

---

# Konstruktion historischer Wörterbücher

Andreas Neumann

---



München 2008



---

# Konstruktion historischer Wörterbücher

Andreas Neumann

---

Ludwig–Maximilians–Universität München  
Centrum für Informations– und Sprachverarbeitung

Magisterarbeit

vorgelegt von  
Andreas Neumann

München, den 02.10.2008

Betreuer der Arbeit: Professor Klaus U. Schulz

# Inhaltsverzeichnis

<b>Danksagung</b>	<b>x</b>
<b>Zusammenfassung</b>	<b>xii</b>
<b>1 Europas kulturelles Erbe bewahren</b>	<b>1</b>
1.1 i2010 Digital Libraries Initiative . . . . .	1
1.2 Improving Access to Text (IMPACT - Projekt) . . . . .	1
1.3 Ziel der Arbeit . . . . .	2
<b>2 Besonderheiten historischer Texte</b>	<b>3</b>
2.1 Layout und Typen . . . . .	3
2.1.1 Typen . . . . .	3
2.1.2 Seitenaufbau . . . . .	5
2.1.3 Durch Layout und Typen verursachte Probleme . . . . .	5
2.2 Zeichensysteme . . . . .	5
2.2.1 Die lateinische Schrift . . . . .	5
2.2.2 Einführung der Minuskel . . . . .	5
2.2.3 Abbraviaturen und Ligaturen . . . . .	6
2.2.4 Probleme, die aus dem Zeichensystem erwachsen . . . . .	8
2.3 Phonemisches Schreiben in einer fremden Schrift . . . . .	8
2.3.1 Vertrauen auf die Sprachkenntnis des Lesers . . . . .	8
2.3.2 Umdeutung freier Zeichen . . . . .	9
2.3.3 Zeichenketten . . . . .	10
2.3.4 Neue Zeichen . . . . .	11
2.3.5 Kontextsensitive Zeichen . . . . .	12
2.3.6 Zeichen verändern vorhergegangene oder nachfolgende Zeichen . . . . .	12
2.3.7 Probleme, die aus der fehlenden Entsprechung von Zeichen und Laut entstehen . . . . .	13
2.4 Sprachwandel und Dialekte . . . . .	13
2.4.1 Sprachwandel . . . . .	13
2.4.2 Dialekte . . . . .	13
2.4.3 Probleme die aus der Sprachveränderung entstehen . . . . .	14
2.5 Orthographie . . . . .	14
2.5.1 Anfänge der Orthographie . . . . .	14
2.5.2 Erste Schritte zur Verinheitlichung der Schriftsprache . . . . .	15

2.5.3	Normierung der Schriftsprache . . . . .	15
2.5.4	Probleme, die aus dem Fehlen einer Orthographie erwachsen . . . .	15
<b>3</b>	<b>Verfahren zur Gewinnung historischen Vokabulars</b>	<b>17</b>
3.1	Manuelle Extraktion des historischen Wortschatzes . . . . .	17
3.1.1	Beschreibung der manuellen Extraktion . . . . .	17
3.1.2	Vor- und Nachteile der manuellen Wortsuche . . . . .	17
3.2	Ausschlußverfahren . . . . .	17
3.2.1	Beobachtungen die zum Ausschlußverfahren führen . . . . .	18
3.2.2	Beschreibung des Ausschlußverfahrens . . . . .	18
3.2.3	Idee zur Umsetzung des Ausschlußverfahrens . . . . .	18
3.2.4	Probleme beim Ausschlußverfahren . . . . .	18
3.3	Verfahren mit hypothetischen Wörterbüchern . . . . .	18
3.3.1	Vorbeobachtungen zum “vorkochen” von Wörtern . . . . .	18
3.3.2	Beschreibung des Verfahrens mit “vorgekochten” Wörterbüchern . .	19
3.3.3	Das “vorkochen” eines Wortes . . . . .	19
3.4	Umsetzung des Verfahrens mit “vorgekochten” Wörterbüchern . . . . .	20
3.4.1	Das Korpus . . . . .	20
3.4.2	Das Wörterbuch . . . . .	20
3.4.3	Ersetzungsmuster bzw. Patterns . . . . .	20
3.4.4	Anwendung der Patterns auf das Wörterbuch . . . . .	22
3.5	Probleme des Verfahrens mit “vorgekochten” Wörterbüchern . . . . .	23
3.5.1	Ambige Einträge . . . . .	23
3.5.2	Überproduktion . . . . .	23
3.5.3	Wörter ohne moderne “Nachkommen” können nicht erkannt werden	23
3.5.4	Abkürzungen . . . . .	24
3.5.5	Größe der erzeugten Lexika . . . . .	24
3.6	Kombination der Verfahren zur Lösung der Probleme . . . . .	24
3.6.1	Lösung für ambige Einträge . . . . .	24
3.6.2	Lösung für die Überproduktion historischer Wörter . . . . .	25
3.6.3	Lösung für Wörter ohne Nachkommen . . . . .	25
3.6.4	Lösung für nicht mehr verwandte Abkürzungen . . . . .	25
3.6.5	Umgehen des Erstellens großer Lexika . . . . .	25
<b>4</b>	<b>Graphical–User–Interface zur Erstellung historischer Wörterbücher</b>	<b>27</b>
4.1	Ein Webinterface . . . . .	27
4.2	Basiskomponenten . . . . .	27
4.2.1	Ruby on Rails . . . . .	28
4.2.2	CISLEX . . . . .	28
4.2.3	VAAM . . . . .	29
4.2.4	Flector . . . . .	29
4.3	Korpus . . . . .	29
4.4	Lexika . . . . .	29

4.4.1	Das Lexikon der historischen Wortformen – HistLex . . . . .	30
4.4.2	Das Lexikon der historischen Wortformen ohne modernen Nachfolger – HistLexWS . . . . .	30
4.4.3	Das Lexikon der historischen Abkürzungen – AbbrevLex . . . . .	30
4.4.4	Das Lexikon der problematischen historischen Wortformen – Hist- ProbLex . . . . .	30
4.4.5	Das Lexikon der Entitäten – EntLex . . . . .	30
4.4.6	Das moderne Zusatzlexikon – ModLex . . . . .	31
4.5	Benennung des GUIs . . . . .	31
<b>5</b>	<b>Arbeiten mit LeXtractor</b>	<b>33</b>
5.1	Die Hauptansicht . . . . .	33
5.2	Wort zum Bearbeiten auswählen . . . . .	34
5.2.1	Hochfrequente Wörter auffinden und bearbeiten . . . . .	34
5.2.2	Am Text arbeiten . . . . .	34
5.3	Ein Wort aus dem hypothetischen Lexikon zum Lexikon der historischen Wortformen hinzufügen . . . . .	36
5.3.1	Auswahl der Lesart . . . . .	36
5.3.2	Zuordnung des Strings zu einem Paradigma . . . . .	36
5.3.3	Hinzufügen des Worts . . . . .	37
5.4	Sonstige Möglichkeiten zum Hinzufügen eines Worts . . . . .	37
5.4.1	Wort zum Lexikon der historischen Wörter ohne Nachfolger hinzufügen	37
5.4.2	Wort zum Lexikon der historischen Abkürzungen hinzufügen . . . . .	37
5.4.3	Wort zum Lexikon der problematischen historischen Wortformen hin- zufügen . . . . .	38
5.4.4	Wort dem modernen Zusatzlexikon hinzufügen . . . . .	38
5.4.5	Wort zum Entitätenlexikon hinzufügen . . . . .	38
5.5	Lexika bearbeiten . . . . .	38
5.5.1	Das Lexikon der historischen Wörter – HistLex . . . . .	38
5.5.2	Das Lexikon der historischen Wörter ohne Nachfahren – HistLexWS	39
5.5.3	Das Lexikon der historischen Abkürzungen – AbbrevLex . . . . .	40
5.5.4	Das Lexikon der problematischen historischen Wörter – HistProbLex	40
5.5.5	Das Lexikon der Entitäten . . . . .	41
5.5.6	Das moderne Zusatzlexikon . . . . .	41
5.6	Korpus . . . . .	41
5.6.1	Korpus – Die Übersicht . . . . .	41
5.6.2	Korpus – Daten ändern . . . . .	42
5.7	Ersetzungsmuster . . . . .	42
5.7.1	Der Menüpunkt Ersetzungsmuster . . . . .	42
5.7.2	Ersetzungsmuster hinzufügen . . . . .	43
5.8	Auswertung . . . . .	43
5.8.1	Lemmata . . . . .	43
5.8.2	Muster . . . . .	44

5.8.3	Korpora . . . . .	44
5.8.4	Lexikonabdeckung . . . . .	45
5.9	Configuration . . . . .	46
5.10	Nutzerverwaltung . . . . .	46
5.10.1	Hauptansicht Nutzerverwaltung . . . . .	46
5.10.2	Neuen Nutzer anlegen . . . . .	47
5.11	Partials . . . . .	47
5.11.1	Konkordanz . . . . .	47
5.11.2	Frequenzlisten . . . . .	47
5.11.3	State . . . . .	48
<b>6</b>	<b>Testen des GUIs</b>	<b>55</b>
6.1	Das Testkorpus . . . . .	55
6.1.1	Die Ausgangsdaten . . . . .	55
6.1.2	Ausgangsdaten im GUI . . . . .	55
6.2	Beschreibung des Ziellexikons . . . . .	57
6.3	Vorgehen . . . . .	57
6.3.1	Das 19. Jahrhundert . . . . .	57
6.3.2	Das 18. Jahrhundert . . . . .	60
6.3.3	Das 17. Jahrhundert . . . . .	63
6.3.4	Das 16. Jahrhundert . . . . .	66
6.4	Ergebnisse . . . . .	69
6.4.1	Zeitaufwand und schwierige Situationen . . . . .	69
6.4.2	Die gewonnenen Ersetzungsmuster . . . . .	69
6.4.3	Das gewonnene Lexikon . . . . .	70
<b>7</b>	<b>Schlußbetrachtung</b>	<b>71</b>
<b>A</b>	<b>Textbeispiele</b>	<b>73</b>
A.1	Textbeispiel 13. Jahrhundert – Under der Linden (Codex Manesse) . . . . .	73
A.1.1	Scan des Originaltextes . . . . .	73
A.1.2	Textinhalt . . . . .	73
A.2	Der Testkorpus . . . . .	74
A.2.1	Was ein Comet sey: woher er komme / vnd seinen vrsprung habe / [...] 74	
A.2.2	ie Hand mit auffgerekten dreyen schwartzen Fingern / So der falsche Eyd bedeut . . . . .	75
A.2.3	Erschreckliche / vnerhörte Mordt That / So sich den 28. Aprilis / dieses 1606. Jahres [] zugetragen / [] . . . . .	75
A.2.4	Beschreibung eines Wunder-Menschen / zu diesen unsern Zeiten entsprungen in der Neapolitanischen . . . . .	75
A.2.5	Brief an Christian Wolff . . . . .	76
A.2.6	Geschichte des 30jährigen Kriegs . . . . .	76
A.2.7	über Pädagogik . . . . .	76



A.2.8	Additional-Artikel zu dem am 21. Oktober 1867 zwischen der Postverwaltung des Norddeutschen Bundes und der Postverwaltung der Vereinigten Staaten von Amerika abgeschlossenen Verträge für die Verbesserungen des Postdienstes zwischen den beiden Ländern, sowie zu dem Additional-Verträge vom 7./23. April 1870. . . . .	77
<b>B</b>	<b>Aufbau der Tabellen</b>	<b>79</b>
B.1	Tabelle Historic . . . . .	79
B.1.1	Basistabelle – Historic . . . . .	79
B.1.2	Tabelle – Lemma . . . . .	79
B.1.3	Tabelle – FlexionClass . . . . .	79
B.1.4	Tabelle – UsedPattern . . . . .	80
B.1.5	Tabelle – Beleg . . . . .	80
B.2	Tabelle – Abbreviation . . . . .	80
B.3	Tabelle – HistoricWithoutAncestor . . . . .	80
B.4	Tabelle – HistoricProblem . . . . .	80
B.5	Tabelle – Entity . . . . .	81
B.6	Tabelle – Modern . . . . .	81
B.7	Tabelle – Korpus . . . . .	81
<b>C</b>	<b>Programmcode</b>	<b>83</b>
C.1	VamInterface . . . . .	83
C.2	FlexionInterface . . . . .	85
C.3	100words . . . . .	87
<b>D</b>	<b>Lexikon des Testkorpus</b>	<b>89</b>
D.1	Testkorpus – Historische Lexika . . . . .	89
D.1.1	Testkorpus – Lexikon der historischen Formen . . . . .	89
D.1.2	Testkorpus – Lexikon der historischen Abkürzungen . . . . .	92
D.1.3	Testkorpus – Lexikon der historischen Wörter ohne Nachfolger . . .	93
D.1.4	Testkorpus – Lexikon der problematischen historischen Wortformen	93
D.2	Testkorpus – Entitätenlexikon . . . . .	94
D.3	Testkorpus – modernes Zusatzlexikon . . . . .	94



# Danksagung

Besonders bedanken möchte ich mich bei Annette Gotscharek und Uli Reffle, die mir bei meiner Arbeit stets als Ansprechpartner zur Verfügung standen, ihre Ideen beisteuerten und mir sonst jedwede denkbare Unterstützung angedeihen ließen.

