times, e.g. in Middle English *Þ* was common.

2.2.7 Abbreviations

Since handwriting dominated the time before and the beginning of the early modern German period abbreviations where much more common than todays and there where many abbreviation symbols, that denoted character sequences. Abbreviations of words were marked by . like todays, but also by *:*, which is not used anymore. From the abbreviation symbols the *&* is e.g. still in use. But in historical texts we also find many others, see Table 2.2.7.

Macron, ¯, or tilde, ˜, above vocals, *ā,ē,ī,ō,ū,* etc., abbreviates a following *m* or *n*, like in *kaīē* (*kainem*). Above *m* and *n* they double the character below, like in *komēn* (*kommen*).
May of these abbreviations are not represented in the Unicode character encoding and therefore not even OCR Software, supporting Unicode, can handle these characters. The limitations of Unicode are further discussed in Section 8.1.1.
Tests with current OCR software suggest that misrecognitions by current OCR software can be classified into the categories:

- reduction to base character, e.g. *e* instead of *ē*
- similar diacritical character, e.g. *é* instead of *ē*
- total misrecognition, e.g. *G* instead of *ē*

2.3 OCR Software

While OCR software for contemporary text and fonts is commonly available, there are only a handful programs available and some are not available to private persons. All of the researched software, shows limitations, especially with older text or low quality input.

2.3.1 Available OCR Software for Blackletter

Following is an overview of the currently publicly known offerings in OCR software.

2.3.1.1 ABBYY FineReader XIX

From ABBYY there is FineReader XIX, a special version of FineReader 7, that includes support for some Blackletter fonts and a lexicon for for “Old German”. As the XIX in the name suggests, the intended historical period is from the beginning of the 19$^{\text{th}}$ century up until 1938. While it is the only commercial offering that is available to private Persons, it is also not cheap, as it costs per page. The smallest offering starts at 343.10 and includes 2.500 pages, or about 0.14 per

page. Larger amounts of paper can be had for about 0.03 per page. Additionally german-dataservice, a company specialized in digitizations, has offerings for 0.03 for larger and 0.04 for smaller amounts.

2.3.1.2 PaperIn Book

PaperIn Book is an OCR product by the swiss ARPA Data GmbH[3] for Blackletter as well as modern fonts. It has very good recognition capabilities for historical text in Blackletter fonts. The program was made available to the CIS (University of Munich) and used e.g. for the recognition of the Allgemeine Deutsche Bibliographie, which is set in a broken font, reaching a recognition rate of nearly 90% of the words.

2.3.1.3 Tesseract

Tesseract is an OpenSource OCR Software developed by HP in the 1990's. The software was later bought by Google and made freely available as OpenSource. It supports broken fonts via training. Trained data for broken fonts is available on the homepage[4] and can be further enhanced.

2.3.2 OCR Software Limitations

The evaluated OCR software is not capable to recognize all of the special symbols and characters found in historical texts. Unicode is one such limiting factor since it does not contain all the necessary characters, see Section 8.1.1, and OCR software can not be expected to work around this limitation. But the real problem is that the available OCR software does not even provide Unicode recognition or output. The software is either limited to characters from the ISO-8859-1 or a similar encoding or a subset of Unicode that does not include many of the characters used in historical texts.

[3] http://www.arpa.ch/
[4] URL http://code.google.com/p/tesseract-ocr/

Chapter 3

Lexica for Postcorrection

The central data basis for OCR postcorrection of historical texts are the lexica. Their quality influences the process most centrally in the lookup phase where the correction candidates are chosen.
Since every token, the post-processing engine encounters is checked against the lexica to know whether it is an actual word or an OCR error. Therefore a missing word in the lexica can lead to a word being classified as OCR misrecognition even though it was perfectly well recognized by the OCR engine. Vice versa an OCR error might be classified to be a correctly recognized word because the lexica contained a word that by chance is identical to the misrecognized string. The latter effect is called **false friends** problem and measured e.g. in Peterson (1986).
In the Correction phase the lexica are critical again because the correction candidates are chosen from lexica or at least derived from them. Further, if the lexica are big, a lot of candidates are likely to be found, which makes it harder to decide which is the best one. On the other hand, if the lexica are too small, they might not contain the best candidate at all.
The **perfect lexicon** for the post-processing therefore would contain all words that the $\text{text}^{\text{Orig}}$ contained and not more. Of course this is usually not the case when OCR is used. But the opposing principles to follow when choosing and/or compiling the lexica are therefore two, one goal should be, making them complete, the other, making them small in relation to the $\text{text}^{\text{Orig}}$.
The currently bigger problem though is obtaining lexica, that contain at least a

Name	*language*	*centuries*	*unique words*	*inflected?*	*variations?*
*Hunspell de_DE**	*German*	*19. - 21.*	*1.500.000*	*yes*	*no*
Hunspell de_DE	*German*	*19. - 21.*	*300.000*	*no*	*no*
DWB	*German*	*15. - 19.*	*300.000*	*no*	*some*
Georges	*Latin*		*63.000*	*no*	*no*
*fr_FR**	*French*	*14. - 17.*	*1.500.000*	*yes*	*no*

Table 3.1: Overview of the basic lexica used.

high percentage words encountered in historical texts. As noted in Section 8.2.1 Language Profiles, summarizing the key features of the text to correct, could help choosing more precise lexica. For now, only lexica for the New High German and Early New High German period were compiled. Recent research suggest more and more that these periods should be further into periods of about 200 year.
Especially in the Early New High German period local **printer languages** showed big differences, mostly influenced by the dialects spoken in the area. So that special lexica for these regional languages in this period could also be valuable for postcorrection.

3.1 Historical Base Lexica

Base lexica are lists of words found in sources like digitized dictionaries or corpora. For the base lexica only sources of confidence are considered, as they shall provide a stable foundation for the hypothetical lexicon described in the next section.
Confidence and availability are the advantages of the base lexica. Their disadvantage is their limited vocabulary, especially when it comes to spelling variations. The main lexicon used is a free, contemporary spell checking. To better cover historical vocabulary that is based on historical lexemes a diachronic lexicon was added. Since linguistic literature suggested a certain amount of foreign vocabulary, two lexica where added for the foreign languages, that had the most influence on the German vocabulary in modern times.
The following table gives an overview of the different base lexica that were used:

3.1.1 Main Lexicon: Hunspell de_DE and de_DE*

The main lexicon, we use, for postcorrection is a general electronic spell checking lexicon for contemporary German. Many tools and programs today include spell checking components and many standalone spell checking programs are available. One of the latest is the Hunspell[1] programm, which is used Open Source and provides the bases for spell checking components in other Open Source programs. The most prominent include the OpenOffice[2] office suite and the next major release, 3.0, of the web Firefox web browser and the Thunderbird E-mail program from the Mozilla Foundation [3].
Hunspell the latest development in the tradition of the UNIX spell checking correction programs *spell*, *ispell*, *aspell* and *myspell*. It supports lexica for many different languages, including German. The Hunspell program and the accompanying lexica was developed by the Budapest Technical University Media Research Centre and is deployed with the office suite OpenOffice[4] and the next major release of the web browser Firefox and the E-mail program from the Mozilla Foundation [5].

A possibly better, although not freely available, alternative would be the electronic CISLEX dictionary, that was developed at the Center for Information and Language Processing (CIS) at the University of Munich.
Hunspell lexica are split up into two files, one file containing a list of base forms with affix flags, the .dic file, and one containing affix rules, the .aff file, to be combined with the base forms. The affix flags in the base forms file after each entry are separated by a forward slash (/). Each character after the forward slash specifies a group of affix rules that can apply to the base form, given that the rules matches.
Entries in the base form file might look like these:

```
verzögern/DIOWXY
verzücken/DIOWXY
```

[1]URL: http://hunspell.sourceforge.net/
[2]URL: http://www.openoffice.org/
[3]URL: http://www.mozilla.com/
[4]URL: http://www.openoffice.org/
[5]URL: http://www.mozilla.com/

```
verächtlich/AC
veränderbar/AU
veränderlich/ACU
veränderlichste/A
```

The affix file contains primarily affix rules to generate inflected and derived words from the base forms. Each rule is separated by a newline and each set of rules is identified by a character. The set is introduced by a header rule that gives information about the type of rule, prefix (*PFX*) or suffix (*SFX_*), the identifier, whether the rules in the set can be combined with other rules, *Y* for yes and *N* for no, and the number of rules in the set.
The affix rules first specify type and identifier, similar to the header rules, then a substring that has to be deleted from the base form, at the beginning for prefix rules and at the end for suffix rules, then a substring that has to be inserted at the same side and finally a Regular Expression[6], that the has to be matched in the original base form for the rule to apply, also at the same side of the base form.
An example of the first lines of a set of suffix rules named *Y*:

```
SFX Y Y 36
SFX Y   n       te          e[lr]n
SFX Y   n       te          [dtw]en
SFX Y   en      te          [^dimntw]en
SFX Y   en      te          eien
SFX Y   n       te          [^e]ien
SFX Y   n       te          chnen
```

Given both of these short samples above, the inflected form *verzögerte* can be computed by combining the base form *verzögern* and applying the first *Y* rule, deleting *n*, which leads to *verzöger*, and appending *te*. This is allowed since the base form has an *Y* flag, the first *Y* rule applies if the pattern *e[lr]n* matches the *ern* suffix of the base form.

To compute the combined lexicon, that contains the inflected forms, Hunspell

[6]It's not a standard Regular Expression, just similar.

provides the unmunch program. We will call the German lexicon obtained with the unmunch [7] program de_DE* and the base form lexicon de_DE.
For the hypothetical lexica the affix file was extended to also cover historical morphology. And the inflected lexicon was again computed with unmunch program.

3.1.2 Diachronic Lexicon: Deutsches Wörterbuch (DWB)

The *Deutsches Wörterbuch* was started by the brothers Grimm, Jacob and Wilhelm, in the year 1854 and continued by other linguists after their death until it was finished in 1960. It is a diachronic dictionary containing about 300.000 keywords and about as many articles, often containing quotations, documenting the use of the words.
In 2004 an electronic version was released on CD-ROM and on the internet by the Kompetenzzentrums für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften at the University of Trier. For this work we gratefully received a copy of the keywords database driving the web version from the Kompetenzzentrum. The dictionary articles were crawled from the web version and used as corpus in testing the lexica.
Two lexica were obtained from the DWB. The first contains all of the ca. 300.000 keywords (lexemes) and the second was compiled from the corpus of the articles and contains about 600.000 tokens.
From the 300.000 keywords less than 20% were contained in the Hunspell de_DE* lexicon and about 21% of the unique tokens from the articles. While this seems to be a low percentage and might lead to questioning of the main lexicon. But tests of the main lexicon with historical texts shows that it covers them relatively well.

3.1.3 Foreign Lexica

Historical German texts show a not insignificant amount of foreign words. The Early New High German period shows strongest influence from Latin, but also from French, Greek and Italien. Latin kept its influence until the 19^{th} century in the New High German period. In the NHG period French won the most influence,

[7] We used version Hunspell 1.1.9.

more than the already declining Latin, but in the end also lost its position in the 20th century, see (von Polenz, 1994, p. 77ff). Greek and Italian never as strong as Latin or French faded even faster than the others. English seems to be the only language that gained influence, while only as late as in 17th century, see (Ganz, 1957, p. 11ff.) and p. 27 - 239, and then only slowly, that still is productive in the 20th century.
For the two dominating foreign languages Latin and French, together responsible for over 80% of the lean words between 1600 and 1800, see also (von Polenz, 1994, p. 77ff), lexica were added for postcorrection. While none of the correction candidates are chosen from these lexica, they can guard against mistakenly correcting correct foreign or lean words. The effect of these lexica might be even stronger with historical languages, since the spelling of foreign words up until the 19th century has often not been adjusted to German spelling conventions.