Mein Dank gilt auch Christoph Ringlstetter und allen die sich zum Testen der Anwendung bereiterklärt haben und mich mit ihrer Fachkenntnis unterstützt haben.

Danken möchte ich überdies Max Hadersbeck, der meine Liebe zum Programmieren weckte und den ich 1 1/2 Jahre lang als Tutor unterstützen durfte.

Hervorheben möchte ich auch Professor Klaus U. Schulz, der es mir ermöglichte diese Arbeit zu schreiben.



# Zusammenfassung

Im Rahmen dieser Arbeit wird eine graphische Benutzeroberfläche (Graphical–User–Interface kurz GUI) entwickelt um historische Lexika zu erstellen. Die Arbeit gliedert sich in sieben Teile:

Der erste Teil zeigt die Beweggründe die zur Beschäftigung mit historischen Texten führen.

Im zweiten Teil wird auf die Besonderheiten historische Texte und die daraus resultierenden Probleme für die Lexikonerstellung eingegangen.

Im dritten Teil werden verschiedene Verfahren zur Lexikonerstellung vorgestellt, verglichen und bewertet. Es wird versucht in einem Syntheseschritt die Stärken der vorgestellten Verfahren zu kombinieren.

Im vierten Teil werden die Kernkomponenten des GUIs beschrieben, mit dem das Verfahren umgesetzt wird.

Der fünfte Teil beschreibt die Funktionen des GUIs.

Der sechste Teil beschreibt die Evaluation des GUIs anhand eines Testkorpus.



# Abbildungsverzeichnis

2.1	Handschriften . . . . .	4
2.2	Incunabeln . . . . .	4
5.1	Wort nach Häufigkeit einfügen . . . . .	49
5.2	Am Text arbeiten um ein Wort zu wählen . . . . .	50
5.3	Auswahl der Lesarten . . . . .	50
5.4	Anzeige der aktuellen Lesart . . . . .	51
5.5	Anzeige des aktuell in Bearbeitung befindlichen Paradigmas . . . . .	51
5.6	Darstellung des Paradigmas mit Score . . . . .	52
5.7	Darstellung der Konkordanz . . . . .	53
5.8	Die Teildarstellung State . . . . .	53





# Tabellenverzeichnis

2.1	Abbiatiuren in modernen Texten . . . . .	6
2.2	Abbiatiuren in historischen Texten . . . . .	6
2.3	Ligaturen . . . . .	7
2.4	Ligaturen in historischen Texten . . . . .	7
2.5	umgedeutete Zeichen . . . . .	9
2.6	Zeichenketten im heutigen Deutsch . . . . .	10
2.7	Zeichenketten im historischen Deutsch . . . . .	10
2.8	Wortzeichen mit Zusatz . . . . .	11
2.9	historische Zeichen mit Zusatz . . . . .	11
2.10	Kontextsensitive Zeichen . . . . .	12
2.11	Doppelung und Dehnung von Vokalen . . . . .	12
2.12	Kürzung von Vokalen . . . . .	12
2.13	Zuordnung Lemma zu graphemischer Repräsentation . . . . .	15
2.14	Zuordnung graphemische Repräsentation zu Lemma . . . . .	16
3.1	Historische Formen des Wortes Tat . . . . .	19
3.2	Ersetzungsmuster . . . . .	19
3.3	vorgekochte Wortformen des Wortes "Tat" . . . . .	20
3.4	Induktives Verfahren zum Auffinden von Ersetzungsmustern . . . . .	22
3.5	Besipiel für ein aus den gewonnenen Ersetzungsmuster erzeugtes Wörterbuch	22
3.6	Ambiguität der erzeugten Vollformen . . . . .	23
3.7	Überproduktion . . . . .	23
6.1	Testkorpor: Übersicht Ausgangsmenge . . . . .	56
6.2	Testkorpor: Übersicht nach Jahrhunderten . . . . .	56
6.3	Testkorpor: 19. Jahrhundert bearbeitet . . . . .	58
6.4	Testkorpor: Übersicht Ausgangsmenge nach Bearbeitung des 19. Jahrhunderts	59
6.5	Testkorpor: Übersicht nach Jahrhunderten nach Bearbeitung des 19. Jahr- hunderts . . . . .	59
6.6	Testkorpor: 18. Jahrhundert bearbeitet . . . . .	61
6.7	Testkorpor: Übersicht Ausgangsmenge nach Bearbeitung des 18. Jahrhunderts	62
6.8	Testkorpor: Übersicht nach Jahrhunderten nach Bearbeitung des 19. Jahr- hunderts . . . . .	62
6.9	Testkorpor: 17. Jahrhundert bearbeitet . . . . .	64
6.10	Testkorpor: Übersicht Ausgangsmenge nach Bearbeitung des 17. Jahrhunderts	65

6.11 Testkorpus: 16. Jahrhundert bearbeitet . . . . .	67
6.12 Testkorpus: Übersicht Ausgangsmenge nach Bearbeitung des 16. Jahrhunderts	68
6.13 Lexikon des Testkorpus . . . . .	70

# 1 Europas kulturelles Erbe bewahren

Europa kann auf eine lange wissenschaftliche Tradition zurückblicken. Wichtige Werke und Entdeckungen werden in Bibliotheken oder ähnlichen Institutionen aufbewahrt. Dort sind sie zwar gut geschützt und können bei Bedarf auch restauriert werden, doch ist der Zugang zu diesen Werken oft zeitaufwendig und schwierig. Um dem Durchschnittsbürger auch am kulturellen Erbe Europas teilhaben zu lassen wurden Projekte gestartet um einen leichteren Zugang zu ermöglichen.

## 1.1 i2010 Digital Libraries Initiative

Mit der i2010 Digital Libraries Initiative hat sich die europäische Kommission zum Ziel gesetzt Europas kulturelles Erbe einer breiten Bevölkerung online zugänglich zu machen.[9]. Im Bereich der Bibliotheken setzt dies eine Digitalisierung des vorhanden Materials voraus. Bis zum jetzigen Zeitpunkt liegt nur c.a. 1% des vorhanden Bestands in digitaler Form vor. Gründe dafür sind vor allem die hohen Kosten und das fehlende “Know-How” bei der Digitalisierung historischer Texte.

Digitalisierung bedeutet hier nicht nur die digitale Fotografie eines Dokuments anzufertigen. Der gesamte Inhalt des Dokuments muss erschlossen werden, um es sinnvoll in eine moderne Bibliotheksumgebung integrieren zu können. Zeitgemäße Dienste, wie automatische Inhaltsklassifikation, Suche in Dokumenten und Querverweise über Hyperlinks sind nur möglich wenn das Dokument auch inhaltlich von einem Computer erschlossen werden kann. Aus diesem Grund wurde das Impact Projekt ins Leben gerufen.

## 1.2 Improving Access to Text (IMPACT - Projekt)

Das Impact-Projekt wurde im Januar 2008 gestartet und ist auf eine Dauer von 4 Jahren angelegt. Auf der Webseite des Projekts werden folgende Hauptziele definiert[12]:

- Den Zugang zu historischen Texten zu verbessern
- Die OCR-Technologie zu verbessern
- Mittel und Wege zu finden mit Problemen umzugehen die aus Unterschieden der heutigen und der damaligen Sprache resultieren
- Hilfe und Anleitungen zur Massendigitalisierung von Texten zu geben

- Material in allen europäischen Sprachen sammeln und bearbeiten zu können

Um diese Ziele erreichen zu können wird im IMPACT-Projekt u.a. folgendes entwickelt[13]:

- OCR-Technologie die mit historischen Texten umgehen kann
- Tools und Lexika die mit der Varianz in historischer Sprache umgehen können

### **1.3 Ziel der Arbeit**

Als Partner im IMPACT-Projekt obliegt es dem Centrum für Informations- und Sprachverarbeitung (CIS) unter der Leitung von Professor Klaus U. Schulz Sprachmodelle und Wörterbücher zu entwickeln. In dieser Arbeit wird daher ein Verfahren und ein Graphical User Interface (GUI) entwickelt, das mit besonderer Rücksicht auf historische Sprachvarietäten einen Nutzer bei der Erstellung historischer Wörterbücher unterstützen soll.

## 2 Besonderheiten historischer Texte

Vor dem eigentlichen Erstellen des Wörterbuchs möchte ich in diesem Kapitel auf die Besonderheiten historischer Texte und die Probleme, die sich daraus bei der Erstellung von Wörterbüchern ergeben, eingehen.

Historische Texte weisen im Vergleich zu modernen Texten sowohl für den unbedarften Leser als auch für die maschinelle Bearbeitungen einige Besonderheiten im Bereich des Layouts (Seitengestaltung, Schrifttypen), des verwandten Zeichensystems und der Interpretation der verwendeten Zeichen (Lautzuordnung und Bedeutung) auf.

Auf Probleme die aus dem verwendeten Druckmaterial, Beschädigung und natürlicher Alterung resultieren wird hier nicht eingegangen. Dazu verweise ich auf die Arbeit von Andreas Hauser.[11]

### 2.1 Layout und Typen

Die äußere Form eines historischen Texts kann sich je nach Alter enorm von einem Text, der den heute gültigen Konventionen folgt, unterscheiden. Über die Jahrhunderte hinweg war die Form eines gedruckten/geschriebenen Dokuments starken Veränderungen ausgesetzt. Neuerungen in der Drucktechnik, neue ästhetische Ansprüche und wirtschaftliche Interessen (zum Beispiel Senkung der Druckkosten) führten zu mannigfaltigen Veränderungen in der äußeren Form eines Textes.

#### 2.1.1 Typen

Dieser Abschnitt befasst sich mit den verschiedenen Schrifttypen die zu bestimmten Zeiten verbreitet waren. Vor dem Siegeszug des Buchdrucks in Europa wurden Texte vor allem handschriftlich verfasst.

Die Grafik 2.1 zeigt einen Auszug aus den damals verwandten Handschriftalphabeten.

Mit der Erfindung des Buchdrucks wurden sog. Typenalphabete verwendet. Die frühen Druckschriften bis 1600, auch Inkunablen oder Wiegedruck genannt[3, S. 65], waren noch wenig einheitlich:

Die Grafik 2.2 zeigt einen Auszug aus den damals verwandten Alphabeten.

Ab dem sechzehnten Jahrhundert werden die Techniken ausgefeilter und die Typen einheitlicher.

Uncial		Alt-Irisch		8. — 9. Jahrh.		9. — 10. Jahrh.		10. Jahrh.	
Grosse	Kleine	Initial	Minusk.	Initial	Minusk.	Initial	Minusk.	Initial	Minusk.
À Á	à	À Á	à á	À Á Á	à æ	À Á Á	à	À	à
Β Β	β	Ḃ	ḃ	Β	ḃ	Β	ḃ	Β	ḃ
Ɔ	ε	Ɔ	ε	Ɔ	ε	Ɔ	ε	Ɔ	ε
Ɔ	ο ϑ	Ɔ	ο	Ɔ	d	Ɔ Ɔ	d	Ɔ	d ο

[6, S.196]

11. Jahrh.		12. Jahrh.		12. — 13. Jahrh.		13. — 14. Jahrh.		14. Jahrh.	
Initial	Minusk.	Initial	Minusk.	Initial	Minusk.	Initial	Minusk.	Initial	Minusk.
À Á	à	À Á	à á	À Á Á	à á	À Á	à	À	à
Β Β	β	Β	ḃ	Β Β Β	ḃ	Β	ḃ	Β	ḃ
Ɔ	ε	Ɔ	ε	Ɔ	ε ε	Ɔ	ε ε	Ɔ	ε
Ɔ	d ε	Ɔ Ɔ	d d	Ɔ Ɔ Ɔ	d	Ɔ d Ɔ	d	Ɔ	d

[6, S.197]

Abbildung 2.1: Handschriften

Fraktur				Schwabacher
London 1476	Paris 1498	Augsburg 1514	Lyon 1538	
À a	À a	À a	À a	À a
Β β	Β β	Β β	Β β	Β β
Ɔ c	Ɔ c	Ɔ c	Ɔ c	Ɔ c
Ɔ d	Ɔ d	Ɔ d	Ɔ d	Ɔ d

[6, S.205]

Abbildung 2.2: Incunabeln

### 2.1.2 Seitenaufbau

Viele den Leser unterstützende Hilfsmittel wie Abstände, Satzzeichen, Einrückungen, Verwendung verschiedener Schriftarten, Wortabstände, Satzzeichen, Abstände zwischen Sinnabschnitten usw. wurden erst nach und nach eingeführt.

### 2.1.3 Durch Layout und Typen verursachte Probleme

Die unterschiedliche Schriften und der ungewohnte Aufbau der Seiten sind vor allem ein Problem für die OCR. Der unvorhersagbare Seitenaufbau erschwert die Segmentierung. Die Zeichenerkennung wird durch unbekannte und uneinheitliche Typen erschwert.

## 2.2 Zeichensysteme

Die in fast allen europäischen Ländern benutzten Zeichen zur (Laut)-Darstellung eines Wortes entstammen einem den Römern entlehnten Alphabet.[6, S. 191] Da dieses Alphabet nicht geeignet war alle Laute der Zielsprache darzustellen wurde es mit der Zeit den Bedürfnissen dieser angepasst.

### 2.2.1 Die lateinische Schrift

Die vorherrschende Form der lateinischen Schriften war die römische Capitalschrift[6, S. 192]. Sowohl römische Eroberungen als auch die spätere Ausbreitung des Christentums trugen zur Verbreitung dieses Schriftsystems in Europa bei.

In der ursprünglichen Form besteht die römische Schrift aus 21 Buchstaben:

**Alphabet:** A B C D E F Z H I K L M N O P Q R S T V X

### 2.2.2 Einführung der Minuskel

Im Mittelalter reichten die 21 ursprünglichen Zeichen des Alphabets den Schreibern nicht mehr aus. Neben neuen Zeichen wurden auch die sogenannten Minuskel oder Kleinbuchstaben eingeführt. Diese finden sich seit dem 8. Jahrhundert. Sie wurden in Alkuins Schule in Tours unter dem Einfluss irischer Mönche entwickelt.[6, S.196]. Das Vorhandene Zeichenninventar besteht nun aus 25 Majuskeln (neu sind hier G,U,W und Y) und 25 Minuskeln.

**Majuskel:** A B C D E F G H I K L M N O P Q R S T U V W X Y Z

**Minuskel:** a b c d e f g h i k l m n o p q r s t u v w x y z

### 2.2.3 Abbreviaturen und Ligaturen

Ab 800 beginnt man auch Abbreviaturen zu verwenden.[6, S. 197] Bis zum Beginn des Buchdrucks im 14. Jahrhundert wuchs deren Zahl ständig an. Seit der Einführung des Buchdrucks mit beweglichen Lettern sinkt die Zahl der Abbreviaturen, da für jede ein einzelner Letter nötig wäre aus Kostengründen.

#### Abbreviaturen

Abbreviaturen erfüllen ähnliche Aufgaben wie Abkürzungen. Im Gegensatz zu diesen handelt es sich aber um ein einzelnes Zeichen, das aus mehreren Zeichen zusammengesetzt sein kann oder um ein Symbol. Im heutigen Zeichengebrauch benutzt man zum Beispiel für Währungen noch Abbreviaturen.

Für Beispiele von Abbreviaturen im heutigen Sprachgebrauch siehe Tabelle 2.1.

Zeichen	Bedeutung
\$	Dollar
£	Britische Pfund
&	und , "Kaufmannsund"
@	heute vor allem in E-Mailadressen, ursprünglich Stückpreis
©	Copyright

Tabelle 2.1: Abbreviaturen in modernen Texten

Die Abbreviaturen der historischen Sprache gehen auf handschriftliche Abkürzungen zurück. Überdies sind sie oft lateinischen Ursprungs werden aber in ihrer übersetzten Bedeutung verwendet. Dies macht ihre Entschlüsselung noch schwieriger. Für einen kleinen Auszug historischer Abbreviaturen siehe Tabelle 2.2.

Diese Abbreviaturen stammen aus Faulmanns Schriftzeichen und Alphabete [6, S.199].


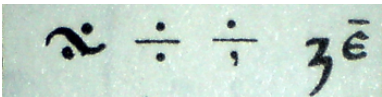

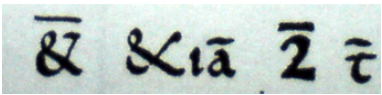
Zeichen	Bedeutung
	commune bzw. Gemeinschaft
	con bzw. mit
	contrahiter bzw. gegen
	cuius bzw. dessen

Tabelle 2.2: Abbreviaturen in historischen Texten



Ligaturen

Hier handelt es sich um die enge Verbindung zweier Zeichen zu einem Zeichen. Man findet sie vor allem im Buchdruck, wo sie einen angenehmeren Textfluss bewirken sollen. Viele Abkürzungen lassen sich auf Ligaturen zurückführen und auch einige moderne Zeichen der Standardschrift gehen auf Ligaturen zurück. Die ursprünglichen Zeichenbedeutungen bleiben ungefähr erhalten.[3][Stichwort: Ligatur] Für Ligaturen die in der modernen Sprache erhalten geblieben sind siehe: Tabelle 2.3.

Für ein Beispiel für Ligaturen in historischen Texten siehe Tabelle 2.4.

Buchstaben	Zwischenform	moderne Form
f l		fl
f i		fi
V V	W	
ae	æ, a <sup>e</sup>	ä
oe	œ, o <sup>e</sup>	ö
ad		@
et		&
cto		%

Tabelle 2.3: Liagturen

Das Beispiel stammt von Faulmann und bezieht sich auf Gutenbergs Bibelschrift [6, S. 203]

















	præ		ser
	po		ss
	pp, pop		st
	ppe		ta
	pre, pri		ter, tur
	pri		th
	pro		the
	prop		ua

Tabelle 2.4: Ligaturen in historischen Texten

### 2.2.4 Probleme, die aus dem Zeichensystem erwachsen

Für die OCR ist der Bereich der Ligaturen und Abbraviaturen ein großes Problem. Abbraviaturen sind schwer zu erkennen, da Überprüfungsmechanismen auf Wortebene, wie zum Beispiel Fehlerwörterbücher, nicht auf sie anwendbar sind (da es sich oft nur um ein einzelnes Zeichen handelt). Desweiteren werden sie nicht einheitlich, sogar schreiberabhängig gebraucht. Ligaturen sind nicht immer eindeutig trennbar und auf ihre Ursprungszeichen zurückführbar. Daher kann es leicht zu einer Fehlinterpretation kommen.

Für die Lexikographie stellt sich die Frage ob und wie (historische) Abbraviaturen und Ligaturen aufzunehmen sind. Da es in keinem genormten (Computer)-Zeichensatz Entsprechungen für diese Zeichen gibt. Wünschenswert wäre eine Auflösung der Ligaturen in ihre Ausgangszeichen und eine Ersetzung der Abbraviatur durch die "ausgeschriebene" Form auf Seiten der OCR.

## 2.3 Phonemisches Schreiben in einer fremden Schrift

In der ursprünglichsten Form des Schreibens in einer Sprache ohne eigene Schrift benutzt man eine Lehn­schrift, im Falle des deutschen war dies Latein. Der Schreiber versucht dabei jedem Phonem seiner Sprache den entsprechenden Laut der Sprache aus der das Schriftsystem stammt, zuzuordnen. Da die Sprachen nicht identisch sind gibt es sowohl Laute der eigenen Sprache die nicht dargestellt werden können, als auch Laute der Sprache, zu der das Schriftsystem ursprünglich gehört, die nicht gebraucht werden. Um alle Laute der eigenen Sprache darstellen zu können, für die es in der schriftgebenden Sprache keine Entsprechung gibt, wurden im Deutschen mehrere Ansätze nebeneinander verfolgt, von denen einige bis heute praktiziert werden.

### 2.3.1 Vertrauen auf die Sprachkenntnis des Lesers

Es wird ein Zeichen gewählt das einem dem darzustellenden Laut ähnlichen Lautwert besitzt. Der Schreiber geht davon aus, dass der die Sprache kennende Leser das Zeichen richtig, das heißt zur Sprache passend interpretiert. Diese Technik wird heute noch bei Dialekttexten angewandt.

Als Beispiel, sowohl für einen historischen als auch dialektalen Text in Umschrift, dient ein Auszug aus dem "Briefwechsel eines bayrischen Landtagsabgeordneten" von Ludwig Thoma.[18, S. 5].

Gelibte Leser

Ich bin der Jozef Filser, kgl. Abgeorneter im Parlamend.

Ich bin gebohren am 16. Sedember 1856 in Mingharding, Bosd daselbst, als der Sohn des Silfester und der Ursuhla Filser. Ich bin fon meinen Beruf Oegonohm und durch das Ferdrauen des Folkes parlamendarrischer Abgeorneter. Ich habe die Schule in Mingharting besucht und auch zu meiner Follkommenheit das

Mzgerhandwerg erlehrt bis ich das elderliche Anwesen iebernahm und es noch besieze.[23]

Jene Stellen die dem Leser als Rechtschreibfehler erscheinen, entpuppen sich bei genauerer Betrachtung als lautähnliche Wiedergabe der dialektalen Aussprache der Wörter.

### 2.3.2 Umdeutung freier Zeichen

Wie in Abschnitt 2.3 erwähnt existieren im Ausgangsalphabet oft unbelegte Zeichen. Diese werden dann mit einer neuen Lautbedeutung versehen. Die Tabelle 2.5 soll die unterschiedliche Belegung einiger Zeichen in einem Sprachvergleich zwischen verschiedenen europäischen Sprachen zeigen, die sich zwar ein Alphabet teilen aber ein unterschiedliches Phoneminventar besitzen. Wie man im Beispiel sieht, sind nicht alle Laute in allen Sprachen vorhanden

Umschrift nach Lepsius in [6, S. 6]			
Zeichen	Französischer Laut	Deutscher Laut	Finnischer Laut
H	´ (stummes H)	h	h
U	ü	u	u
J	dsch	y	y
W	–	w	–
V	w	f	v

Tabelle 2.5: Umgedeutete Zeichen, Französisch – Deutsch – Finnisch

oder zumindest durch andere Zeichen dargestellt. So hatte zum Beispiel das französische keine Verwendung für den Laut *h*. Der Freiraum wird für das “stumme h” genutzt. Der laut *w* wird im Französischen nicht als *w* sondern mit dem in dieser Sprache zuvor unbelegten Zeichen *v* dargestellt. Eine Einführung des Zeichens *w* war somit nicht nötig. Dehnt man diese Reihen aus wird schnell klar dass unterschiedliche Sprachen die gleichen Alphabete benutzen können, aber die Lautzuordnung zwischen Graphem und Phonem nicht immer die selbe ist.

### 2.3.3 Zeichenketten

Will man das Zeicheninventar nicht erhöhen kann man eine Zeichenfolgen benutzen um einen Laut darzustellen. Im heutigen Deutsch findet man unter anderem folgende Zeichenketten die einem einzigen Laut entsprechen: Siehe Tabelle 2.6.

In historischen Texten war diese Technik Laute auszudrücken weiter verbreitet. Überbleib-

Zeichenkette	Laut [IPA] oder Umschrift
sch	[ʃ]
st	[ʃt] oder [st]
ck	[k]
pf	[pf]
tz	[ts]
ch	[x] oder [ç]
au	[aʊ]
ai,ei	[ai]

Tabelle 2.6: Zeichenketten im heutigen Deutsch

sel findet man heute noch in Kreuzworträtseln bei der Darstellung von ä,ö,ü und ß. Siehe dazu Tabelle 2.7

Zeichenkette	heute verwendetes Zeichen
ae	ä
oe	ö
ue	ü
ss	ß
iu	j

Tabelle 2.7: Zeichenketten im historischen Deutsch

### 2.3.4 Neue Zeichen

In den seltensten Fällen werden Zeichen ersonnen die mit dem Schriftsystem in das sie eingebettet werden sollen nichts zu tun haben. Sie beruhen oft auf Ligaturen (siehe auch 2.2.3) oder Überbleibseln aus Ligaturen. Aus diesen "Überbleibseln" sind auch die diakritischen Zeichen entstanden.