3.1.3.1 Georges Latin Lexicon

Since many words in historical German texts are from Latin, although the main text is written in German, a Latin lexicon was added to the set of lexica. Because of it's good coverage of the Latin used in German, its scientific repute over the last 200 years and electronic availability Georges (2004) was chosen. While much more work could be done to obtain a more evolved Latin correction dictionary for historical texts, using inflection and handling variations e.g., this work will only make use of this basic lexicon without any modifications.

3.1.3.2 Hunspell fr_FR Lexicon

For the same reasons as the Latin lexicon, see previous section, a lexicon for French was used. A contemporary French lexicon was chosen from the freely available Hunspell lexica, that also provide our main lexicon for German. All base forms containing a ', as in *c'est*, were removed since the tokenizer, used by the postcorrection system, handles the character as a delimiter, see Section 4.1. After this the base forms were expanded to obtain the inflected lexicon (fr_FR*), in the same way as with the Hunspell de_DE* lexicon in Section 3.1.1.

3.2 Hypothetical Lexica

Hypothetical lexica are computed from base lexica using transformation rules. The intended result is a more complete lexicon at the cost of confidence in the generated entries. The transformation rules either encode morphological derivations or variations in spelling. A rule might, e.g., encode a spelling variation like the use of *c* instead of *k*, in realizing the words *korrekt* and *correct*. While *korrekt* is a form found in the base lexica *correct* is not and would be generated for the hypothetical lexica.

3.2.1 Spelling Variations

The problem of the spelling variations in historical texts was described in Section 2.2.1. Spelling variations are handled in the lookup, see Section 4.2, with fuzzy matching and special weights for the varying substrings. The special weights must also be taken care for in the correction phase and what we call patch mode, see Section 4.3.
The variations could also be directly handled in the lexica. To handle these variations in the lexicon, word forms for the variant spelling would have to be generated. For practical reasons this was not done, since even applying few rules can explode a lexicon, beyond what can be efficiently used in memory.
While this part of the hypothetical lexicon is therefore not directly compiled into the lexicon, but encoded as special weights in the fuzzy matching of the lookup. The same results could be obtained by expanding the lexicon, using the weights as generating rules. Therefore the rules set or weights, depending on how one sees it, is handled here.

3.2.1.1 Spelling Variations in New High German

This period still shows some diachronic spelling variations but has mostly lost variations between different writers. Many of the differences to modern spelling are limited to certain morphemes, like *bey* $\rightarrow$ *bei*, that can be found in many derived words like *dabey*, *beyspiel* etc. A very common difference is also *th* $\rightarrow$ *t*. Some variations, like *c* $\rightarrow$ *k* and *c* $\rightarrow$ *z*, are limited to foreign words, like

redaction → *redaktion*.

Some of the variations the nearly 40 different weights used for the New High German period and examples of words, where they apply are shown in the table below:

Variation	*Example*
th → t	rath → rat
ey → ei	bey → bei
d → t	teutsch → deutsch
ph → f	westphälischen → westfälischen
c → z	citat → zitat
c → k	exact → exakt
ss → ß	ausserhalb → außerhalb
ß → s	dieß → dies
ah → a	niemahls → niemals
oh → o	verlohren → verloren
o → oh	obwol → obwohl
u → v	vnd → und
v → f	bevestigen → befestigen
e → ä	erwegen → erwägen
l → ll	gleichfals → gleichfalls
kk → ck	stükken → stücken

3.2.1.2 Spelling Variations in Early New High German

The period from about the 14^{th} to the 17^{th} century shows much more additional variations than the subsequent periods. It also shows especially many synchronic variations. Most of these are due to local differences but they are not limited to them. Even a single text can show many variations, even in the common words, e.g. *frauen* ('women') is also realized as *frawen* in Johanes Virdungs *Practica deutsch*, printed in 1522 by Anastasius Nolt. The variations between different texts and printers or regions and text type can even be more rich. The Munich Corpus for Early New High German (MCF) contains the following spellings for the word *ihm* ('him'): *jhm*, *jhme*, *yhm*, *ym* and *yme*. Proper names can show

similar variations in the same corpus, e.g. *Jerusalem* ('Jerusalem') is spelled: *Hierusalem*, *Hyerusalem*, *Jerusalem* and *Jherusalem*. Often more than one variation is found in a word, e.g. in *abentheurlich* → *abenteuerlich*, both *e* → *ϕ* and *h* → *ϕ* are found.

While a complete list of spelling variations might be impossible to obtain, rules for frequent spelling variations were handcrafted based on the MCF and descriptions of variations mainly in Reichmann and Wegera (1993) and Moser (1929). All rules devised for New High German period are also used for the Early New High German rule set. The rules special to ENHG are about 50, so that for ENHG about 90 rules are used all in all. Some of the rules special to ENHG together with examples from the MCF are listed in the table below:

Variation	*Example*
e → ä	schendet → schändet
ae → e	laeben → leben
t → d	freuntlich → freundlich
i → ei	freudenrich → freudenreich
ich → ig	billich → billig
j → i	jm → im
ck → k	volck → volk
ue → ü	schueler → schüler
w → u	fraw→ frau
y → j	yede → jede
y → i	schyff → schiff

3.2.2 Historical Inflection

The basic lexica do not contain the complete set of inflected forms, that were realizable in historical German. An example for this is *gehet*, which is an historical form in the paradigm of the lexeme *gehen* and is derived as *geht* in modern German. To generate these forms, the affix rules file of the Hunspell lexicon was expanded. The Hunspell lexicon is provided as two files, one containing a base form list with flags that mark the affix rules that apply and a file containing the affix rules. Therefore to generate historical inflected forms, the affix file has to be

expanded. The question is, how to obtain the rules that must be added?

3.2.2.1 Manual Inflection Extraction from Historical New High German

One approach to obtain historical inflection rules is to manually check which words from a historical corpus are not contained in the inflected modern lexica. lexicon. By inspecting these words, someone familiar with the modern language will be able to find historical forms of modern words, that only differ in inflection. A list of unique words that could not be looked up in the Hunspell de_DE*, using the New High German variation weights, was compiled from the New High German corpora.
The only morphological changes with influence on the inflected vocabulary found, were reductions of Schwa *e* [8] in inflection suffixes of verbs and nouns:

1. Reduction of Schwa *e* in verb inflection suffixes, as in *geruhet* → *geruht*
2. Reduction of Schwa *e* in noun inflection suffixes, as in *dache* → *dach*

Based on these findings, the affix file of the Hunspell de_DE dictionary was expanded by about ten rules, to also generate the forms with Schwa. The hypothetical lexicon computed from the was named Hunspell de_DE_19th, because of these suffixes seemed to be only be productive until the 19^{th} century. Besides these small changes, the inflection morphology seemed to be very stable in the New High German period.

3.2.2.2 Automatic Morphology Extraction from Early New High German

Since there is no electronic affix file for Early New High German morphology and it had to be assumed that this period might show more morphological changes, an automatic approach to morphology extraction was tried.
A number of algorithms have been devised to extract the morphological rules of many languages. An overview and evaluation of such techniques can be found in Hafer and Weiss (1974). In recent years these techniques have been applied to many languages , e.g. Arabic in Rogati *et al.* (2003), Lee *et al.* (2003), Finnish

[8] An unstressed *e*.

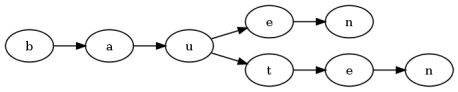


Figure 3.1: A simplified trie as used in Goldsmith's algorithm.

in Creutz and Lagus (2002), Indo-germanic languages in general in Goldsmith (2001) and many in the specific.

3.2.2.2.1 Goldsmith's Algorithm Because of the good support for indo-germanic languages claimed by Goldsmith's and the availability of an implementation of his algorithm, a program called Linguistica, the Goldsmith algorithm was used to automatically extract morphology from the historical corpora. To add further modifications, the algorithm was independently implemented.
The general idea in Goldsmith's algorithm, is to build a trie from all the words in a corpus. The Trie is then searched for nodes with special features. The node must contain a certain number of children, while its parent has only that node as child. A simplified version of such a trie can be seen in Figure 3.1. The path leading to such a special node from the root is then identified as "stem" and the paths leaving from the node to leaf nodes are the suffixes. An example of such signatures is pairs of stem and suffixes is:

$$bau - \begin{cases} -e \\ -en \\ -t \\ -te \\ -ten \end{cases}$$

Stems with the same suffix set are sorted together into, what he calls signatures, that are pairs of the stems with the same suffix and the suffixes, like:
$(stem_1, stem_2, stem_3), (suffix_1, suffix_2, suffix_3, suffix_4)$
The signatures are then filtered by rules that shall ensure that the signature really encodes stems and suffixes, e.g. signatures that only have one stem will be discarded. Further optimizations are then applied to further ensure the quality of the signatures. As the list of signatures was already small enough for manual valida-

tion, after filtering out signatures where either the stem list or the suffix list had less than two members, the further optimizations were not implemented.
The results were less good than Goldsmith's, as it only found a handful of signatures. German shows some features that play against the Goldsmith algorithm like allomorphic stems, which are explicitly not handled as Goldsmith states himself. Further the historical German language with it's variations seemed to have increased this problem. A solution for these problems is given in the next section.

3.2.2.2.2 Enhancing Goldsmith's Base Algorithm by Preclustering The algorithm given by Goldsmith, described in the previous section, is not intended to handle allomorphic stems, as shown by the German language and historical German. Allomorphs are morphemes that are realized as different morphs, e.g. in *baum* ('tree') and *bäume* ('trees'), the two morphs *baum* and *bäum* form are allomorphs. Historical German is especially problematic since each spelling variation in the stem is producing another allomorphic stem.
The Goldsmith algorithm was therefore modified. Instead of working on a trie representing all the words in a corpus, the corpus is first divided into clusters of words within a certain threshold edit distance. Then the Goldsmith algorithm is run on each of the cluster, like each cluster was its own corpus. All the stems found in the trie are then discarded besides the one with the highest number of suffixes. The suffixes of the discarded stems are added to suffixes of the one, that was kept. For example, given a cluster containing the words *baum, baumes, bäume, bäumes, bäumlein* the Goldsmith algorithm finds the two stems *baum-* with the suffixes *-ø, -es* and *bäum* with the suffixes *-e, -es, -lein*. Our modification choses the stem *bäum* and merges the suffixes to *-ø, -es, -e, -es, -lein*. After this has been done for all the clusters, the Goldsmith algorithm continues with sorting the stems together with the same suffix lists to obtain signatures.
The modified algorithm was then used on the Munich Corpus for Early New High German to automatically extract the morphology from the corpus. Looking at at the results the modified algorithm found mostly derivational suffixes, that were not different to contemporary ones. The main difference being again the ones already manually identified for New High German. The following is a snippet of the output of the algorithm, showing signatures containing these suffixes:

```
{(besichtig-, bestetig-, betreuff-, entschuldig-, ...), (-en, -et, -t
{(abkauff-, klopff-, veriag-), (-en, -et, -ten)}
{(meld-, vereinig-, versoehn-), (-en, -et, -ung)}
{(anzeyg-, entdeck-, handthab-, vberreich-, verkünd-), (-en, -t, -unc
{(beklag-, gecreutzig-), (-et, -t, -ten)}
```

Because of the high coverage of the extracted inflection morphology by the Hunspell de_DE_19th, nothing special was done to handle the found morphology. Until further, more in depth research, we assumed the morphology has not changed substancially.

3.3 Evaluation of the Lexica on Historical Corpora

The quality of the lexica for postcorrection of historical texts was measured by their coverage of different historical corpora. First the different corpora that were used are presented and then the coverage of the corpora by the lexica is tested.

3.3.1 Historical Corpora

Corpora are collections of texts that cover certain topics, e.g. there are medical corpora that contain texts that can be assigned under medical topics. Corpora relevant to postcorrection of historical texts for corpora can be classified by three domains time, which describes the historical periods they cover, location, which mostly describes the dialects or printer languages covered, and the text category, which describes the kind of text, e.g. scientific or prose.

Because of the early stage some of the corpora and postcorrection of historical texts itself, the corpora were only divided by the time domain. For the Early New High German period an incomplete and preliminary version of the Münchner Corpus for Early New High German was made available by the Institute for German Philology (University of Munich). For the subsequent period from 1650 to 1800 the GerManC corpus was used, covering currently only news paper texts. To enhance the coverage of the pivotal time between the Early New High German

period and the New High German period, two corpora were compiled from Wikisource texts, one covering the years from 1601-1650 and one to cover the years from 1651-1700. The central New High German period was covered by another corpus obtained from Wikisource texts.
The following table gives an overview of the corpora used, the period they cover and the unique words, they contain:

Short Name	*Main Period Covered*	*Words*	*Unique Words*
MCF	*1500 - 1600*	*110000*	*17000*
GerManC	*1650 - 1800*	*100000*	*19000*
WS_1601-1650	*1601 - 1650*	*13000*	*4300*
WS_1651-1700	*1651 - 1700*	*5800*	*2200*
WS_18th	*1701 - 1800*	*86000*	*12000*

Table 3.2: Overview of the corpora the period they cover, the words and unique words they contain.

3.3.1.1 Münchner Corpus für Frühneuhochdeutsch (MCF)

The Münchner Corpus für Frühneuhochdeutsch[9] (Munich Corpus for Early New High German) by the Institute for German Philology (University of Munich) is a yet only partially published corpus of 63 chosen texts from the Early New High German period between 1350 and 1650. Of the 63 texts elven have already been keyed by trained linguists. They keyed mostly without special characters.
The corpus was carefully chosen from a variety of printers and dialectic regions and also from different text types. Therefore the corpus should show a broad spectrum of the features of the printed language of its period. An overview of the titles with year, printer and location can found in Table 3.3.1.1.

[9]URL: http://demo.fruehneuhochdeutsch.is.guad.de/

Year	*Printer*	*Location*	*Author / Title*
1495	*Gregor Böttiger*	*Leipzig*	Eyn libliche histori von vier Kaufleuten
1516	*Heinrich von Neuß*	*Köln*	Tundalus Ritter
1521	*Pamphilus Gengenbach*	*Basel*	*Eberlin von Günzburg:* Dz lob der pfarrer
1521	*Johann Schöffer*	*Mainz*	Meintzisch hoffgerichts Ordnung zu allen anderen gerichten dienlich
1522	*Anastasius Nolt*	*Speyer*	*Johanes Virdung:* Practica deutsch
1530	*Paul Kohl*	*Regensburg*	*Hans Lutz:* Grundige vnd warhafftige bericht
1540	*Alexander Weissenhorn*	*Augsburg*	Dyll Vlnspiegel
1548	*Gervasius Stürmer*	*Erfurt*	*Kaspar Aquila:* Eyn sehr hoch noetige Ermanung
1564	*Ulrich Neuber*	*Nürnberg*	*Johannes Mathesius:* Leichpredigt
1586	*Johann Scharfenberg*	*Breslau*	Christliche Bekaentnis
1589	*Christoph Rab*	*Herborn*	*Johannes Piscator:* Kurtzer be-richt von des Herren Abendmal

Table 3.3: The eleven texts in the pre-released Munich Corpus for Early New High German.

It includes a database containing linguistic annotations for each token in the corpus. The database contains the following fields for the token, modern translation, historical lexeme, modern lexeme, part of speech, grammatical extra information.

3.3.1.2 GerManC

GerManC[10] is an ongoing project to provide a historical German Corpus for the years 1650 to 1800 by the School of Languages, Linguistics and Cultures in the University of Manchester. While the final corpus will contain texts from eight categories, in its current initial stage only the news paper category is finished.
The newspaper category consists of word samples of about 2000 words from different news papers from five regions (North German, West Central German, East Central German, West Upper German, East Upper German) of three fifty years 1650 to 1700, 1701 to 1750 and 1751 to 1800. All in all it makes up about 100000 words and 19000 unique words.
The texts were obtained by double keying and encoded in the TEI[11] XML format in a third step. The TEI markup contains annotations that mark words as person names, places, foreign words and many more.

3.3.1.3 Wikisource

Wikisource[12] (WS) is a project supported by the Wikimedia Foundation[13], well-known for the Wikipedia encyclopedia, for building a digital text library of books, unrestricted by copyright. Because of this requirement it contains mostly historical books, where the copyright protection expired. It is an ongoing project - but many smaller and some bigger historical texts have been completed. As most projects supported by the Wikimedia Foundation, Wikisource uses the Mediawiki software platform, an elaborate wiki, which provides collaborative editing and publishing over the web, using a special markup language, called wikitext[14].

[10]URL: http://www.llc.manchester.ac.uk/research/projects/germanc/
[11]URL:http://www.tei-c.org/
[12]URL: http://wikisource.org/
[13]URL: http://wikimediafoundation.org/
[14]URL:http://en.wikipedia.org/wiki/Wiki

The books are digitized either by using OCR software or keying and then marked up using wikitext, to encode some of the style and structural information of the texts. The Wikisource policy requires two proof reading cycles by different persons. A Mediawiki based web interface supports the whole process, beginning with the initial upload of the OCR or keying output. Proofreading and correction is usually done in a web browser. The Mediawiki software records each change to the text and allows to retrieve any previous version.

3.3.1.3.1 Problems Using Wikisource Wikis are intended for managing marked up text for presentation in the web. They do not provide interfaces or markup that is intended for representing historical documents, like TEI XML. Storing historical texts in a wiki leads to a number of problems, especially for the use as electronic corpus.

First, each wiki platform specifies its own wikitext dialog, so that often, the only parser understanding the particular markup is the one provided by the particular software, which often is implemented for HTML output only. Which can make extracting text difficult.

Second, wikis like Mediawiki, are structured in articles or pages that are interlinked, like other Hypertext. Historical texts, on the other hand, show the classical hierarchical structure found in paper documents. This is only a little mediated by the categories system. Categories are tags associated with articles. They can be structured in a hierarchy but the hierarchy has to be defined by the project. Wikisource tags all articles, belonging to a book, with the book's title. Articles represent parts of the books, either pages or entries. For example for a book containing poems the associated articles might represent the poems, even so some pages of the book may contain more than one poem per page or some poems might span more pages.
The development of the category system is still at an early stage in Wikisource and many issues are still open. E.g. a translation made in the 20^{th} century in contemporary language of a Dante text from the 14^{th} century might still be tagged as 14^{th} century.

Third, the meta data specific to the content, must also be encoded in wikitext, mak-

ing it difficult to separate the content, meta data and editorial extra information. Relevant meta data for Wikisource content is e.g. the year the text was published. Now while the wikitext, encoding the books content, and meta data relevant to the Mediawiki software is stored in a database, the meta data, relevant to the content, is stored inside the wikitext, which cannot easily be queried from the database interface or use the database's indices.

Fourth, storing the wikitext and all it's revisions in the database, instead of e.g. only the meta data, is also questionable as it bloats the database and makes it slow, while not providing much advantage over storing the wikitext in a filesystem and possibly only the differences for each revision, like many source code management systems like CVS do. The database snapshot from XXX is XXX.

3.3.1.3.2 Wikisource Corpora Since the Wikisource project, especially the German subdivision, tries to digitize texts as close to the original as possible and sensible Wikisource can be a great resource for historical texts. Care has to be taken when compiling corpora for certain periods though. Many of the texts are editions, and, as we discussed in Section 1.3, editions are often heavily normalized and often do not represent the original language of the text. Therefore only original texts have been chosen from wikisource.
For the 17^{th} century two corpora have been compiled, since the the 17^{th} century is divided into two parts. The first 50 years belong to the Early New High German period and the second to the to New High German period. But such a sharp border for a dynamic system like language is seldom telling the whole truth. Indeed many scholars draw different borders, but from the six different periodisations by the spelling criteria analyzed in (Roelcke, 1995, p. 439-441) only one draws no border in the 17^{th} century and only one of the others not in 1650. A thorough analyzes of this century therefore seems to be necessary and so more fine grained corpus was chosen for this century, that covers as many years as possible. Therefore the two copora were compiled from leaflets and nearly for each decade one or more leaflets could be found. Table 3.3.1.3.2 shows the year, word count, code in the VD17 and title contained in the corpus we name WS_1601-1650, that covers the first 50 years. And Table 3.3.1.3.2 shows the same data for the corpus, that covers the second 50 years, and which was name WS_1651-1700. WS_1601-

1650 contains 13763 words and 4307 unique words, WS_1651-1700 5838 words and 2159 unique words.
Finally a third corpus was compiled for the 18^{th} century. Table 3.3.1.3.2 shows the same data as the previous tables for this corpus, that was named WS_18th. This corpus contains 85650 words and 11655 unique words. It was avoided to chose prose for this corpora, as prose often does give a good representation of the language at the time it was written.

Year	*Words*	*VD17 Code*	*Title*
1602	*1544*	*23:328302M*	*Kurtze Beschreibung vnd Erzehlung von einem Juden mit Namen Ahaßverus*
~1620	*385*	*12:204271T*	*Der Mainaidt*
~1620	*612*	*1:089798E*	*Deß gwesten Pfaltzgraf offne schuldt*
1621	*4761*	*39:125065Q*	*Achterklärung über Friedrich von der Pfalz*
1621/22	*733*	*12:666426W*	*Deß guten Geldes Grabschrifft*
1626	*551*	*12:649185T*	*Abbildung/ neben kurtzem Bericht/ welcher gestalt den 15 April. Anno 1626. der Hertzog zu Friedland/ die Manßfeldische Armee von der Elbbrucken zu Dessa abgetrieben/ zertrennt/ und guten theils erlegt hat*
1629	*1116*	*12:668618Z*	*Ware Abbildung deß in Anno 29 Jars den 2 May zu Nurmberg ankommenten Elephanten*
1635	*399*	*14:001601B*	*Newauffgerichtete Verträwliche Brüderschafft eines Frantzösischen vnd teutschen Soldatens*
1640	*886*	*1:088817W*	*Der großmächtige Nasen Monarch*
~1650	*789*	*23:677393W*	*Die faule Haußmagd*
~1650	*399*	*23:647312B*	*Newer Korb voll Venuskinder*
~1650	*528*	*23:659251B*	*Der alten und neuen Manns- und Weiber-Tracht*
1650	*620*	*23:244818C*	*Peinliche Anklag unnd hitzige Antwort eines zornigen Schneiders und warhafften Hirtens*

Table 3.4: Leaflets from Wikisource from 1601 to 1650.

Year	*Words*	*VD17 Code*	*Title*
1660	*494*	*1:091614L*	*Die Weiber-Treu der Frauen zu Weinsberg*
1662	*141*	*23:233013P*	*Auf den Kornschnitt*
1675	*630*	*23:674758S*	*Kurtzweilige Beschreibung des Baurn-volcks jhrer Rockenstuben / vnd was darinnen für schöne Possen getrieben werden*
1680	*304*	*3:320019V*	*Steckbrieflich gesuchte Mord-Brenner*
1689	*868*	*3:651231D*	*Die bestraffte Frauenzimmer Hauben-Mode*
1693	*824*	*3:651630T*	*Eigentliche und warhaffte Abbildung Eines erschröcklichen und grausamen Meer-Drachens*
~1700	*998*	*3:651316K*	*Der Frauen und Weiber Privilegia*
~1700	*976*	*3:632977M*	*Eigendlicher Abris des heiligen Grabes zu Görlitz*
~1700	*414*	*3:658749D*	*Eine Wunderns-würdige Miß-Geburth eines Kindes*
1700	*189*	*23:684018F*	*Verordnung zur Herabsetzung des Bierpreises wegen der guten Ernte*

Table 3.5: Leaflets from Wikisource from 1651 to 1700.