Ein Beispiel dafür sind zum Beispiel die deutschen Umlaute [3, Stichwort: Umlaut]: Ursprünglich war es verbreitet den Laut [ɛ] als *ae* zu schreiben, später wurde das *e* hochgestellt und es war üblich *a<sup>e</sup>* zu schreiben, das <sup>e</sup> wurde dann über das *a* gesetzt und später zu <sup>¨</sup> verkürzt. Somit steht das <sup>¨</sup> für das Zeichen [e]. Die anderen Umlaute des Deutschen entstanden auf dem selben Weg.

Für Beispiele siehe Tabelle 2.9.

Zeichen	Zusatz	Neues Zeichen	Laut [IPA]	Wortbeispiel
A	¨	Ä	[ɛ]	Ärger, Ähre
a	¨	ä	[ɛ]	Bär, Käse
O	¨	Ö	[œ] oder [ø]	Öko, Öl [4, S.26–30]
o	¨	ö	[œ] oder [ø]	Körner, Föhn
U	¨	Ü	[(ɪ)ʏ] oder [y]	Übel
u	¨	ü	[(ɪ)ʏ] oder [y]	süß, Sünde

Tabelle 2.8: Wortzeichen mit Zusatz

Neben den heute noch verwendeten Umlauten existieren in historischen Texten noch andere Darstellungen:

Zeichen	Zusatz	neues Zeichen	Wortbeispiel	heutige Darstellung
a	–	ā	ā	am oder an
e	–	ē	ē	em oder en
m	–	ṁ	ṁ	am oder an
a	e	a <sup>e</sup>	angra <sup>e</sup> ntzende	angrenzende
u	o	o <sup>u</sup>	blu <sup>o</sup> men[22]zu <sup>o</sup> [22]	Blumen
u	e	u <sup>e</sup>		

Tabelle 2.9: historische Zeichen mit Zusatz

### 2.3.5 Kontextsensitive Zeichen

Ein Zeichen kann abhängig von der Umgebung mehreren Lauten entsprechen. Dies trifft zum Beispiel auf das c zu. Je nach folgendem Vokal wird es als  $[\widehat{ts}]$  bzw.  $[z]$  (vor e und i) oder  $[k]$  (vor a,e,u und Konsonanten) aufgefasst. Siehe Tabelle 2.10 für Beispiele.

Wort	Lautwert von c
CIS	$\widehat{ts}$
Computer	'k
Clothilde	'k
Cent	z

Tabelle 2.10: Kontextsensitive Zeichen am Beispiel c

### 2.3.6 Zeichen verändern vorhergegangene oder nachfolgende Zeichen

Eine Doppelung des Zeichens oder das Einfügen eines zusätzlichen Zeichens wird benutzt, um Längenunterschiede der Vokale bei der Aussprache zu kennzeichnen.

Für Beispiele siehe Tabelle 2.11.

Wort	Mechanismus
Aal	Doppelung
Ahle	Dehnungs-h
Floß	ß
sieben	Dehnungs-e
Itzehoe	Dehnungs-e
Bleckede	Dehnungs-c

Tabelle 2.11: Doppelung und Dehnung von Vokalen

Wort	Mechanismus
Hass	ss
Halle	ll
Mann	nn
Hacke	ck

Tabelle 2.12: Kürzung von Vokalen

### 2.3.7 Probleme, die aus der fehlenden Entsprechung von Zeichen und Laut entstehen

Alle oben genannten Mechanismen zur schriftlichen Darstellung eines Lautes existieren nebeneinander und wurden von Schreibern vor Einführung einer Orthographie beliebig benutzt und kombiniert. Somit ist die eindeutige Zuordnung eines Schrifzeichens/einer Zeichenfolge zu einem Laut/einer Lautfolge nicht zu gewährleisten. Dies macht eine (sichere) maschinelle Zuordnung eines historischen, geschriebenen Wortes über dessen Lautgestalt zu dessen modernen Pedant sehr schwierig.

## 2.4 Sprachwandel und Dialekte

Sprachwandel und Dialekte fügen zusätzliche Unsicherheiten bei der Zuordnung von historischen Wortformen zu ihren Nachkommen hinzu.

### 2.4.1 Sprachwandel

Sprachwandel bezeichnet die diachrone Veränderung der Sprache. Nach Glück ist

... jede natürliche Sprache [...] [dem] Prozess des Sprachwandels unterworfen [und er betrifft] ... alle Ebenen eines Sprachsystems. [8, Stichwort: Sprachwandel]

An anderer Stelle nennt er die betroffenen Bereiche genauer:

... Phonologie, Morphologie, Syntax [...], Semantik und [...] Pragmatik [8, Stichwort: Sprachwandel]

. Somit verändern sich alle Bereiche einer lebendigen, natürlichen Sprache sofern sie nicht künstlich daran gehindert werden (zum Beispiel durch eine Festlegung der Schreibung, siehe Abschnitt 2.5) kontinuierlich.<sup>1</sup>

### 2.4.2 Dialekte

Als Dialekt wird eine synchrone Sprachvarietät bezeichnet. Nach Knoop ist ein Dialekt

Besondere Sprech- (und z.T. auch Schreib-)weise innerhalb einer National- oder Standardsprache. Die Besonderheit erstreckt sich auf alle Sprachebenen [...] hat aber v.a. in Lautung und Wortschatz eine deutliche Ausprägung ... [8, Stichwort: Dialekt]

An anderer Stelle schreibt Knoop:

---

<sup>1</sup>Hier sei zum Einstieg auf das Buch von Gerhart Wolff "Deutsche Sprachgeschichte von den Anfängen bis zur Gegenwart" und zur Vertiefung die Bücher "Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart" und "Geschichte der Deutschen Sprache" von Peter von Polenz verwiesen.

Die prägenden Kennzeichen verbinden sich mit dem Geltungsbereich des D[ialekts] dahingehend, dass dieser räumlich eingegrenzt werden kann ... [8, Stichwort: Dialekt]

Vor allem bei frühen Texten fehlt eine Standardsprache und der Schreiber eines Textes versucht die Lautung seiner Region durch die in Abschnitt 2.3 beschriebenen Techniken zu verschriftlichen.

### 2.4.3 Probleme die aus der Sprachveränderung entstehen

Sprachwandel und Dialekte potenzieren die möglichen Zuordnungen einer historischen Zeichenfolge zu einem modernen Wort. War das zu lösende Problem das aus der fehlenden Schrift erwuchs nur die nicht eindeutige Zuordnung eines Lauts zu einem Zeichen, besteht nun auf der Seite des Lautes keine absolute Sicherheit mehr. Das heißt um eine richtige Zuordnung zwischen Zeichenkette und Wort in einem bestimmten Text garantieren zu können ist es unabdingbar dessen Entstehungsort und die Entstehungszeit zu kennen. Für die Lexikonerstellung bedeutet dies, dass alle Varianten die durch Dialekte und Sprachwandel möglich sind, aufgenommen werden müssen.

## 2.5 Orthographie

Ein historischer Text kann auf den heutigen Leser mitunter befremdlich wirken. Der heutige Leser und Interpreteur von Schriftzeichen ist es gewohnt dass *ein* Wort *ein* bestimmtes Schriftbild hat, dieser Zusammenhang trifft mit steigendem Alter eines Textes immer weniger zu. Der Zusammenhang zwischen Schriftbild und Wort wird durch die Normierung des Schreibens erreicht oder wie es Augst formuliert: [8, Stichwort: Orthographie]

[Orthographie bezeichnet die] ... Gesamtheit der (amtlich) normierten Schreibkonventionen unter Einschluß der Interpunktion ...

Doch auch diese Konventionen mussten sich erst entwickeln und sind bis heute (siehe u.a. Rechtschreibreform 1996) steten Veränderungen ausgesetzt. Dieser Abschnitt gibt einen kurzen Überblick über die Entwicklung der Schreibkonventionen der deutschen Sprache.

### 2.5.1 Anfänge der Orthographie

Vor der Entstehung einer deutschen Orthographie und Schriftsprache bediente man sich des Regelsystems des Lateinischen:

Die Graphie einer Sprache oder eines Dialekts bildet historisch heraus [,] immer in Anlehnung an [eine] schon verschriftete fremde Sprache. Die heutigen westeuropäischen Sprachen sind in Anlehnung an das Lateinische verschriftet; die eine oder andere Inkonsistenz der heutigen Schreibung (zum Beispiel im Deutschen die Dehnungszeichen) erklärt sich aus der nicht ganz passgerechten Übernahme. [8, Stichwort: Orthographie]



Dieses System ist nur praktikabel wenn alle Schreiber und Leser des Lateinischen mächtig sind. Überdies weist es Lücken auf, da es den besonderen Bedürfnissen der deutschen Sprache nicht Rechnung trägt. Das gleiche gilt für alle anderen Sprachen die so verschriftlicht wurden. Überdies besteht keine Garantie dafür dass die Schreiber strittige Fälle einheitlich behandeln.

### 2.5.2 Erste Schritte zur Verinheitlichung der Schriftsprache

Mit der Ablösung des Latein als Amtssprache beginnt eine langsame Angleichung der deutschen Verschriftlichung. Maßgebend dafür sind zum einen die Kanzleisprachen:

Die [...] Kanzleisprachen streben trotz regionaler Unterschiede einen einen Überregionalen Ausgleich an, ihnen wird deshalb für die Herausbildung der Neuhochdeutschen Schrift eine wichtige Rolle zugeschrieben ... [8, Stichwort:Kanzleisprache].

Zum anderen Luthers Bibelübersetzung, die ein weit verbreitetes Schriftwerk schafft, das als Referenz für andere Texte dienen kann.

### 2.5.3 Normierung der Schriftsprache

Mit den Nationalisierungsbewegungen und der deutschen Reichsgründung 1871 strebte man auch eine Angleichung und Normierung der Schriftsprache an. Die anzuwendenden Regeln wurden in der "orthographischen Konferenz" von 1901 beschlossen, doch die Umsetzung sollte noch Jahre in Anspruch nehmen. Besonders Doppelschreibungen (also nebeneinander existierende, gleichberechtigte Schreibungen) waren Grund für Diskussion. Als Standardwerk konnte sich der Duden aufgrund eines staatlichen Erlasses schon 1903 in Deutschland durchsetzen.[5, S. 3-8]

### 2.5.4 Probleme, die aus dem Fehlen einer Orthographie erwachsen

Das fehlen einer einheitlich Schreibkonvention ist das Kernproblem bei der Erstellung historischer Lexika. Ein Lemma kann viele verschiedenen graphemische Repräsentationen besitzen: siehe Tabelle 2.13.

Aber eine graphemische Repräsentation kann auch auf mehrere Lemmata zutreffen: siehe

Lemma	graphemische Repräsentation
<i>ihm</i>	yn,jn,yhn,jhn
<i>abschrift</i>	abschriefft,abschrift
<i>tun</i>	thun, thuen

Tabelle 2.13: Zuordnung Lemma zu graphemischer Repräsentation

Tablle 2.14.

Daher muss ein Lexikon der historischen Sprache mehr ambige Einträge aufweisen als in

graphemische Repräsentation	mögliche Lemmata
<i>syn</i>	sein, sinn
<i>wysen</i>	waisen, weisen, weien, wiesen, wissen
<i>glantz</i>	glanz, glans, klans, clans

Tabelle 2.14: Zuordnung graphemische Repräsentation zu Lemma

einem Wörterbuch der modernen Sprache üblich wären um alle möglichen Zusammenhänge zu erfassen.

## **3 Verfahren zur Gewinnung historischen Vokabulars**

Ziel dieses Kapitels ist es ein Verfahren zu finden, das mit den Besonderheiten historischer Texte ( Kapitel 2) umgehen kann um damit ein (Korrektur)–Wörterbuch mit historischen Wörtern zu erstellen.

### **3.1 Manuelle Extraktion des historischen Wortschatzes**

Das klassische manuelle Verfahren um ein historisches Wörterbuch zu erstellen.