Year	*Words*	*Title*
1746	*17444*	*C. F. Gellerts sämmtliche Schriften. Erster Theil. Leipzig, bey M. G. Weidmanns Erben und Reich, und Caspar Frisch, 1769.*
1753	*556*	*Viehseuchenverordnung Ravensburg 1753*
1769	*31459*	*Die Ursache des Einschlagens vom Blitze*
1784	*2619*	*Beantwortung der Frage: Was ist Aufklärung*
1787	*24584*	*Nöthige Belehrung und Warnung für Jünglinge und solche Knaben, die schon zu einigem Nachdenken gewöhnt sind*
1797	*8687*	*Ausführliche Vorschriften zur Blitz-Ableitung an allerley Gebäuden*

Table 3.6: Corpus for the period from 1701 to 1800 obtained from Wikisource.

3.3.2 Coverage of the Corpora by the Lexica

The performance of a lexicon for postcorrection can be measured by the coverage of test corpora. To get an estimation of the performance of the base lexica and the hypothetical lexicon, the coverage of each of the lexica by each other and of the historical corpora by each of the lexica was measured.

The percentages of the not inflected lexica covered by each of the other not inflected lexica is shown in Table 3.3.2. As can be seen, the coverage is as at maximum 18%. And as could be expected the highest coverage is between the contemporary spell checking lexicon, Hunspell de_DE, and the diachronic, DWB keywords lexica. Interestingly the number is lower than might have been expected. There could be two reasons for this. Either historical, language covered in the DWB, shows a high percentage of lexemes out of use today, or the DWB covers many low frequency lexemes. More insight on this can be gained below when the corpora are test. All other not inflected lexica do not share more than 7% with others.

More interesting than the coverage between the lexica is the coverage of historical corpora, shown in table Table 3.3.2 for the ENHG corpora and in Table 3.3.2 for the NHG corpora. In difference to the above table the inflected lexica are used, if available. The row *not words* shows the number of tokens in the corpus that either are punctuation, numbers or words of length one or two, because these will not be handled by postcorrection.

	Hunspell de_DE	*DWB keywords*	*Georges*	*Hunspell fr_FR*
Hunspell de_DE	*100%*	*18%*	*5%*	*7%*
DWB keywords	*17%*	*100%*	*1%*	*2%*
Georges	*1%*	*0%*	*100%*	*1%*
Hunspell fr_FR	*2%*	*0%*	*2%*	*100%*

Table 3.7: Percentages covered of the not inflected base lexica in the first row by the others in the first column.

	MCF	*WS_1600-1650*
not words	*9%*	*7%*
*Hunspell de_DE**	*50%*	*59%*
*Hunspell de_DE_19th**	*51%*	*60%*
Hunspell de_DE_19th + ENHG weights*	*81%*	*85%*
DWB keywords	*22%*	*35%*
Georges	*1%*	*2%*
*Hunspell fr_FR**	*3%*	*4%*

Table 3.8: Coverage of the Early New High German corpora in the first row by the basic lexica, inflected if available, in the first column.

In the NHG period the contemporary spell checking lexicon, Hunspell de_DE, covers between 70% and 78% percent of the corpora, the lexicon with added historical inflexions always outperforms it by one or two percent. The best lexicon though is the one, additionally using special weights to handle spelling variation in the NHG period. Discounting the *not words*, only six to ten percent of the corpus is not covered by this lexicon. For ENHG the performance of this lexicon is similar, while the other perform worse. This indicates, that ENHG mostly differs because of the high variation rate from NHG.
Of course that means that still about 10% of the words in the corpora are not covered. Since neither domain specific lexica nor proper name lexica were used, a certain percentage of the corpora is expected to not be covered, this is a low number.

	GerManC	*WS_1651_1700*	*WS_18th*
not words	*9%*	*9%*	*11%*
*Hunspell de_DE**	*70%*	*73%*	*78%*
*Hunspell de_DE_19th**	*71%*	*74%*	*80%*
Hunspell de_DE_19th + NHG weights*	*82%*	*81%*	*83%*
DWB keywords	*26%*	*47%*	*38%*
Georges	*2%*	*2%*	*1%*
*Hunspell fr_FR**	*4%*	*4%*	*3%*

Table 3.9: Coverage of the New High German corpora in the first row by the basic lexica, inflected if available, in the first column.

Chapter 4

Automatic Postcorrection

Automatic postcorrection tries to correct the text digitized by the OCR phase. This is possible because of the assumption that the text that was digitized, the $\text{text}^{\text{Orig}}$, had certain features. Especially it is assumed that the text was written in a certain language. The more knowledge about the language and vocabulary used in the original is available, the better postcorrection can work. Of course this is only true as long as the text is not random. If the OCR was used to recognize random strings, then no (linguistic) OCR postcorrection would be possible.
Automatic postcorrection, as the phase after the text has been digitized by the OCR Software, is itself divided into phases. The first phase is tokenization, where a tokenizer parses the output format of the OCR software and breaks it down to tokens, that might represent words. Which is followed by the lookup phase, where a candidate list for correction is composed of words looked up in the lexica. And finally the correction phase where tokens, that are determined to be misrecognitions are corrected, given that a convincing correction candidate was found in the lookup phase.

4.1 Tokenization

The input to the postcorrection process is the output of the OCR software. Depending on the OCR software, different output formats are available. Common formats that are interesting to postcorrection are plain text, containing the least

extra information, but commonly available, XML/HTML mark up, containing extra information but often encoded in a way special to a single OCR program, and ORF, a text format with added information.
The postcorrection process acts on words and so the OCR output must be **tokenized**, broken down to tokens, that in this case usually represent words or punctuation marks, unless the OCR software's output format already marks the word boundaries.
A very basic tokenizer will simply split the input on delimiting characters or character sequences in a regular language, like Regular Expressions. More advanced might even be based on a context free grammar.
We only used to a very simple tokenizer that identifies words separated by a space or the following delimiter characters:
*tn./,;:”’()[]{}?!_“ˆ|&=<>%§«»#**

Section 2.2.7 might give hints for more advanced tokenizers with better handling for the many abbreviations found in historical texts and Section 2.1.4 discusses punctuation issues in those texts.

4.2 Lookup Using Fuzzy Matching

In the lookup phase the tokens provided by the tokenizer are looked up in the lexicon to gather a list of correction candidates for the correction phase. To get this list of candidates fuzzy matching is used on the lexica. In a simple setup, all the words from the lexicon within Levenshtein distance one might be chosen as candidates. More advanced setups might use weights for common OCR misrecognitions such as a *c* for an *e* or to handle spelling variations found in historical texts.
Fuzzy matching is used to match strings that are only similar instead of identical. To measure similarity, a distance measure is applied. A basic string distance measure is the **Levenshtein distance**, Levenshtein (1966), which measures distance by the edit operations on the character level applied to one string needed to obtain the other string. The allowed edit operations are character deletion, character insertion and character confusion, each with a cost of one. That means for the two words *und* and *unnd* one insertion operation is needed, an insertion of *n*. So the

		f	*r*	*a*	*w*	*e*	*n*
	0	*1*	*2*	*3*	*4*	*5*	*5 + y*
f	*1*	***0***	*1*	*2*	*3*	*4*	*4 + y*
r	*2*	*1*	***0***	*1*	*2*	*3*	*3 + y*
a	*3*	*2*	*1*	***0***	*1*	*2*	*2 + y*
u	*4*	*3*	*2*	*1*	***x***	*x + 1*	*x + 1 + y*
e	*5*	*4*	*3*	*2*	*x + 1*	***x***	*x + y*
n	*5 + y*	*4 + y*	*3 + y*	*2 + y*	*x + 1 + y*	*x + y*	***x***

Table 4.1: Matrix for Levenshtein-Distance between *frawen* and *frauen* with special costs x for u $\rightarrow$ w and y for n $\rightarrow$ ø with $1 > x >= 0$, $1 > y >= 0$. The least cost path is in bold.

distance is one for this pair.

4.2.1 Brill and Moore Model

Some modified algorithms use different costs for different operations and/or different characters involved. More advanced algorithms work with statistical weights for the operations and/or allow the operations to work on more than single characters. Brill and Moore (2000) generalized these modifications to use **weighted substring to substring operations**. The length of the substrings is only limited by the length of the involved strings. The first substring being empty ($\emptyset$) models an insertion of the second substring. And the second being empty a deletion.
A common implementation[1] of the Levenshtein Distance algorithm family uses a dynamic programming approach which involves a matrix to record the costs. An example of such a matrix for the strings *frauen* and *frawen* with special costs x for the replacement operation $u \rightarrow w$ and y for $\emptyset \rightarrow n$ and $1 > x >= 0$, $1 > y >= 0$ can be seen in Figure 4.2.1.
With Brill and Moore's model the example above could have even more specific weights. Instead of weights for the single character operations $u \rightarrow w$, there could be weights for operations on substrings representing the diphthongs used: $au \rightarrow aw$.

[1]For a number of sample implementations, that make use of a matrix, see e.g. URL: http://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=159325879#Implementations

4.2.2 Computing the Weights

The step from three edit operations and uniform weights, as in the original Levenshtein distance, to statistical weights comes at the price of having to obtain these weights. The method proposed by Brill and Moore needs handcrafted pairs of training samples. For these pairs the edit distance is computed first in the way of the original Levenshtein distance. Additionally the operations needed to get from one string to the other are recorded.
For the example above that would yield the following operations:
$f \rightarrow f$, $r \rightarrow r$, $a \rightarrow a$, $u \rightarrow w$, $e \rightarrow e$, $n \rightarrow n$.
The operations are then aggregated to operations on substrings up to a limit given as parameter *maxSubstringLength*. For *maxSubstringLength = 3* the following additional aggregated operations are the result:
$fr \rightarrow fr$, $ra \rightarrow ra$, $au \rightarrow aw$, $ue \rightarrow we$, $en \rightarrow en$
$fra \rightarrow fra$, $rau \rightarrow raw$, $aue \rightarrow awe$, $uen \rightarrow wen$
A table is maintained that holds the frequencies of the operations over the set of training pairs, depending on the position of the operation in the two strings: beginning, end or none of both of the strings.
From this frequency table the statistical weights are then computed by relating the frequency of each operation to the frequency of all operations with the same first substring. E.g. if $au \rightarrow aw$ was counted four times and $au \rightarrow ar$ once the resulting weight for $au \rightarrow aw$ would be $\frac{4}{5}$.
Training pairs consisting of a Early New High German word and its corresponding contemporary, or translation if you will, are available in the Munich Corpus for Early New High German's database, see Section 3.3.1.1. Form these training pairs weights were computed that should give low weights to edit operations needed to obtain the contemporary entry in a lexicon, given a ENHG variation.

4.2.3 Special Correction Weights

As discussed in Section 4.2.2, the lookup in the dictionary can use fuzzy matching with statistical weights to obtain a list of candidates that.
Besides typical OCR misrecognitions that can be found in different OCR software, some of them default to output a special character, in the case where the

recognition was not possible. PaperIn Book e.g. often defaults to ˜ to mark such positions. Therefore a special weight table was added for such characters, when such a software was used. The weight tables allows transition from the special character to any other character with a very low weight.