#### **3.1.1 Beschreibung der manuellen Extraktion**

Ein historischer Text wird manuell nach Worten der historischen Sprache durchsucht. Der Leser/Extraktor muss jedes Wort des Texts lesen und entscheiden ob das gelesene Wort dem aktuellen Wortschatz oder dem historischen Wortschatz zuzuordnen ist oder ob es sich um eine Entität handelt. Hat er ein historisches Wort gefunden trägt er es in eine Liste ein.

#### **3.1.2 Vor– und Nachteile der manuellen Wortsuche**

Die Ergebnisse der manuellen Suche sind meist sehr hochwertig. Es besteht dennoch die Gefahr Lesarten zu übersehen, da der Leser das Wort fast nur im Kontext des aktuellen Textes berücksichtigen kann, da die synchrone Arbeit mit mehreren (hundert) Texten aufgrund des Zeitaufwands meist nicht möglich ist. Falsch und Fehlklassifikationen sind fast ganz auszuschließen.

Das größte Problem einer gänzlich manuellen Erstellung eines Lexikons der historischen Wortformen mit diesem Verfahren sind der extreme Zeitaufwand und die hohen Kosten die damit verbunden sind.

### **3.2 Ausschlußverfahren**

Die Idee hinter diesem (automatischen) Verfahren besteht darin die Auswahlmenge so lange einzuschränken bis nur noch das gewünschte Ergebnis übrig bleibt.

### 3.2.1 Beobachtungen die zum Ausschlußverfahren führen

Historische Texte sind für einen Leser der heutigen Zeit relativ gut lesbar. Ein großer Teil der Worte gehört zum passiven Wortschatz, das heißt ist kein historisches Wort. Verständnisprobleme beim Leser, also ihm unbekannte Wörter sind entweder Entitäten (Personen, Orte, Zeitangaben...) oder historische Wörter. Entfernt man nun Entitäten und moderne Wörter aus dem Text erhält man den historischen Anteil eines Texts.

### 3.2.2 Beschreibung des Ausschlußverfahrens

In diesem Ansatz betrachten wir einen Text als eine Menge von Wörtern. Nach dem entfernen aller Wörter die dem modernen Wortschatz zuzuordnen sind bleiben nur noch Entitäten und Wörter die dem historischen Wortschatz zuzuordnen sind. Entfernt man nun auch die Entitäten erhält man die historischen Wörter.

### 3.2.3 Idee zur Umsetzung des Ausschlußverfahrens

Durch Erkennungsalgorithmen (zum Beispiel Lokale Grammatiken) oder Aufzählung könnten Entitäten auf Grund ihres Umfelds erkannt werden. Durch Nachschlagen in einem Wörterbuch (zum Beispiel CISLEX) können alle Worte die Teil des modernen Wortgebrauchs sind aufgefunden werden.

### 3.2.4 Probleme beim Ausschlußverfahren

Da die Schreibung der Entitäten auf die selbe Art und Weise wie bei Wörtern des historischen Wortschatzes variiert ist es sehr aufwendig (wenn nicht gar unmöglich) eine vollständige Aufzählung zu erreichen. Die Erkennungsverfahren sind auf ein bekanntes "Umfeld" des Wortes angewiesen und funktionieren bis jetzt nur bei aktuellen deutschen Texten.

## 3.3 Verfahren mit hypothetischen Wörterbüchern

Mangels eines besseren Terminus wird ein hypothetisches Wort<sup>1</sup> als "vorgekocht" bezeichnet.

### 3.3.1 Vorbeobachtungen zum "vorkochen" von Wörtern

Ein Großteil der Wörter des aktuellen Wortschatzes läßt sich auf eine (bzw. meist mehrere) Vorgängerformen im historischen Wortschatz zurückführen.

---

<sup>1</sup>Als hypothetisches (historisches) Wort wird eine Wortform bzw. graphemische Repräsentation eines Worts bezeichnet, welches nach bestimmten Regeln erzeugt wurde und noch nicht anhand eines Beispiels belegt werden konnte

Moderne Wortform	historische Wortform
die Tat	die That die Tath die Thate die Dat

Tabelle 3.1: Historische und moderne Formen des Wortes "Tat"

Die auftretenden Unterschiede zwischen historischen Formen und aktueller Form sind systematisch. Ersetzt man ein oder mehrere bestimmte Zeichen durch ein oder mehrere andere Zeichen kann man ein Wort des modernen Wortschatzes auf seinen historischen Vorläufer zurückführen:

Moderne Wortform	Ersetzungsmuster	historische Wortform
die Tat	$T \rightarrow Th$ an Position 1 $T \rightarrow Th$ an Position 3 $T \rightarrow Th$ an Position 1, $\epsilon \rightarrow e$ an Position 4 $T \rightarrow D$ an Position 1	die That die Tath die Thate die Dat

Tabelle 3.2: Historische und moderne Formen des Wortes "Tat" mit Ersetzungsmuster

### 3.3.2 Beschreibung des Verfahrens mit "vorgekochten" Wörterbüchern

Im ersten Schritt wird ein Lexikon erstellt in dem alle wahrscheinlich möglichen historischen Formen eines modernen Wortes enthalten sind. Dies geschieht indem alle Wörter eines Lexikons der heutigen Sprache "vorgekocht" werden. (genauer zum "vorkochen" in 3.3.3).

Ein historischer Text wird nun anhand dieses "vorgekochten" Lexikons nach Kandidaten für historische Wörter durchsucht.

Entspricht ein gefundenes Wort im historischen Text einem Eintrag im "vorgekochten" Lexikon werden die möglichen modernen Nachfahren des Wortes ermittelt.

### 3.3.3 Das "vorkochen" eines Wortes

Ein ähnliches Verfahren wird zuerst in der Magisterarbeit "OCR Postcorrection of Historical Text" von Andreas Hauser erwähnt.[11, S. 29].

Das Verfahren wurde im Rahmen des Hauptseminars "Computerlinguistische Probleme bei der Digitalisierung historischer Texte" verfeinert. Alle für das Wort in Frage kommenden Ersetzungsmuster bzw. "Patterns" werden auf ein modernes Wort angewandt: Dies soll am Beispiel des Wortes Tat gezeigt werden:

Jedes Pattern wird, oder wird nicht an jeder Position des Wortes an dem es zutrifft angewandt, bis alle möglichen Varianten des Wortes erzeugt worden sind. Die aus diesem

Wort aus modernem Lexikon	Tat
Ersetzungsmuster	$a \rightarrow aa$ $a \rightarrow ah$ $t \rightarrow th$ $t \rightarrow d$
Erzeugte Wortformen	thaat, tath, daath, thahth, that, tahth, dahd, thath, dath, daad, dat, thaath, dad, taht, daat, tad, taat, thaht, taath, tahd, thahd, dahth, daht, taad, thaad, thad, tat

Tabelle 3.3: vorgekochte Wortformen des modernen Wortes "Tat"

Verfahren gewonnen Wörter werden als Kandidaten für historische Wörter in ein Lexikon eingetragen.

### 3.4 Umsetzung des Verfahrens mit "vorgekochten" Wörterbüchern

Das Verfahren setzt zu Anfang ein Wörterbuch der modernen Sprachform, eine Liste von Ersetzungsmustern bzw. Patterns und ein Korpus voraus.

#### 3.4.1 Das Korpus

Um ein Lexikon historischer Wörter erstellen zu können ist eine Basis aus historischen Texten notwendig.

#### 3.4.2 Das Wörterbuch

Um die historischen Wortformen erzeugen zu können ist eine Basis aus modernen Wortformen nötig. Es müssen alle Wortformen vorhanden sein, das heißt aus dem Lexikon müssen alle Vollformen eines Wortes zu entnehmen sein.

#### 3.4.3 Ersetzungsmuster bzw. Patterns

Für das Erstellen einer Liste mit Ersetzungsmustern gibt es mehrere Vorgehen:

##### Praeskriptives Vorgehen

Anhand linguistischer und pragmatischer Betrachtungen (Lautverschiebungen usw.) wird eine Liste mit Patterns erstellt. Für einen Überblick siehe 2.2 und 2.3.

### Induktives Vorgehen

Man beginnt mit einer leeren Liste und einem historischen Text. Aus dem historischen Text entfernt man, zum Beispiel mithilfe eines modernen Wörterbuchs, alle bekannten Wörter. Ein menschlicher Betrachter prüft die verbliebenen Worte ob sie sich durch Transformationen, also Ersetzungen von Buchstaben oder Buchstabenfolgen, auf entsprechende moderne Wortformen zurückführen lassen. Ist dies der Fall wird das Ersetzungsmuster in der Liste notiert. Diesen Schritt wiederholt man beliebig oft, wobei nur neue Transformationen in der Liste notiert werden.

Das Verfahren soll hier beispielsweise an einem Vers von Walther von der Vogelweide (siehe A.1) gezeigt werden:

### Der Ausgangsvers

Vnder der linden  
an der heide  
da vnser zweier bette was  
da mugent ir vinden  
schone beide  
gebrochen blu<sup>o</sup>men vnd gras  
vor dem walte in einem tal  
tandaradei schone sanc dui nahtegal.  
[24, Vers 1]

### Erster Verarbeitungsschritt

Nach entfernen aller modernen Wörter bleiben folgende Zeichenfolgen übrig: (Zeilen werden zur besseren Orientierung beibehalten)

Vnder  
  
vnser  
mugent ir vinden  
  
bluomen vnd  
walte  
tandaradei sanc dui nahtegal.

### Zweiter Bearbeitungsschritt

Nun werden systematische Zusammenhänge zwischen modernen und historischen Wortformen gesucht. Siehe dazu Tabelle 3.4.

Worte	moderne Entsprechung	Ersetzung um zu einem modernen Wort zu gelangen
Vnder	Unter	$v \rightarrow u, d \rightarrow t$
vnser,vnd	unser,und	$v \rightarrow u$
mugent	möget	$u \rightarrow ö, ent \rightarrow et$
ir	ihr	$i \rightarrow ih$
vinden	finden	$v \rightarrow f$
bluomen	blumen	$uo \rightarrow u$
walte	walde	$t \rightarrow d$
tandaradei	tandaradei	–
sanc	sang	$c \rightarrow g$
dui	die	$ui \rightarrow ie$
nahtegal	nachtigall	$ah \rightarrow ach, e \rightarrow i, l \rightarrow ll$

Tabelle 3.4: Induktives Verfahren zum Auffinden von Ersetzungsmustern

historische Vollform	moderne Vollform	Pattern und Position der Ersetzung
iech	ich	$i \rightarrow ie$ an Position 1
ihch	ich	$i \rightarrow ih$ an Position 1
jch	ich	$i \rightarrow j$ an Position 1
y ch	ich	$i \rightarrow y$ an Position 1
...	...	...

Tabelle 3.5: Beispiel für ein aus den gewonnenen Ersetzungsmustern erzeugtes Wörterbuch

## Ergebnis

Die gefundenen Zusammenhänge werden nun in Regeln übersetzt und als Ersetzungsmuster gespeichert. Man beachte dass sich die Richtung der Anwendung geändert hat, da die Ersetzungsmuster dazu dienen aus modernen Wörtern historische zu erzeugen, die gefundenen Muster aber historische Wörter auf ihre “Nachfahren” abbilden:

**Ersetzungsmuster**  $ach \rightarrow ah, d \rightarrow t, et \rightarrow ent, f \rightarrow v, g \rightarrow c, i \rightarrow e, ie \rightarrow ui, ih \rightarrow i, ll \rightarrow l, ö \rightarrow u, u \rightarrow uo, u \rightarrow v$

Die Stärke dieses Verfahrens zur Patternengewinnung ist seine Erweiterbarkeit und schnelle Durchführbarkeit.

### 3.4.4 Anwendung der Patterns auf das Wörterbuch

Die gefundenen Ersetzungsmuster werden auf das Wörterbuch angewandt um das vorgekochte Wörterbuch zu erzeugen. Dies könnte so aussehen wie in Tabelle 3.5.



moderne Vollform	Ersetzungsmuster	erzeugte historische Vollform
zeit	$i \rightarrow y$ an Position 3	zeyt
zeiht	$i \rightarrow ih$ an Position 3	zeyt
seit	$s \rightarrow z$ , an Position 1, $i \rightarrow y$ , an Position 3	zeyt
seiht	$s \rightarrow z$ , an Position 1, $i \rightarrow ih$ , an Position 3	zeyt
seiet	$s \rightarrow z$ , an Position 1, $ie \rightarrow y$ , an Position 3	zeyt
seid	$s \rightarrow z$ , an Position 1, $i \rightarrow y$ an Position 3 $d \rightarrow t$ , an Position 4	zeyt

Tabelle 3.6: Ambiguität der erzeugten Vollformen

## 3.5 Probleme des Verfahrens mit “vorgekochten” Wörterbüchern

Trotz der guten Anwendbarkeit treten einige Probleme bei diesem Verfahren auf:

### 3.5.1 Ambige Einträge

Die gleiche historische Zeichenfolge kann aus unterschiedlichen Lemmata mit unterschiedlichen Ersetzungsmustern erzeugt werden. Dies entspricht aber auch der Realität der historischen Texte. Siehe dazu Tabelle 3.6.

### 3.5.2 Überproduktion

Durch das Anwenden von Ersetzungsmustern auf moderne Wörter werden neben historischen Wörtern auch wiederum modernen Wörter erzeugt. Für Beispiele siehe Tabelle 3.7.

moderne Ausgangswort	Ersetzungsmuster	fälschlich produziertes historisches Wort
der	$e \rightarrow i$	dir
mut	$u \rightarrow uh$	muht
zeit	$z \rightarrow s$	seit

Tabelle 3.7: Überproduktion

Diese werden fälschlich als historische Wörter klassifiziert.

### 3.5.3 Wörter ohne moderne “Nachkommen” können nicht erkannt werden

Wörter die aus dem Sprachgebrauch ausgeschieden sind, wie zum Beispiel

- allhier
- hierob

- dero
- dannenhero

haben keinen “Nachkommen” im modernen Wortschatz und können deshalb nicht durch die Anwendung von Patterns erzeugt werden.

### 3.5.4 Abkürzungen

Ungewöhnliche und nicht mehr gebräuchliche Abkürzungen können mit dem Verfahren nicht erkannt werden. Hier einige Beispiele:

- hrn. für Herrn
- fl. fr Florint bzw. Gulden
- se. für seiner
- ao. für anno

### 3.5.5 Größe der erzeugten Lexika

Mit der Wortlänge und der Anzahl der Ersetzungsmuster nimmt die Größe des Lexikons mit “vorgekochten” Wortformen enorm zu. Zum Beispiel sieht Andreas Hauser in seiner Arbeit zur OCR Nachkorrektur auf Grund der enormen Größe davon ab das Lexikon direkt zu erzeugen. [11, S. 29]

## 3.6 Kombination der Verfahren zur Lösung der Probleme

Für sich genommen hat jedes der zuvor genannten Verfahren seine Stärken und Schwächen. Durch ihre geschickte Kombination ist es jedoch möglich die Schwächen der Verfahren abzumildern bzw. zu beseitigen.

Als Ausgangspunkt dient das Verfahren mit “vorgekochten” Lexikon, das mit Techniken der anderen beiden Verfahren verbessert wird.

### 3.6.1 Lösung für ambige Einträge

Die Ambiguität, die bei der Zuordnung auftritt, lässt sich maschinell bei der heutigen Textbasis nicht lösen. Deshalb ist es am einfachsten eine semi-automatische Lösung anzustreben; also den Benutzer diese Auflösung übernehmen zu lassen.

### 3.6.2 Lösung für die Überproduktion historischer Wörter

Diese Idee stammt aus dem Ausschlußverfahren. Schließt man moderne Wörter von vornherein von der Analyse aus können sie nicht fälschlich als historisch klassifiziert werden. Dies kann erreicht werden indem man sowohl ein modernes Lexikon, als auch das "vorgekochte" Lexikon verwendet und jedes Wort mit einem Lookup in beiden Lexika prüft. Findet sich der Eintrag bereits im modernen Lexikon ist es unwahrscheinlich, dass es sich um ein Wort des historischen Wortschatzes handelt.

### 3.6.3 Lösung für Wörter ohne Nachkommen

Da es sich um eine relativ kleine Gruppe von Wörtern handelt, ließe sich das Problem durch bloße Aufzählung lösen. Die Erkennung dieser Gruppe muss jedoch manuell durch einen menschlichen Nutzer erfolgen.

### 3.6.4 Lösung für nicht mehr verwandte Abkürzungen

Diese Gruppe ist von Wörtern ist nicht sehr groß, kommt aber realativ häufig vor. Das Problem lässt sich mit einem zusätzlichen Abkürzungslexikon lösen. Dieses muss jedoch manuell befüllt werden.

### 3.6.5 Umgehen des Erstellens großer Lexika

Diese Problem wurde von Ullrich Reffles Vaam gelöst. Siehe Abschnitt 4.2.3.



## 4 Graphical–User–Interface zur Erstellung historischer Wörterbücher

Es werden die Stärken der im vorigen Kapitel vorgestellten Verfahren kombiniert, um eine graphische Benutzeroberfläche zu schaffen, die es erlaubt, effizient Wörterbücher der historischen Sprache erstellen zu können. In diesem Kapitel werden die einzelnen Bausteine des GUIs beschrieben.

### 4.1 Ein Webinterface

Die enormen Fortschritte die die Webprogrammierung in den letzten Jahren gemacht hat, erlauben es, das GUI als Webinterface zu implementieren. Die sich daraus ergebende Vorteile sind unter anderem:

**Plattformunabhängigkeit** Das GUI ist Plattformunabhängig und kann je nach Vorliebe des Nutzers vom Computer aus (sei es Linux, Windows, Mac OS X), dem Mobiltelefon oder der Spielekonsole das GUI bedienen. Die einzige Voraussetzung ist ein standardkonformer Webbrowser.

**Mobilität** Das Interface ist weltweit erreichbar und kann unabhängig vom Standort des Nutzers aufgerufen werden. Das erlaubt sowohl das Arbeiten von zu Hause aus als auch von unterwegs.

**Simultaneität** Mehrere Nutzer können gleichzeitig mit dem GUI arbeiten.

**Zentralität** Die Daten liegen zentral auf einem Server. Dies erlaubt es die Datensicherheit besser zu gewährleisten und ermöglicht alle Nutzer mit dem gleichen Datenbestand arbeiten zu lassen.

**Skalierbarkeit** Die erlaubte Nutzeranzahl hängt stärker von der gegebenen Infrastruktur als vom Programm ab.

### 4.2 Basiskomponenten

Die in diesem Abschnitt beschriebenen Bestandteile des GUIs bilden das Grundgerüst der Anwendung und sind sozusagen Voraussetzung für die übrigen Tools.

### 4.2.1 Ruby on Rails

Das Webframework *Ruby on Rails*<sup>1</sup> stellt die grundlegenden Mechanismen zur Darstellung im Webbrowser und der Verbindung der einzelnen Komponenten zur Verfügung. Besondere Stärken des von David Heinemeier Hansson entwickelten Frameworks sind:

**Model-View-Controller-Konzept** Das Model-View-Controller-Konzept (kurz MVC) wird komplett umgesetzt. Durch Trennung von Darstellung, Modell und Steuerung ist es möglich flexibel zu Arbeiten. Die Fehleranfälligkeit des Programms wird dadurch verringert.

**ActiveRecord** ActiveRecord ist ein Datenbanklayer der es erlaubt Datenbanken als Objekte und Einträge als Instanzen zu behandeln. Neben der natürlichen Einbindung von Datenbankobjekten in die Programmiersprache erlaubt ActiveRecord das Benutzen unterschiedlicher Datenbanktypen ohne in die Programmlogik eingreifen zu müssen.

**Scaffolding** Scaffolding ist eine Technik die es erlaubt sehr schnell anhand eines Datenbankmodells eine lauffähige Testumgebung zu erstellen. Dazu wird ein Gerüst (engl. scaffold) aus Controllern, Modellen und Ansichten erstellt, das den eigenen Bedürfnissen angepasst werden kann.

**REST** Unterstützung des Representational-State-Transfer nach Roy Fielding. [2, S. 407] Dies hat eine Vereinheitlichung in der internen Programmkommunikation zur Folge.

**Erweiterbarkeit** Das Framework ist darauf ausgelegt mit eigenen Modulen, Bibliotheken und Komponenten erweitert zu werden. Daher ist es einfach eigenen Programmcode und Fremdkomponenten einzubinden.

**Ruby** Das Framework ist in der von Yukihiro Matsumoto erdachten Scriptsprache *Ruby*<sup>2</sup> geschrieben. Aufgrund der großen Community um Ruby existieren zahlreiche Pakete für wiederkehrende Aufgaben die verhindern dass man “das Rad ständig neu erfinden” muss.

### 4.2.2 CISLEX

Das CISLEX ist ein am Centrum für Informations- und Sprachverarbeitung entwickeltes Vollformenlexikon der deutschen Sprache. Ziel des CISLEX-Projekts ist es ein weitgehend

theorieneutrales, vollständiges Wörterbuch der deutschen Sprache

zu schaffen[7, S. 1].

Der Aufbau und das Postulat der Theorienneutralität sind die perfekten Voraussetzungen dafür es in eigenen Programmen einzusetzen. Das CISLEX bildet die Grundlage für das Wörterbuch der modernen Sprache, das von Vaam erzeugte hypothetische Lexikon und wird auch vom Flector benutzt.

---

<sup>1</sup><http://www.rubyonrails.org/>

<sup>2</sup><http://www.ruby-lang.org>

### 4.2.3 VAAM

Das von Ulrich Reffle entwickelte Tool VAAM realisiert das hypothetische Lexikon. Reffle Beschreibt die Arbeitsweise von Vaam wie folgt:

Das [...] hypothetische historische Lexikon setzt sich wie bereits dargestellt aus allen orthographischen Varianten zusammen, die sich anhand eines modernen Lexikons und einer spezifizierten Menge von Rewrite-Mustern produzieren lassen. Dieses hypothetische Lexikon ist in der Praxis um ein Vielfaches zu groß, um explizit im Speicher gehalten zu werden. Stattdessen dient ein eigens hierfür entwickelter Algorithmus dazu, das Lexikon zur Laufzeit anhand von modernem Lexikon und Rewrite-Mustern zu simulieren. Das Modul Vaam (Variant-Aware Approximate Matching) übersetzt hierzu das moderne Lexikon sowie die Rewrite-Muster in endliche Automaten, um bei der Verarbeitung einer Anfrage effizient auf das moderne Lexikon zugreifen und gleichzeitig über die mögliche Anwendung von Rewrite-Mustern buchführen zu können.[15]

Das Interface zu Vaam wird per Ruby vom Modul `vam_interface.rb` bereitgestellt. (C.1)

### 4.2.4 Flector

Das von Annette Gotscharek entwickelte Flexionsprogramm<sup>3</sup> erlaubt es aus einer einzigen Form alle möglichen Paradigmen eines Wortes zu generieren. Es ist in der Lage sowohl mit einfachen Formen als auch mit Komposita zu arbeiten[10, S. 30]. Das Interface zum Flektionsprogramm wird im Modul `flexion_interface.rb` bereitgestellt. (C.2)

## 4.3 Korpus

Das Korpus ist die Textsammlung auf deren Basis das Lexikon erstellt wird. Es ist zum einen der Betrachtungsgegenstand als auch eine Quelle für Belege. Um größtmögliche Flexibilität zu gewährleisten, wird das Korpus als eine Mischung von Plain-Text-Dateien und Datenbank implementiert. Der Datenbankteil der Implementierung (siehe B.7) dient zur Verwaltung der Dateien und zum Verzeichnen von interessanten Daten wie Ursprungs-ort, Alter des Texts, Genre, Autor usw.. Die Texte selbst werden im Wurzelverzeichnis der Anwendung unter *korpora/* als reine Textdateien abgelegt, um eine Bearbeitung und Auswertung mit vom GUI unabhängigen Tools zu erlauben.

## 4.4 Lexika

Das Programm nutzt diverse Lexika und befüllt dies auch wiederum mit Daten.

---

<sup>3</sup> Genauerer zur Arbeitsweise des Tools in "Ein System zur Erstellung eines Kompositalexikons für das Deutsche"

#### 4.4.1 Das Lexikon der historischen Wortformen – HistLex

Das Lexikon *HistLex* ist das Kernlexikon der Anwendung. In ihm werden die vom Nutzer erarbeiteten historischen Wortformen gespeichert. Zu jeder historischen Wortform werden das moderne Lemma, die Flexionsklasse aus dem CISLEX, die verwendeten Ersetzungsmuster und die Belegstellen vermerkt. Anstatt der Belegstellen kann der Benutzer auch einen Konfidenzwert angeben. Seine Realisierung findet das Lexikon in der Tabelle *Historic*. (Siehe B.1 im Anhang)

#### 4.4.2 Das Lexikon der historischen Wortformen ohne modernen Nachfolger – HistLexWS

In diesem Lexikon werden alle historischen Wortformen gespeichert, die mit dem Verfahren des “Vorkochens” nicht beschrieben werden können, da keine moderne “Nachfolgeform” existiert. Gespeichert werden die historische Zeichenfolge und eine Worterklärung. Die Umsetzung erfährt das Lexikon in der Tabelle *HistoricWithoutAncestor*. Siehe Abschnitt B.3 im Anhang.