4.3 Correction in the Presence of Spelling Variations

The main task of OCR postcorrection is the correction of the words misrecognized by the OCR Software. The decision whether a word is a misrecognition or valid is decided based on correction candidate list compiled in the lookup phase. If the token is determined to be misrecognized, a correction candidate is chosen from the list.
In a simple setup the correction candidate in the list with the lowest edit distance to the token, that is postcorrected, is chosen and if token and the candidate is identical, no correction is done. But if the candidate differs from the token, the token is replaced by the candidate in the output.

For texts that contain high variation rates that make it hard to provide a lexicon including all the variations, a special **patch mode** for correction was developed. The patch mode only corrects the parts of the word that have the highest costs conversion costs, as provided by the distance function for the lookup (see also section 4.2).
For example word$^{\mathrm{Orig}}$ = *ewangelisten*, word$^{\mathrm{OCR}}$ = *ewangclisten* and the best match in the dictionary is *euangelisten*. Without patch mode the word would be miscorrected to the word found in the dictionary, *euangelisten*. So not only the misrecognized *c* would be corrected but also the correctly recognized *w*.
But since the distance function from the lookup provides not only the distance cost but also a list of substring to substring conversions and their cost. Using a *maxSubstringLength* of 3 we might get:[2]
ˆe>ˆe = 0.00, u>w = 5.98, ang>ang = 0.00, e>c = 9.91, l>l = 0.00, ist>ist = 0.00, en$>en$ = 0.00

[2]Remember that ˆ and *$* are used to mark beginning and end of string.

The cost of the variation, *u>w = 5.98*, is much lower than the cost for the misrecognition, *e>c = 9.91*. Therefore if only those substring to substring conversions are applied, that are above a certain threshold, say 8.0 for the example, a new string is created, that is made of parts of both strings and we call the patched word. For the example this would be *ewangelisten*, which is exactly the text$^{\mathrm{Orig}}$ and hence a correct correction.

4.4 Implementation

An automatic post-processing system was implemented[3]. As input the OCR output is expected and the output is a corrected version. The toolchain is written in Java and many parts are implemented in a modular way, that makes it easy to hook in modules that implement new ideas.

4.5 Modular Correction Methods

For each encountered token each postcorrection module is given the chance to handle the token, given that no previous module has already handled the token. Each method returns a status from the following:

UNKNOWN No status was set.

DECLINED The method did not handle the token.

SKIP The method recommends to skip the token without further processing.

CORRECT The token could be verified as a correct word.

CORRECTION The token was not a word and corrected.

INCORRECT The token is probably not correct.

The modules are split into two categories validation methods and correction methods. Validation methods are responsible for validating the tokens and decide whether further processing is helpful.

[3]The software is available at URL: http://www.splashground.de/~andy/OCRC

The following validation methods have been implemented:

Skip Numbers returns SKIP for tokens that match a regular expression representing Roman, e.g. *XVI* and Arabic numbers.

Skip Whitespace returns SKIP for tokens that match a regular expression representing whitespace characters, e.g. a tab character.

Skip Punctuation returns SKIP for tokens that match a regular expression representing punctuation characters, e.g. a dot.

Skip Too Short returns SKIP for tokens with string length below a certain threshold, defaults to ≤ 2.

Skip Capital Words returns SKIP for tokens beginning with a capital letter. This module is not used by default, but can be useful for texts, where mostly proper names are capitalized and postcorrection underperforms.

Exact Lexicon Trie returns CORRECT if the lower cased token is found in the dictionary. This method can be used multiple times with different lexica.

Fuzzy Lexicon Trie returns CORRECT, if the lower cased token is found by fuzzy matching in the lexicon. This is useful if variations are are not expanded in the lexica. See the next section for details.

Fuzzy Corrector Trie returns CORRECTION, if the lower cased token is found by fuzzy matching with special correction weights within a certain edit distance.

As indicated by the name the lexica are held in a trie data structure, as proposed by Brill and Moore (2000), describing the fuzzy matching technique, that is described in Section 4.2.1. To speed up the fuzzy matching over lexica as large as the ones described in Section 3, containing more than a million words, an approach as described in Oflazer (1995) is used, where the edit distance for each path in the trie is only computed once, and reused for words with the same prefix.

4.6 Command Line Usage

The usage in the command line is rather complex as all the dictionaries and weights have to be given as arguments. Therefore wrapper scripts were added, that provide a sensible choice for each of the two periods. The one wrapper script is called *run-corrector-nhg* for the New High German Period and *run-corrector-enhg* for the Early New High German Period. Files that are to be corrected are given with the *-f* argument. The output of the program is the corrected text. An example usage would be:

```
run-corrector-nhg -f ADB.txt
```

Chapter 5

Groundtruth Data and Preleminary Results

Groundtruth data in the OCR context is the text that should be recognized - the $\text{text}^{\text{Orig}}$. This data is necessary for test setup to determine the quality of the output of the OCR software or postcorrection system.
Groundtruth data can be obtained via keying, OCR together with manual postcorrection, or a digital master, from which the output was produced. Since a digital master is of course usually not available for historical texts, which were produced in pre-digital ages.
If a historical text has been digitized though, then it can be reprinted to produce new ground truthh, then it can be reprinted to produce new ground truth. This is especially useful when the text would be interesting to test but the original scans or paper copies are lost, or of bad quality. Producing groundtruth data this way, is of course inferior to the use of authentic material, but it can be a great resource to test the postcorrection quality, besides the artificial setup.
For the preliminary tests on the groundtruth the handcrafted weights to handle spelling variations, described in Section 3.2.1 were used. Depending on the period the text was from either the Early New High German ones or the New High German ones were chosen. For all test the following validation lexica were used: Hunspell de_DE_19th*

5.1 Dyll Vlnspiegel Reprinted

The photo copies of the original print, that were available for this text, were of low quality. None of the OCR programs performed well enough to make postcorrection sensible.
Therefore the first 20 pages of the electronic version of the *Dyll Vlnspiegel* text from the MCF was reprinted with a current laser printer in a modern font. The reprint was then scanned and the Ocrad OpenSource OCR was used to reread the text. This OCR was chosen because it shows high misrecognition rates even on modern fonts. The intention was to simulate higher misrecognition rates of OCR on historical documents.
Results:

	Original	*OCR*	*Postcorrection*	*Corrections*
Correct Words	*8386*	*6342*	*6377*	*351*
Percentage	*100%*	*75.6%*	*76.0%*	*4.2%*

While the text as read by the OCR shows 6342 correct words, the postcorrected text shows 6377 correct words, an improvement of 35 words. This shows that postcorrection can improve OCR read text, even when the recognition rate is as low as 75%.

5.2 Allgemeine Deutsche Biographie

The Allgemeine Deutsche Biographie (ADB), ADB (1875), is a biography, a Who is Who, of about 26.000 Germans, in 45 volumes. It was published from from 1875 to 1910 and is printed in a Blackletter font. The scans were high-quality scans made available by the Bavarian State Library, consisting of 20 sample pages, covering the introduction to the ADB and none of the bibliographic entries. A sample scan can be seen in Figure A.3 in the appendix.
Results:

	Original	*OCR*	*Postcorrection*	*Corrections*
Correct Words	*5906*	*5284*	*5298*	*114*
Percentage	*100%*	*89.5%*	*89.7%*	*1.9%*

Nearly 99% percent of the words in the 20 sample pages are contained in the de_DE_19th, when using the NHG spelling variation weights. The spelling variations found in the sample pages are listed here:

Count	*Variation*
1	t → ˆd
1	i → ie
1	ph → f
1	ß → ss
2	c → z
2	mm → m
2	ä → a
3	wol → wohl
4	c → k

5.3 Zimmerische Chronik

The *Zimmerische Chronik* (Zimmern Chronicle) excessively describes the family history of the von Zimmern, a noble family situated in Meßkirch in the south-west of Germany, see Jenny (1959). It was written from about 1559 to 1566 by Froben Christoph von Zimmern (1519-1566) and is preserved in two handwritings, Manuscript A and B, which were the foundation of the later editions by Karl August Barack, which was used as groundtruth. The manuscripts do not provide a title and *Zimmerische Chronik* is only the name it is commonly known as. Manuscript B contains 1581 pages, a sample can be seen in Figure 1.1 in Section 1.3.

The master for the scanned images was the second edition of Karl August Barack published in 1881. The edition is printed in an Antiqua. The scanned images are provided by the University of Freiburg, Germany, a sample page can be seen in Figure A.1 in the appendix. The groundtruth data is available from Wikisource and Tesseract was used to produce the OCR read text.

Results:

	Original	*OCR*	*Postcorrection*	*Corrections*
Correct Words	*7183*	*5841*	*5873*	*343*
Percentage	*100%*	*81.3*	*81.8%*	*4.8%*

Again postcorrection showed some improvement. This time though the maximum allowed distance for choosing a correction candidate had to be lowered from the default value of 1.2 to 1.0.

5.4 Schedel'sche Weltchronik

The so called Schedel'sche Weltchronik, also Nuremberg Chronicle or liber chronicarum, written by Hartmann Schedel and printed in 1493 in Nuremberg by Anton Koberger, Schedel (1493), is a lavishly made chronicle of the world as understood at the time. It is one of the valuable early prints, called incunabula, which is attributed to those books printed in Europe before the year 1501. It is printed in Schwabacher Blackletter font and colorized.
A scan of a facsimile edition of an original print was double keyed by Wikisource. To obtain the text$^{\mathrm{OCR}}$, Tesseract was used on high quality scans on scans from a private version of the facsimile edition. A sample page from this facsimile edition can be seen in Figure A.2 in the appendix.
Since the intention of the editors from de.wikisource.org was not to produce perfect groundtruth data but a readable edition, they chose to normalize some characters to modern standards. On the introduction page[1] they note the following derivations from the source:

- long s is normalised to *s* (round s).
- The following abbreviations are expanded :
  - ā ē ī ō ū ⇒ an/am en/em in/im on/om un/um
  - ñ ⇒ en/nn
  - vñ ⇒ vnd

[1] URL: http://de.wikisource.org/w/index.php?title=Schedel%E2%80%99sche_Weltchronik&oldid=194302

  - d' $\Rightarrow$ der
  - &c $\Rightarrow$ etc.

- Characters written above other characters (e.g. a) are encoded as superpositions (like a^e).
- Some editorial annotations were added.

One sample page has been chosen from the few pages on Wikisource that were marked as "corrected". The page with the highest recognition rates was chosen. Judging by the low recognition rates, incunabula, that is early prints from before 1500, might still be out of reach of contemporary OCR.

Results:

	Original	*OCR*	*Postcorrection*	*Corrections*
Correct Words	*1168*	*384*	*385*	*48*
Percentage	*100%*	*32.9%*	*33.0%*	*4.1%*

Even though the correction rate was as low as 32.9% postcorrection did not turn the result to the worse and even improved it.

Chapter 6

Semi-automatic Postcorrection

Recognition rates reached by OCR even together with automatic postcorrection are not perfect and can often be far less than 100%. While for some tasks like information retrieval such inaccurate digitized texts already can be valuable, as demonstrated by Google Book Search[1] for example, many tasks require higher recognition rates of near 100% and therefore need to be manually or semi-automatically post corrected.

6.1 Graphical Postcorrection Editor

As neither OCR nor postcorrection on historical texts reaches perfection, an essential part of the text digitization is the correction by hand. To visualize the correction results and to also to bring the benefits from the researched automatic postcorrection to semi-automatic postcorrection, a graphical postcorrection program[2] was prototyped, that interfaces the postcorrection tool chain. The program gives visual feedback about the automatically corrected parts of the text and highlights words that the postcorrection toolchain can neither validate nor correct.
The program works with Ocrad description files, developed by the programmers of the Ocrad OpenSource OCR software. The format give information about each recognized character and its position in the original image file. The editor can use

[1] URL:http://books.google.com/
[2] The software is available at URL: http://www.splashground.de/~andy/OCRC

this information together with the original image file to present the recognized text and above each recognized character an image of the original character is shown. The format can be converted to and from other similar formats, like Tesseract's box format. Therefore the editor can produce training data for the Tesseract OCR. While manually correcting the recognized text, the image above each character shall help to avoid errors, especially those that speakers of contemporary German might make by unintentionally normalizing the words. There are some advantages compared to a paper copy lying on the desk or an image beside an editor on the screen. For once, characters are brought to focus instead of words, therefore the normalizing effect of the human correctors is alleviated. Second, the time to move the eyes from the character image to the typing position is far less, which is more comfortable and fast and also alleviating the normalization effect.
Sometime it can be helpful to see the whole, unsegmented line of text in the original image. Below the editor is a second pane that is vertically divided, which shows the whole unsegmented, original image. Focusing the currently edited part has not yet been implemented.

6.2 Historical Spell Checking Lexicon for OpenOffice

As one of the Base Lexica was based on a lexicon in the Hunspell format, which is also used for spell checking in the OpenOffice office suite, it was a simple step to provide back a hypothetical lexicon for the office suite. Figure 6.2 shows the OpenOffice user interface when post correcting with one of the Hypothetical historical lexica provided by us.
Many of the people working manually postcorrecting historical texts are much more comfortable using an office suite for the task than with the interfaces provided with OCR software. The built-in spell checking user interfaces in the office suites usually put red lines under the unknown words, to quickly identify these. More detailed information cannot directly be shown, e.g. there are no lines of different colors than red available to encode different confidence levels of the postcorrection software.

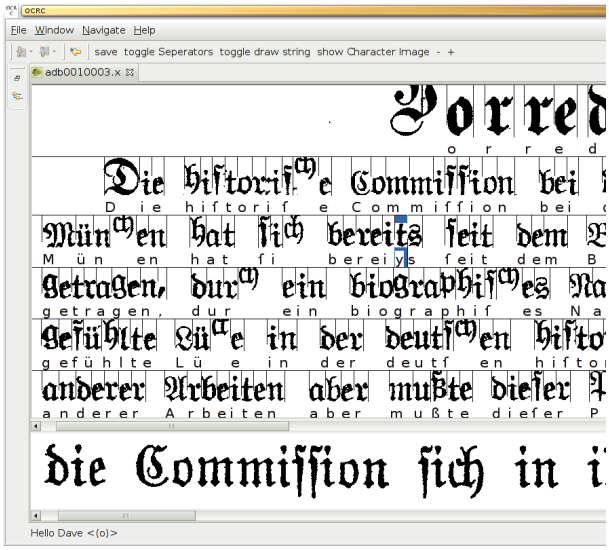


Figure 6.1: A custom postcorrection program based on our research.

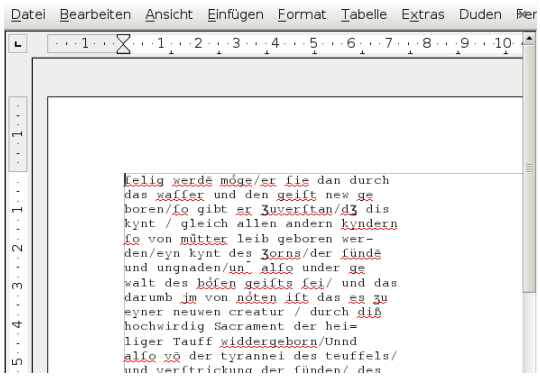


Figure 6.2: OpenOffice using one of the Hypothetical historical spell checking lexica.

Although the Hunspell format itself, offers only limited capabilities to rank the candidates. It only allows to list substring confusion pairs, that are then preferred in the correction candidate list, e.g. one could list pairs that are often confused by the OCR software like *ii* for *ü*.

Chapter 7

Tools

A number of programs and tools were written to support research in postcorrection of historical texts and obtaining and converting useful data. To provide an insight in technical details, that might be relevant to interpret the numbers and statistics of the previous chapters, and to give a short overview to interested users, the tools shortly discussed.

7.1 Tool for Extraction of the Characters Used

Since historical texts show a lot of special characters, one question that should was to be answered was about the characters that were used in the corpora. Therefore a little tokenizer was written that prints each Unicode character on a line. Counting the occurrence of each character and printing the statistics can be achieved using basic UNIX tools like *sort* and *uniq*. To reformat the statistics as columns, to make the output better readable, the UNIX tool *column* can be used.
The whole command line to analyze a text, *text.txt*, would then look like this:

```
run-tokenizer-character.sh test.txt | sort | uniq -c | column
```

The output might then look like:

```
  10 -     219 E     363 S    3815 h    1328 v
  15 0     170 F     108 T    4912 i     956 w
```

19	1	261	G	22	U	166	j	4	x
15	2	290	H	243	V	474	k	285	y
8	3	129	I	264	W	2731	l	832	z
6	4	128	J	1	Y	1612	m	405	ß
14	5	237	K	57	Z	7801	n	202	ä
8	6	135	L	3717	a	1699	o	1	æ
4	7	178	M	1325	b	254	p	300	ö
7	9	108	N	2532	c	6	q	434	ü
240	A	54	O	3522	d	4918	r	4	
186	B	112	P	10920	e	3837	s		
158	C	2	Q	1343	f	4214	t		
310	D	228	R	2139	g	2121	u		

7.2 Tool for HTML to Text Conversion

None of the tools available for HTML to text conversion was able to handle all of the special characters encoded in historical HTML sources, like the web version of the Deutsches Wörterbuch, although the encoding follows a simple standard scheme. Therefore a little tool based on the *hpricot* HTML parsing library available for the Ruby scripting language was written to convert HTML to plain UTF-8 encoded text.
The tool requires only one argument, the HTML source, e.g. *test.html*:

```
html2text.rb test.html
```

7.3 Tool for Extraction of Texts from Wikisource

As described in Section 3.3.1.3.1 dealing with Wikisource as a corpus can be problematic especially because meta data queries are difficult to realize. One might, e.g., want to extract all texts belonging to a category like *17. Jahrhundert* (‘17$^{\text{th}}$ century’) that also belong the category *fertig* (‘ready’), which means, that the two phased correction of the text is completed. While there is a web tool provided by Wikimedia, which is still in development and can be unstable at times,

direct queries to a local installation of the downloadable Wikisource database are preferable in many situations. So a tool was written that allows for such queries. The basic usage is:

```
mediawiki_get category "17. Jahrhundert"
```

The first argument specifies the type of query category or revision. Category queries fetch all articles in the specified category and revision queries fetch certain revisions of an article from the database, e.g. the first revision or the latest.

Chapter 8

Conclusion

Postcorrection of historical texts has to deal with a number of problems. The main problems were outlined and it was shown how a number of problems could be solved.
At first, one of the biggest problems seemed to be the lack of an inflected historical lexicon, as the coverage of e.g. the MCF by the contemporary spell checking lexicon was lower than 30%. But after using a fuzzy matching approach in the lexicon lookup with handcrafted substring to substring edit weights to handle spelling variations, the coverage improved up to more than 80% for all periods covered. Taking into account that no domain specific lexica and no proper name lexicon was used and that often about 10% of text where tokens, like punctuation marks, that are not compiled into lexica and cannot be postcorrected anyways currently, we showed that a contemporary spell checking lexicon can very well handle historical language, provided that some inflection rules are added and the variations are dealt with, e.g. by using fuzzy matching and handcrafted weights.

A custom toolchain for OCR postcorrection of historical texts has been implemented to test the lexica and the handcrafted variation rules in a realistic setup. Then preliminary, unoptimized results where obtained on real data. It was shown that even in this preliminary setup postcorrection shows improvements in the recognition rates over the unprocessed OCR output.

Because neither automatic postcorrection nor OCR does not work perfectly and might never, because e.g. spelling errors and very rare spelling variations close to

more frequent forms in spelling, or even spelling errors in the historical texts itself, semi automatic or manual postcorrection will always be needed, but is especially important as long as the recognition rates on broken fonts are still below 90%.
Therefore a postcorrection editor to support human postcorrectors has been prototyped that run the automatic postcorrection toolchain in the background. The automatic corrections are highlighted, so the human can check their reasonability. Words that the toolchain is unsure about, but refrains from automatically correcting, are also highlighted in a different color, for the same reason.
Not all problems could be solved though. For some it was not possible to find a solution since current software technology outside of postcorrection is responsible, e.g. the limitations of OCR software and Unicode with respect to the special characters found in historical texts. These problems are described in the following section.
Some other problems have not been solved or not completely been solved because they require much more scientific research and more detailed knowledge about historical language and ways to apply it. This is discussed in Section 8.2 below.

8.1 Problems with todays technology

8.1.1 Unicode is not enough

Unicode [1] a standard for character encoding. It provides for each character in the standard a unique number. In the current version 5.0 it contains characters used in languages all over the world.
Even though contains about 100.000 characters, it still misses some of the characters encountered in historical texts. A big problem currently is that some of the abbreviations and diacritics are not represented in Unicode at all, see also Section 2.2.7. There are some initiatives by scholars under way to have these included in Unicode. E.g. there is the Medieval Unicode Font Initiative[2](MUFI) trying to get special characters found in medieval Latin included in the Unicode Standard. In the beginning of Unicode there was a strong opposition to include such “rarely”

[1]URL: http://www.unicode.org/
[2]URL: http://helmer.hit.uib.no/mufi/

used characters and symbols in a standard mostly sponsored by industry and governments, but in recent revisions of the standard many of the requested characters have found there way in, although the process for a handful of characters can take years.
But even if representation in current Unicode is possible, the support in current applications is often limited. Many diacritics e.g. can only be constructed using Combining Characters (Consortium (2006)), which almost no software, including web browsers, handles correctly. Many characters are also not present in many fonts.

8.1.2 Better integration with OCR software.

Currently postcorrection is a process disconnected from the OCR software. While OCR software integrates some post-processing itself, this is only very basic, especially for historical languages. The interfaces offered by OCR software are all very basic. Often only the dictionary the OCR software uses can be changed. No OCR software, we evaluated, included an interface for advanced techniques like the approximate search with substring to substring edit operations with different weights as described in section 4.2.
On the other hand there are promising Open Source[3] OCR engines available that can be modified to include these techniques, at the cost of having to study the projects source code. Examples would be ocrad, gocr, Tesseract and OCRopus, which includes a version of Tesseract improved by the German Research Center for Artificial Intelligence (DFKI). Tesseract is especially interesting because it can be trained.
Another way to better integrate with an OCR engine would be to build one from LeCun's lenet5, see LeCun *et al.* (1998), an Opensource Neural Net that was used for the NIST contest in handwritten digit recognition and showed very good recognition rates. It is also Open Source and packaged with the Lush programming language, a Lisp dialect. Since it has been proven that it can deal with such difficult recognition problems, it might also be able to deal better with Gothic

[3]See URL: http://www.opensource.org/ for more information on Open Source. Basically it means that the programming source coded for programs is provided and modifications to it are allowed, allowing for easy customization.

fonts and might even be extended to handle handwritten texts, which could give access to the many historical texts only available as such.

8.2 Future Work

While it was shown that basic OCR postcorrection for historical texts is possible and shows results, that are better than plain OCR, many open problems that can improve the quality of postcorrection are still open to further research.