#### 4.4.3 Das Lexikon der historischen Abkürzungen – AbbrevLex

In diesem Lexikon werden Abkürzungen gespeichert, die in historischen Texten Verwendung finden. Dazu wird die aufgetretene Zeichenfolge, deren Langform und Belegstellen gespeichert.

Das Lexikon wird in der Tabelle *Abbreviation* realisiert. (Siehe B.2 im Anhang)

#### 4.4.4 Das Lexikon der problematischen historischen Wortformen – HistProbLex

Alle historischen Wortformen die zwar erklärt aber aus irgendwelchen Gründen nicht in eines der anderen Lexika eingefügt werden können werden in diese Lexikon eingetragen. Gespeichert werden der historische String, eine moderne Entsprechung und eine Anmerkung des Nutzers. Realisiert wird das Lexikon in der Tabelle *HistoricProblem*. (Siehe B.4 im Anhang)

#### 4.4.5 Das Lexikon der Entitäten – EntLex

In diesem Lexikon werden Entitäten gespeichert. Neben dem String<sup>4</sup> wird die Klasse der Entität gespeichert. Handelt es sich überdies um eine historische Schreibvariante wird die moderne Entsprechung zusätzlich abgespeichert. Es wird in der Tabelle *Entity* realisiert. (Siehe B.5 im Anhang)

---

<sup>4</sup>String und Zeichenkette werden hier analog verwendet.



#### 4.4.6 Das moderne Zusatzlexikon – ModLex

Diese Lexikon dient dazu Wörter der modernen Sprache aufzunehmen, die im CISLEX nicht enthalten waren. Es wird durch die Tabelle Modern realisiert. (Siehe B.6 im Anhang)

### 4.5 Benennung des GUIs

Das GUI wurde LeXtractor getauft. Der Name setzt sich aus “Lexikon” und “Extractor” zusammen und bedeutet zu deutsch etwa “Lexikonextrahierer”, was wiederum den Aufgabenbereich des GUIs recht gut umreißt.



# 5 Arbeiten mit LeXtractor

In diesem Kapitel werden die Funktionen des GUIs beschrieben.

## 5.1 Die Hauptansicht

Direkt nach dem einloggen befindet sich der Nutzer in der Hauptansicht des Programms. Über die zentral angelegten Links kann er von hier aus die verschiedenen Funktionen des GUIs aufrufen.

Folgende Funktionen stehen dem Nutzer zur Verfügung:

**Home** Die Indexseite; hier erhält der Nutzer Information über die verschiedenen Lexika. Unter anderem wie viele Einträge in den Lexika enthalten sind und wie viele Einträge der Nutzer selbst angelegt hat.

**Wort hinzufügen** Ein Wort dem Lexikon hinzufügen. Der Punkt wird unter 5.2 bis 5.4 genauer erläutert.

**Lexika** Die Lexika betrachten und direkt bearbeiten. Beschrieben in 5.5.

**Korpus** Unter diesem Menüpunkt kann der Nutzer Informationen zu Texten im Korpus abrufen, diese bearbeiten oder den Text aus dem aktuellen “Workflow” entfernen. und bearbeiten. Beschrieben in 5.6.

**Ersetzungsmuster** Informationen zu den Ersetzungsmustern abrufen oder Ersetzungsmuster erstellen. Beschrieben in 5.7.

**Auswertung** Unter diesem Menüpunkt sind verschiedene Programme vereint, die es u.a. erlauben, die Lexikonabdeckung zu berechnen, Zusammenhänge zwischen Ersetzungsmustern und Wörtern aufzuzeigen oder Voraussagen über noch möglicherweise auffindbare Wortformen zu treffen. Eine genauere Beschreibung der verschiedenen Funktionen findet sich unter 5.8.

**Letzte Aktion** Bringt den Nutzer zur zuletzt ausgeführten Funktion zurück.

**Hilfe** Hier kann der Nutzer sich mit dem Administrator der Anwendung in Verbindung setzen, wenn er Probleme hat.

**Logout** Dient zum verlassen des GUIs.

Dem Administrator stehen zusätzlich unter dem Menüpunkt Administration folgende Funktionen zur Verfügung:

**Configuration** Erlaubt die Anpassung des GUIs. Mehr dazu in 5.9.

**Nutzerverwaltung** Alles was mit dem Anlegen und Entfernen von Benutzern zu tun hat. Beschrieben in 5.10.

## 5.2 Wort zum Bearbeiten auswählen

Dem Nutzer stehen zwei Einstiegspunkte zum Sammeln von historischen Wortformen zur Verfügung.

### 5.2.1 Hochfrequente Wörter auffinden und bearbeiten

Dem Nutzer werden zwei frequenzgeordnete Listen präsentiert, eine mit den historischen Wortformen über alle (aktiven) Texte des Korpus und eine zweite mit als unbekannt klassifizierten Wörtern.

Dazu wird jeder String, jedes Texts in mehreren Lexika nachgeschlagen: Der erste Lookup läuft über die Lexika der Anwendung (HistLex, AbbrevLex, HistProbLex, HistLexWS, ModLex, EntLex). Findet sich der String dort wird er aus der Liste entfernt. In einem zweiten Lookupschritt werden das hypothetische Lexikon Vaam und das moderne Lexikon CISLEX befragt. Taucht das Wort im CISLEX auf wird es aus der Liste entfernt. Liefert Vaam eine Interpretation wird es der Frequenzliste der historischen Wortformen hinzugefügt. Liefern alle Lookups kein Ergebnis wird der String als unbekannt klassifiziert und der Liste der unbekannten Wörter hinzugefügt.

Anhand der Liste mit wahrscheinlich historischen Wortformen kann der Nutzer Wörter wählen, um sie in das historische Lexikon einzutragen. Die Liste der unbekannten Wortformen dient vor allem zum Auffinden von häufig genannten Entitäten oder Wörtern, die vom Verfahren nicht abgedeckt werden, wie Abkürzungen und Sonderbildungen.

Diese Arbeitsweise eignet sich vor allem für den Einstieg, um das Lexikon mit den wichtigsten/häufigsten Wortformen zu füllen.

Um die Auswahl etwas einzugrenzen, hat der Benutzer die Möglichkeit eine Schwelle anzugeben bis zu dieser die Wörter angezeigt werden sollen.

### 5.2.2 Am Text arbeiten

Der zweite Einstiegspunkt zum sammeln von Wörtern setzt an der Textbasis an. Der Nutzer wählt einen Text aus dem Korpus. Dieser Text wird anhand verschiedener Lexika abgearbeitet und die Wörter entsprechend ihres Typs markiert. Folgende Typen sind möglich:

**Modernes Wort** Das Wort gehört zum Wortschatz des modernen Deutsch. Es hat einen Eintrag im CISLEX oder im modernen Zusatzlexikon. Es wird nicht markiert.

**Historisches Wort** Das Wort gehört zum Wortschatz des historischen Deutsch. Es hat einen Eintrag im historischen Lexikon, HistLex. Es wird markiert und mit zusätzlichen Informationen zu seiner möglichen Herleitung aus verschiedenen modernen Lemmata versehen.

**Problematisches historisches Wort** Das Wort gehört zum Wortschatz des historischen Deutsch. Es hat einen Eintrag im historischen Sonderlexikon, HistProbLex. Es wird als historisch markiert und mit zusätzlichen Informationen zu seinen Nachfolgern und den von den Eintragenden gemachten Anmerkungen versehen.

**Historische Abkürzung** Es handelt sich um eine Abkürzung aus dem historischen Wortschatz. Sie hat einen Eintrag im Lexikon für historische Abkürzungen, AbbrevLex. Sie wird als historisches Wort markiert, zusätzlich wird die Langform der Abkürzung mit angegeben.

**Entität** Es handelt sich um eine Entität. Es gibt einen passenden Eintrag im Entitätenlexikon, EntLex. Das Wort wird als Entität gekennzeichnet. Zusätzlich wird der Typ der Entität und wenn es sich um eine historische Schreibweise handelt auch die moderne Form mit angegeben.

**Wahrscheinliche historisches Wort** Es handelt sich wahrscheinlich um ein Wort des historischen Wortschatzes. Der String wurde im hypothetischen Lexikon, Vaam, gefunden. Es wird als Kandidat für historische Wortformen gekennzeichnet und mit einem Link versehen der eine weiter Bearbeitung erlaubt (5.3).

**Unbekanntes Wort** Die Zeichenfolge konnte in keinem Lexikon gefunden werden. Das Wort wird als unbekannt markiert.

Für ein Beispiel siehe Grafik 5.2.2. Neben der farblichen Markierung wird, um den Benutzer die Arbeit zu erleichtern, eine Frequenzliste zu den historischen Kandidaten und unbekannten Wörtern des Texts eingeblendet. Diese erlaubt einen schnellen Überblick über die interessanten Wörter des Texts. Sie kann auch als Einstiegspunkt genutzt werden. Zusätzlich wird eine Statistik zum Text geführt, die zeigt in welchem Verhältnis die einzelnen erkannten Tokens zueinander stehen. Die Statistik enthält folgende Angaben:

- Anzahl der Tokens gesamt
- Anzahl der modernen Wörter und eine Prozentangabe welcher Anteil des Textes vom modernen Wörterbuch abgedeckt wird
- Anzahl der bekannten historischen Wörter, Abkürzungen und Sonderformen, sowie eine Prozentangabe wieviel vom Text aus diesen Wörtern besteht
- Anzahl der Entitäten plus Prozentangabe welchen Anteil sie am Text haben

- Anzahl der wahrscheinlich historischen Wörter und eine Prozentangabe, die den Anteil dieses Typs am Text anzeigt.
- Anzahl der unbekannten Wörter und eine Prozentangabe

Dies erlaubt dem Benutzer abzuschätzen, wieviel von diesem Text schon abgebeitet wurde und ob er noch eine lohnende Basis zur Wortgewinnung darstellt.

Diese Methode eignet sich vor allem um spezielle Fragen zu klären oder sich gezielt mit den Besonderheiten eines Textes auseinanderzusetzen.

## 5.3 Ein Wort aus dem hypothetischen Lexikon zum Lexikon der historischen Wortformen hinzufügen

Das Hinzufügen eines Wortes zum Lexikon läuft in mehreren Schritten ab.

### 5.3.1 Auswahl der Lesart

Im ersten Schritt werden dem Nutzer alle möglichen Interpretation auf Stringebene des Wortes präsentiert, d.h. es werden alle möglichen Lemmata mit den dazugehörigen Ersetzungsmustern angezeigt, die den historischen String auf einen, oder eventuell mehrere Nachkommen abbilden können. Der Nutzer kann nun die Interpretation, die im sinnvoll erscheinen auswählen und mit dem nächsten Schritt beginnen. Um dem Nutzer einen Überblick zu verschaffen und ihn bei der Auswahl der Lesarten zu unterstützen, wird zusätzlich eine nach Jahreszahl und Texten geordnete Konkordanz über den Korpus eingeblendet.

### 5.3.2 Zuordnung des Strings zu einem Paradigma

In diesem Schritt werden ambige Bedeutungen aufgelöst. Dem Nutzer werden alle möglichen Paradigmen, die anhand der gewählten Lesarten erzeugt werden können, nacheinander präsentiert. Dabei kann er sich stets "vor" und "zurück" bewegen, d.h. Interpretationen überspringen oder zuvor schon bearbeitete anders behandeln. Das GUI zeigt an bei welcher Lesart sich der Nutzer gerade befindet (siehe Grafik 5.3.2) und bei welchem Paradigma (siehe Grafik 5.3.2) der Nutzer sich befindet.

Um den Nutzer bei der Auswahl und Zuordnung zu unterstützen, wird das gesamte Paradigma mithilfe von **Flector** und dem **CISLEX** erzeugt und für jede Vollform des Paradigmas wird die Häufigkeit ihres Vorkommens im Korpus ermittelt. Die Summe aller Vorkommen wird aufaddiert, um einen "Score" zur Bewertung des Paradigmas zu erhalten. Das gesamte Paradigma wird zur besseren Lesbarkeit als Tabelle dargestellt. (Siehe Grafik 5.3.2).