8.2.1 Language Profiles

As already discussed in Section 3 the perfect lexicon would solely consist of all the words of the $\text{text}^{\text{Orig}}$. While this seems impossible, a lexicon as close as possible to text is desirable. One idea to accommodate this is to first analyze the text^{OCR} very closely, gaining as much information about the text as possible and the noisy conditions allow. Interesting fields of such a **language profile** might be: fonts, year of first edition, publishing year, dialect, type of text etc.

8.2.2 Whitespace Misrecognitions

Many of the OCR errors are whitespace misrecognitions that lead to either a word split, like *ge spilt* instead of *gespilt*, or word merges, like *derHund* instead of *der Hund*.
One idea to solve this problem could be based on shifting the tokens to recognized from words, delimited by whitespace and punctuation marks, to lines which are delimited by newline characters. Lookup and Correction would still work similar but string distances would be computed for lines, that is sequences of words. Therefore the computational complexity might not allow such a system to still run in acceptable times.

8.2.3 Postcorrection of Keyed Texts

While some of the techniques we used is mostly useful for the postcorrection of OCR output, many can also be applied directly, or with little adaption, to keyed

texts. As shown in Section 6.2, the lexica compiled could be e.g. be used as spell checking lexica in the OpenOffice office suite. More research needs to be done whether typical typing errors in historical texts can also be automatically corrected or whether most of some are hard to detect false friend errors.

Appendix A

Sample Documents

A number of sample scans of some historical books follow.

[1] Wo die Cimbri erstlich gewonet und was lender sie eingenomen, auch wie sie die Römer angriffen, die mermals geschlagen, doch letstlich von inen überwunden worden.

Es ist zu wissen, das vor jaren die Cimbri ain mechtigs, streitbars und sighafts volk gewesen, auch vil grofser, gevärlicher krieg ain lange zeit gefüert, mechtige königreich und lender eingenomen, diselben mit gewalt erobert und ingehabt haben. Ire vätterliche, angebornne erste sitz und wonungen sein gewesen in cimbrischen Chersoneso, so ain landtschaft teutscher nation gegen mitternacht, die weit in das mer sich zeucht und zwischen dem brittannischen und teutschen Oceano gelegen, diser zeit Hollstain und Schleswig genannt wurdt. Aufs diser landtschaft sein ir ain wolgerustes hör sambt weiben und künden zogen, ungevärlich hundert jar vor dem fürtreffenlichen Homero, so do gewesen nach anfang der welt zwaitausendt neunhundert und vierzig, und vor der gepurt unsers seligmachers tausendt neunhundert und fünf jar. Die haben in kurzer zeit ain grofsen thail Europæ und Asiæ durchstraifet; dann demnach sie in ain grofse macht erwachsen, ist ain aufruor und burgerlicher krieg under inen, wie gewonlichen beschicht, so ain reich am höchsten schwebt, entstanden, derhalben ain grofse und streitbare anzall volks sambt irem könig Liddamio `ire angebornne wonungen verlassen, newe länder zu erobern. Dise haben nachmals unsäglichen schaden vast allen septentrionalischen lendern zugefüegt; sein dergestalt bifs an den meotischen see und Pontum exinum komen, alda sie den Chersonesum und vast alle lender, darumb gelegen, eingenomen, ain mechtige statt darin gebawt, Cymericum genannt,

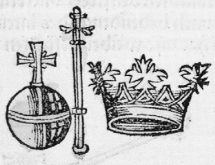
*

[1] seitenzahl der handschrift B. 1 Wo] der anfang der chronik bis gegen ende von s. [8] fehlt in A. 18 grofsen] hs. gorfsen. 27 exinum] d. i. euxinum.

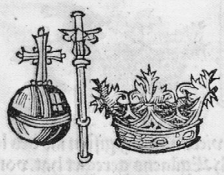
Figure A.1: Sample page of the 1881 edition of the Zimmern Chronicle by Karl August Barack.

Von dem thurn babilonie

NEmroth ein rys vñ sterckst der hannd ward nach absterbē noe sēnes vranherē mit begirde zeherschē ange zündet also das er dē gewalt der herschūg an sich bracht. des selbē reichs anfāg hat sich angehebt in dē feld Sēnaar. daselbst het der selb allergetürstigst vnd redsprechendlichst man ein versamlūg. vnd das er die mēschē vō gottes forcht abforderte so riet er in das sie zigel machtē vñ mit fewer kochten vnd einen hohen thurn paweten des gipfel oō höhe bis an dē himel rürete. gleich als würdē sie dar durch steigē in dē himel. do sie nwn dē thurn paweten vnd sich mit grosser irer hohfart wider got erhuben do hat got iren frefel vnd stoltzmütigkeit mit diser ainer ainigen straff also geslagen das die zwayundsibitzig völker die alda zusamē komen vnd auß den dreyn sünē Noe abgestigen waren vnd alle ain einigs gezūng hetten in soul zerstrewūg der zungen getrennet wurden das eyner des andern stym̄ nit versteen mochte. Dise zesamenblasung oder pūntnus ist also entlöset das sie auff allē amplick der erden zerstrewet wurden. Aber an welchem end diser thurn gestanden sey ist wenig menschen offenbar. Sie sagen im anfang bey dem fluss eufrates sey ein edle kauffmās oder gewerb stat der Caldeer Baldach genāt do selbst sprechen die inwoner das nit verr vō der stat ein grosser stainhauff vnd zurüdung gesehen werde. dohin die menschen vor scharpffen felsen vnd vergifften thiren nit komen mogen. vnd maynen das der thurn daselbst gewessen sey vnd von dannen alle ding in ir stat auß babilonia gefürt sind wordē. Beda spricht diser thurn sey. M. c. lxxiij. schritt hoh gewesen. vnd von weylen vnd zu weylen an der höhe ein eingezwengt. vñ diser thurn wardt genent Babel. das ist zerteilung oder schendung. dañ wywol das gezung alles erdreichs daselbst gewest ist. so hat doch der herr sie auff den amplick aller gegent zerstrewt.



DAs reich Scytharū hat in der gegent gein mitternacht anfang genomen. do hat erstlich geregirt Thanay. vō im ist also genant Thanays der gross berümbt fluss der in die pfürschen (die man paludē meotidem haist) fleüsset. vnd diss gegent wird von dem selben fluss thanais getaylet. Ein tail bleibt in Europa. ō ander streckett sich in asiam. der tail europe endet sich gein Traciam. vnd gepiret wenig thier. vnd bleibt vom fluss vnuerletzet. aber der tail der in asia gein dem aufgang raichet hat mancherlay volcks. vnd gemainlich alle solche cytische völker füren pogen so sie rewten vnd neren sich mit des pflugs sünder der thier die sie iagend fahen. vnd wiewol diss reich das eltist ist. ydoch nach dē es eins groben volcks ist so wirdt es vnder den vier fürnemlichen vnd vordern reichen nit gerechent. Aber diss cytisch volck hat nye einigem menschen im streyt gewichen. Sunder es hat darumb den kōnig der persier geiagt. Cyrum todgeslagen. Syphrionam des grossen alexanders herfürer abgetilgt. Vesoim den kōnig der Egyptier land mit allem seinem her vñ kriegßzeūg abgetriben vnd in die flucht gebracht Asiam zu dreyen malen mit streyt ernydergeworffen vnd im vil iar zinsper gemacht. Auß den selben scytiern sind vil außgegangē die grosse ding geübt haben. Zuerst Amazones die hohberümbtē weiber. durch die scheinpere taten in kriegen bescheen sind. Bactriani vñ parthi sind auß ine komē. Auch der groß Attila vñ ander. der pannoniā vndertrücket vnd aquileiā vmkeret vnd in teütschen landē vil verwüstung machet ist von in abgestigen. Zelior bis ō hungern kōnig der wider den kaiser Justinianū auffstund hat auß scythia seinen vrsprung. die hungern Cathelani vnd alle gothi sind auß den scythiern entstanden. Auch die dani vnd türcken. Diss gegent hat auß Mago des Noe enicklein anfang gehabt. vñ das volck ist grob das weder rechts noch gleichs heltet. Slangen vnd abgotterei hat es geeret. eingewickelt mit vil vnordenlichen begirden.



DAs reich der assyrier in der gegent des aufgangs hat im. xxv. iar des lebens Barnch (als Eusebius sagt) seynen anfang genomen. das dañ vor den andern allen das treffentlicher vnd berümbter gewesen ist. vnd. M. ccc. ij. iar von dem ersten kōnig Belo bis auff Sardanapallum den letsten kōnig vnder. xxxvij. kōnigen gereichßnet hat. Assyria ist ein gegent Asie die sich vom aufgang an den fluss eufrates vnd vom nidergang an vnser meer vnd an Egipto endet. aber von mitternacht hat sie armeniam vnd Capadociam. vnd von mittentag arabiā vñ diss ist Syria.

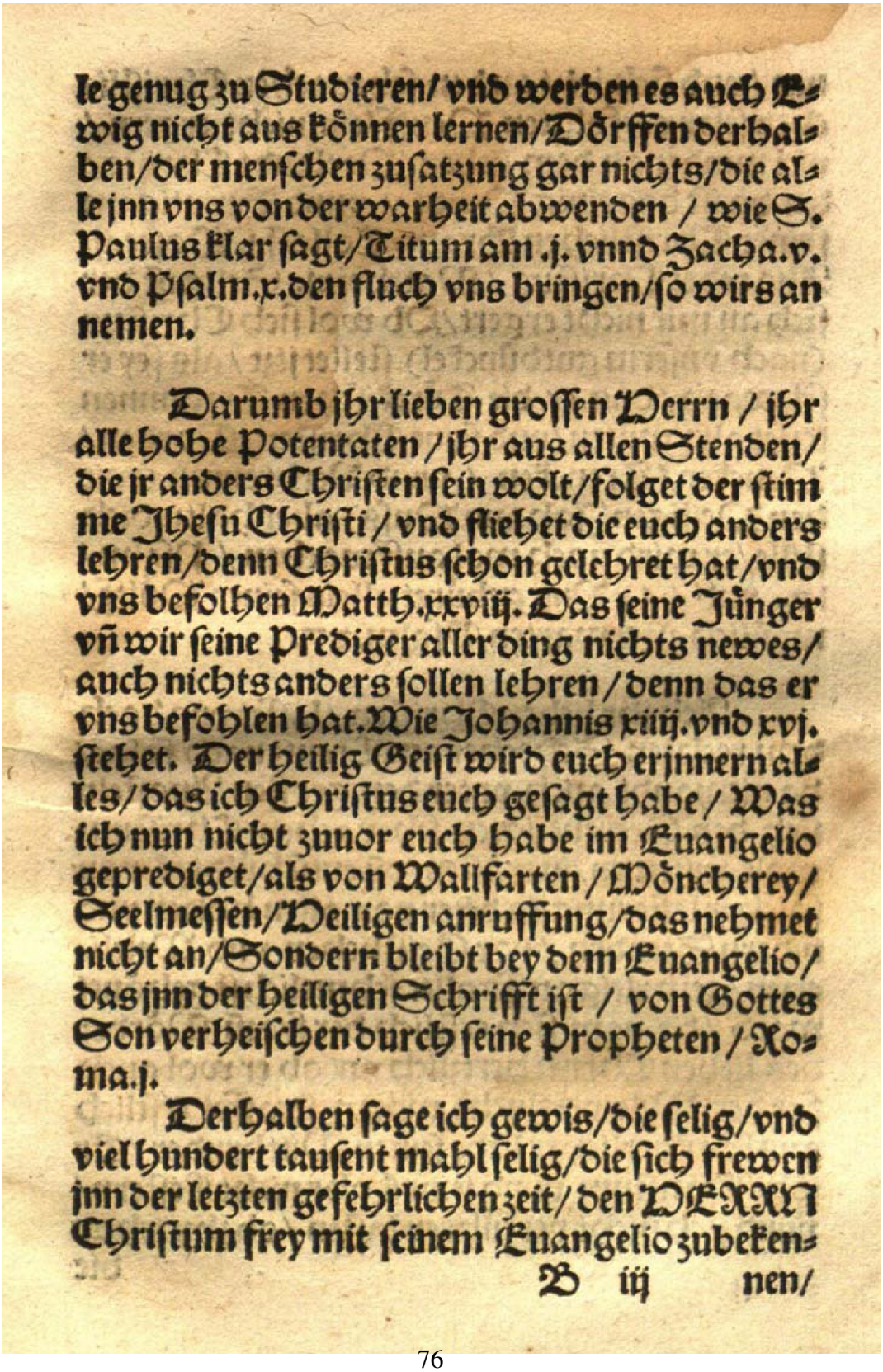
Figure A.2: Sample page of a facsimile edition by Taschen, 2001, of the Nuremberg Chronicle written by Hartmann Schedel and Stephan Füssel.

Vorrede.

Die historische Commission bei der kgl. Akademie der Wissenschaften in München hat sich bereits seit dem Beginn ihrer Arbeiten mit dem Gedanken getragen, durch ein biographisches Nachschlagewerk für Deutschland eine längst gefühlte Lücke in der deutschen historischen Litteratur auszufüllen. Angesichts anderer Arbeiten aber mußte dieser Plan einstweilen zurückgestellt werden, bis die Commission sich in ihrer Jahressitzung von 1868 in der Lage sah, ihn, dem Antrage ihres Vorsitzenden, des Geh. Reg.-Raths L. v. Ranke und des Reichsraths Dr. v. Döllinger folgend, wieder aufzunehmen. Es ward zunächst der mitunterzeichnete Frhr. v. Liliencron mit der Leitung betraut und auf Grund der von ihm gemachten Vorschläge wurden sodann in der Jahressitzung von 1869 die Grundzüge des Unternehmens berathen und festgestellt. Zur Uebernahme des Verlags und Druckes entschloß sich in Würdigung der nationalen Bedeutung des Werkes die Verlagsbuchhandlung Duncker und Humblot in Leipzig.

Das unter dem Namen einer „Allgemeinen deutschen Biographie" herauszugebende Werk war nach den Beschlüssen der historischen Commission zugleich für den wissenschaftlichen Gebrauch des Gelehrten und für die Gesammtheit der Gebildeten zu berechnen. Dem ersten Zweck muß dadurch genügt werden, daß die Biographien so weit wie irgend möglich auf die Kreise auch solcher Personen ausgedehnt werden, welche ein ausschließlich oder doch überwiegend nur wissenschaftliches Interesse haben und daß dem Nachschlagenden das wissenschaftliche Material vorgeführt oder durch Nachweisungen zugänglich gemacht wird. Um des zweiten allgemeineren Zweckes willen aber muß vor Allem denjenigen Biographien, welche auf eine weiter ausgebreitete Theilnahme rechnen können, die Aufgabe gestellt werden, ihren Inhalt in gemeinfaßlicher Darstellung und in wohllesbarer Form zu geben. Der Staatsmann ist nicht dem Historiker allein, der Theologe, der Philosoph, der Jurist, der Künstler u. s. w. nicht nur für seine Fachgenossen darzustellen, sondern sie Alle sollen dem Verständniß des Gebildeten überhaupt entgegengebracht werden. Aufgenommen werden sollen aber in die Biographie alle bedeutenderen Persönlichkeiten, in deren Thaten und Werken sich die Entwickelung Deutschlands in Geschichte, Wissenschaft, Kunst,

Figure A.3: Sample page of scan of the Allgemeine Deutsche Biographie (ADB) made available by the Bavarian State Library to the CIS (University of Munich).


le genug zu Studieren/ vnd werden es auch E-
wig nicht aus können lernen/ Dörffen derhal-
ben/ der menschen zusatzung gar nichts/ die al-
le jnn vns von der warheit abwenden / wie S.
Paulus klar sagt/ Titum am .j. vnnd Zacha.v.
vnd Psalm.x. den fluch vns bringen/ so wirs an
nemen.

Darumb jhr lieben grossen Herrn / jhr
alle hohe Potentaten / jhr aus allen Stenden/
die jr anders Christen sein wolt/ folget der stim
me Jhesu Christi / vnd fliehet die euch anders
lehren/ denn Christus schon gelehret hat/ vnd
vns befolhen Matth.xxviij. Das seine Jünger
vñ wir seine Prediger aller ding nichts newes/
auch nichts anders sollen lehren / denn das er
vns befohlen hat. Wie Johannis xiiij. vnd xvj.
stehet. Der heilig Geist wird euch erjnnern al-
les/ das ich Christus euch gesagt habe / Was
ich nun nicht zuuor euch habe im Euangelio
geprediget/ als von Wallfarten / Möncherey/
Seelmessen/ Heiligen anruffung/ das nehmet
nicht an/ Sondern bleibt bey dem Euangelio/
das jnn der heiligen Schrifft ist / von Gottes
Son verheischen durch seine Propheten / Ro-
ma.j.

Derhalben sage ich gewis/ die selig/ vnd
viel hundert tausent mahl selig/ die sich frewen
jnn der letzten gefehrlichen zeit/ den HERRN
Christum frey mit seinem Euangelio zubeken-
B iij nen/

Figure A.4: Sample page of a scan of Kaspar Aquila's *Eyn sehr hoch noetige Ermanung* printed by Gervasius Stürmer in 1548 the Bavarian State Library.

Abrevations

ADB Allgemeine Deutsche Biographie

BLV Bibliothek des Literarischen Vereins Stuttgart

ENHG Early New High German

NHG New High German

MCF Münchner Corpus für Frühneuhochdeutsch (Munich Corpus for Early New High German)

WS Wikisource, this usually refers to the German subdivision, URL: http://de.wikisource.org/

$\mathbf{W}^{\mathrm{Orig}}$ **/ text**$^{\mathrm{Orig}}$ The original word / text on paper.

$\mathbf{W}^{\mathrm{Orig}}$ **/ text**$^{\mathrm{OCR}}$ The word / text as (mis)recognized by the OCR software

$\mathbf{W}^{\mathrm{Corr}}$ **/ text**$^{\mathrm{Corr}}$ Postcorrected word / text.

Bibliography

Allgemeine Deutsche Biographie. Historische Commission bei der Königl. Akademie der Wissenschaften, Leipzig, 1875.

Johann Christoph Adelung. *Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart*, volume 5. Leibzig, 1774.

Rolf Bergmann and Dieter Nerius, editors. *Die Entwicklung der Großschreibung im Deutschen von 1500 bis 1700*, volume 2. C. Winter, Heidelberg, 1998.

Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

Unicode Consortium, editor. *The Unicode Standard, Version 5.0*. Addison-Wesley Professional, 5th edition, 2006.

Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. *CoRR*, cs.CL/0205057, 2002.

Werner Doede. *Bibliographie deutscher Schreibmeisterbücher von Neudörffer bis 1800*. Hamburg, 1950.

Peter F. Ganz. *Der Einfluss des Englischen auf den Deutschen Wortschatz*. Erich Schmidt, Berlin, 1957.

Karl E. Georges. *Lateinisch-Deutsch, Deutsch-Lateinisch*. Directmedia Publishing, 1. edition, 2004.

John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2), 2001.

Jacob Grimm and Wilhelm Grimm. *Deutsches Wörterbuch. Elektronische Ausgabe der Erstbearbeitung*. Zweitausendeins, 2004.

Ursula Götz. *Die Anfänge der Grammatikschreibung des Deutschen in Formularbüchern des frühen 16. Jahrhunderts: Fabian Frangk - Schryfftspiegel - Johann Elias Meichßner*. PhD thesis, Heidelberg, 1992.

Margaret A. Hafer and Stephen F. Weiss. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10(11-12):371–385, 1974.

Beat Rudolf Jenny. *Graf Froben Christoph von Zimmern*. Jan Thorbecke, Lindau und Konstanz, 1959.

Kluge. Etymologisches Wörterbuch der Deutschen Sprache. Elmar Seebold, 24 edition, 2002.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. Language model based arabic word segmentation. In *ACL*, pages 399–406, 2003.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, 1966.

Gebhard Mehring. Schrift und schrifttum. In *Schwäbische Volkskunde*, volume 7, Stuttgart, 1931. Silberburg.

Virgil Moser. *Frühneuhochdeutsche Grammatik*, volume 1. Carl Winter, Heidelberg, 1929.

Hans Moser. Geredete graphie. zur entstehung orthoepischer normvorstellungen im frühneuhochdeutschen. *Zeitschrift für deutsche Philologie*, 1987.

Karin Müller. *"Schreibe, wie du sprichst"*. PhD thesis, Frankfurt am Main ; Bern ; New York ; Paris, 1990.

Kemal Oflazer. Error-tolerant finite state recognition with applications to morphological analysis and spelling correction. *CoRR*, cmp-lg/9504031, 1995.

James L. Peterson. A note on undetected typing errors. *Commun. ACM*, 29(7):633–637, 1986.

Oskar Reichmann and Klaus-Peter Wegera, editors. *Frühneuhochdeutsche Grammatik*. Niemeyer, Tübingen, 1993.

Thorsten Roelcke. *Periodisierung der deutschen Sprachgeschichte*. de Gruyter, Berlin, New York, 1995.

Monica Rogati, J. Scott McCarley, and Yiming Yang. Unsupervised learning of arabic stemming using a parallel corpus. In *ACL*, pages 391–398, 2003.

Hartmann Schedel. *Schedl'sche Weltchronik*. Anton Koberger, Nürnberg, 1493.

Christian Stetter. Orthographie als normierung des schriftsystems. 1994.

Peter von Polenz. *Deutsche Sprachgeschichte*, volume 2. de Gruyter, Berlin, New York, 1994.

Christiane Wanzeck. *Die Kompositionsbildung im Frühneuhochdeutschen. Eine Studie zu den Entwicklungstendenzen und deren Faktoren.* unpublished.

LEBENSLAUF

21.2.1978	Geboren in München, Germany
1997	Abitur Gymnasium Gilching
1997 – 1998	Military Service
1998 – 2000	Elektrotechnick Studium an der TU München
1999 – 2001	Teil der Gründergruppe der IC4B AG, http://www.ic4b.de/, als System Administrator und Programmierer
2001 –	Studium an der Ludwigs-Maximillians-Universität München Computer Linguistik, Informatik, Germanistische Linguist http://www.lmu.de/
2002 –	System Administrator bei der Rechnerbetriebsgruppe am Institut für Informatik an der LMU
2002	Gründung von splashground, meiner privaten Computer Firm
2003	Organisationsteam Summerschule "New Frontiers in Science http://www.ehims.de/
2003	System Administrator am CIS (LMU), http://www.cis.uni-mu
2005 –	Werkstudent der Grid Computing Gruppe Leibniz-Rechenzent http://www.lrz.de/
5.1.2006	Heirat mit Maria Piskareva-Vassilieva
2006 –	HiWi am CIS
25.6.2006	Geburt des ersten Sohnes, Alexander Maximilian Hauser

Erklärung über die selbständige Abfassung der Magisterarbeit

Hiermit erkläre ich, die eingereichte Magisterarbeit selbständig ange
und andere als die angegebenen Quellen und Hilfsmittel nicht verwende
haben. Meine eingereichte Magisterarbeit ist nicht anderweitig als
Prüfungsleistung verwendet worden und die eingereichte Magisterarbeit
nicht in der deutschen oder in einer anderen Sprache als Veröffentlic
erschienen.

München, 4.10.2007 Andreas W. Hauser