Zusätzlich wird eine Konkordanz über die Textbasis eingeblendet, um ein Paradigma damit belegen zu können. Hat sich der Benutzer für ein Paradigma entschieden, kann er, wenn er einen passenden Beleg gefunden hat, das Wort zum historischen Lexikon hinzufügen.

Gibt es für ein Paradigma keine Belege hat der Nutzer die Möglichkeit einen Konfidenzwert anzugeben, in dem er die Qualität des Paradigmas bewertet und das Wort trotz fehlender Belegstelle(n) zum Lexikon hinzufügen kann.

### 5.3.3 Hinzufügen des Worts

Hat sich der Nutzer für eine Lesart und ein Paradigma entschieden und eine oder mehrere Belegstelle(n) bzw. einen Konfidenzwert angegeben, kann er das Wort in die Datenbank speichern/ in das Wörterbuch eintragen. Gespeichert werden dabei der historische String, die Lesart, das Paradigma, die Belegstellen bzw. Konfidenzwerte der Nutzernamen und der Zeitpunkt des Eintrags.

Sind noch Interpretationen offen (Lesarten oder Paradigmen) werden diese dem Nutzer zur weiteren Abarbeitung präsentiert und der Vorgang beginnt bei 5.3.2 erneut. Sind alle Interpretationen abgearbeitet, wird der Nutzer zur zuletzt ausgewählten Aktion zurückgebracht.

## 5.4 Sonstige Möglichkeiten zum Hinzufügen eines Worts

Handelt es sich um kein historisches Wort oder ein historisches Wort welches sich nicht im hypothetischen Lexikon befindet, stehen dem Nutzer eine Reihe zusätzlicher Funktionen zur Verfügung:

### 5.4.1 Wort zum Lexikon der historischen Wörter ohne Nachfolger hinzufügen

Hier kann der Nutzer Wörter in das moderne Zusatzlexikon eintragen. Dieses dient dazu Wörter der modernen Sprache aufzunehmen die nicht im CISLEX zu finden sind. Der Nutzer wird nach klicken auf diesen Button auf eine Eingabemaske geleitet, in der er folgende Angaben machen kann:

<b>H string</b>	Die in Bearbeitung befindliche Zeichenfolge (wird automatisch ausgefüllt)
<b>Beschreibung</b>	modernes Synonym, Anmerkungen

Zusätzlich werden die gewählten Belege angezeigt.

### 5.4.2 Wort zum Lexikon der historischen Abkürzungen hinzufügen

Hier kann der Nutzer historische Abkürzungen eintragen. Der Nutzer wird nach klicken auf diesen Button auf eine Eingabemaske geleitet in der er folgende Angaben machen kann:

<b>H string</b>	Die in Bearbeitung befindliche Zeichenfolge (wird automatisch ausgefüllt)
<b>Langform</b>	Die "ausgeschriebene" Form der Abkürzung

Zusätzlich werden die gewählten Belege angezeigt.

### 5.4.3 Wort zum Lexikon der problematischen historischen Wortformen hinzufügen

In dieser Maske kann der Nutzer historische Wortformen eintragen die in kein anderes Lexikon passen. Der Nutzer wird nach klicken auf diesen Button auf eine Eingabemaske geleitet in der er folgende Angaben machen kann:

<b>H string</b>	Die in Berarbeitung befindliche Zeichenfolge (wird automatisch ausgefüllt)
<b>M String</b>	Die moderne Wortform
<b>Anmerkung</b>	Besonderheiten des Worts, Vorschläge für Ersetzungsmuster

Zusätzlich werden die gewählten Belege angezeigt.

### 5.4.4 Wort dem modernen Zusatzlexikon hinzufügen

Durch einfaches Klicken wird das Wort dem Lexikon hinzugefügt.

### 5.4.5 Wort zum Entitätenlexikon hinzufügen

Der Nutzer wird nach klicken auf diesen Button auf eine Eingabemaske geleitet in der er folgende Angaben machen kann:

<b>E string</b>	Die in Berarbeitung befindliche Zeichenfolge (wird automatisch ausgefüllt)
<b>M String</b>	Die moderne Wortform, wenn es denn eine gibt
<b>Type</b>	D ropdownmenü mit folgenden möglichen Belegungen: <i>geo, person, group, name, div</i>

Zusätzlich werden die gewählten Belege angezeigt.

## 5.5 Lexika bearbeiten

Hier kann der Nutzer direkt in den Lexika arbeiten. Je nach Typ des Lexikons stehen ihm unterschiedliche Funktionen zur Verfügung, wie das manuelle Löschen, Hinzufügen und Editieren der Einträge.

### 5.5.1 Das Lexikon der historischen Wörter – HistLex

In diesem Lexikon befinden sich die historischen Wortformen die mit dem in 3.6 beschriebenen Verfahren ermittelt und dann wie in 5.3 beschrieben, gewonnen worden sind.

#### Hauptansicht – HistLex

Die Ansicht zeigt alle Einträge des historischen Lexikons alphabetisch sortiert an, die mit einem bestimmten Buchstaben beginnen. Folgende Buchstaben stehen zur Wahl:



*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 ö ü ß*

In der Hauptansicht erhält der Benutzer folgende Angaben:

**HistoricString** Das historische Wort bzw. die Zeichenfolge

**Lemma** die moderne Zeichenfolge, der Nachfolger des historischen Wortes

**FlexionClass** Die Flexionsklasse des Lemmas (vereinfachte Darstellung)

**Pattern** Die angewandten Ersetzungen um aus dem flektierten Lemma das historische Wort zu erzeugen

**Beleg** Die Belegnummer in der Belegtafel

**Gesichert** Konfidenzwert des Eintrags, bei belegten Einträgen stets 1

**User** Voller Name des Nutzers der den Eintrag gemacht hat

In dieser Ansicht hat der Nutzer darüberhinaus den Zugriff auf die Funktionen **Show**(Anzeigen) und **Destroy**(Löschen).

### Ansicht Show – HistLex

In dieser Ansicht erhält der Nutzer detailliertere Angaben zum zuvor gewählten Eintrag:

**FlexionClass** Volle Darstellung der Flexionsklasse

**Belege** Die Belegstellen werden dynamisch aus dem Korpus extrahiert und angezeigt.

### Destroy – HistLex

Diese Funktion steht dem Benutzer fakultativ zur Verfügung. D.h. der Nutzer ist nur berechtigt einen Eintrag zu löschen, wenn

- a) Der Nutzer ihn selbst angelegt hat
- b) Der Nutzer zur Gruppe der Administratoren gehört

## 5.5.2 Das Lexikon der historischen Wörter ohne Nachfahren – HistLexWS

In diesem Lexikon befinden sich die historischen Wortformen die keinen "Nachfolger" im modernen Sprachgebrauch haben. Der Nutzer erhält folgende Angaben:

**HistoricString** Der Eintrag, die Zeichenfolge des historischen Worts

**Description** Beschreibung des Worts, modernes Synonym

**Beleg** Beleg in der Belegetabelle

**User** Nutzer, der den Eintrag angelegt hat

Der Benutzer kann sich über **Show** Details zum Eintrag anzeigen lassen (Belegstellen) und wenn er den Eintrag angelegt hat oder zur Gruppe der Administratoren gehört diesen per **Destroy** löschen.

### 5.5.3 Das Lexikon der historischen Abkürzungen – AbbrevLex

In diesem Lexikon befinden sich Abkürzungen die in historischen Texten gebräuchlich waren. Die Hauptansicht zeigt folgende Einträge:

**HistoricString** Die Zeichenfolge, der Eintrag

**Long** Die Langform des Eintrags

**Beleg** Beleg in der Belegetabelle

**User** Nutzer, der den Eintrag angelegt hat

Der Benutzer kann sich über **Show** Details zum Eintrag anzeigen lassen (Belegstellen) und wenn er den Eintrag angelegt hat oder zur Gruppe der Administratoren gehört diesen per **Destroy** löschen.

### 5.5.4 Das Lexikon der problematischen historischen Wörter – HistProbLex

Die Einträge dieses Lexikons konnten aus diversen Gründen nicht in die anderen Lexika aufgenommen werden. Es dient als Sammelstelle für strittige Fälle und Sonderformen die zwar erklärbar, aber zu selten sind, um deren Mechanismen in das hypothetische Lexikon zu integrieren.

Der Nutzer erhält in der Hauptansicht folgende Angaben:

**HistoricString** Die Zeichenfolge, der Eintrag

**ModernString** Moderne Zeichenfolge des Worts

**Anmerkung** Grund des Eintrags im Sonderlexikon, Ersetzungsmuster die das Wort erklären könnten usw.

**Beleg** Beleg in der Belegetabelle

**User** Nutzer, der den Eintrag angelegt hat

Der Benutzer kann sich über **Show** Details zum Eintrag anzeigen lassen (Belegstellen) und wenn er den Eintrag angelegt hat oder zur Gruppe der Administratoren gehört diesen per **Destroy** löschen. Ist er zu neuen Erkenntnissen über das Wort gelangt kann er den Eintrag über **Edit** verändern.

### 5.5.5 Das Lexikon der Entitäten

In diesem Lexikon werden Strings gespeichert die Entitäten bezeichnen. In der Hauptansicht werden dem Nutzer folgende Informationen präsentiert:

**EntityString** Die Zeichenfolge, der Eintrag

**ModernString** Moderne Zeichen des Worts (optional)

**Type** Typ der Enität

**Beleg** Beleg in der Belegetabelle

**User** Nutzer, der den Eintrag angelegt hat

Der Benutzer kann sich über **Show** Details zum Eintrag anzeigen lassen (Belegstellen) und wenn er den Eintrag angelegt hat oder zur Gruppe der Administratoren gehört diesen per **Destroy** löschen. Mit **Edit** kann der Eintrag verändert werden.

### 5.5.6 Das moderne Zusatzlexikon

Dieses Lexikon dient zur Erweiterung des CISLEX. In der Hauptansicht werden alle Einträge aufgelistet. Der Nutzer kann wenn er Administrator ist oder den Eintrag angelegt hat das Wort aus dem Lexikon entfernen.

## 5.6 Korpus

In diesem Teil des GUIs können die Daten der einzelnen Texte des Korpus betrachtet und mit entsprechenden Rechten auch bearbeitet werden.

### 5.6.1 Korpus – Die Übersicht

In dieser Ansicht werden zu jedem Text des Korpus folgende Informationen angezeigt:

**Name des Textes** Überschrift oder erste Zeile des Dokuments. Durch klicken auf den Link gelangt man in den Bearbeitungsmodus **Wort nach Text hinzufügen** (siehe 5.2.2).

**Author** Der Autor des Textes.

**Year** Das Veröffentlichungsdatum der Textvorlage.

**Place** Erscheinungsort der Textvorlage.

**Genre** Die Textart der der Text zuzuordnen ist.

**Source** Die Quelle aus der das Dokument bezogen wurde. Durch klicken auf den Link gelangt man direkt zur originalen Textquelle.

**Quality** Qualität des Textes. 1 gut, 0 unbekannt.

**Active** Zeigt an, ob der Text zur Zeit im GUI benutzt wird

**Complete** Sind alle relevanten Wörter des Textes in die Lexika eingetragen worden? Die Belegungen sind fertig oder unfertig.

Über den Menüpunkt **Details anzeigen** kann der Nutzer zusätzliche Informationen zum Text abfragen; z.B. wer ihn eingestellt hat. Ein Nutzer mit Administratorrechten hat zusätzlich Zugriff auf die Funktion **Daten ändern**.

### 5.6.2 Korpus – Daten ändern

Unter Daten ändern hat der Nutzer folgende Möglichkeiten:

- Die Daten zu den Texten (Autor, Genre usw.) zu ändern
- Die Quelldatei, in der der Text gespeichert ist, anzugeben oder zu ändern
- Den Text zu aktivieren oder zu deaktivieren.
- Einen Text als abgearbeitet zu markieren

Das Deaktivieren eines Textes entfernt den Text aus dem normalen Programmfluss d.h. er wird nicht mehr für die Suche nach neuen Wörtern und für neue Belegstellen benutzt löscht diesen aber nicht.

## 5.7 Ersetzungsmuster

Hier kann der Nutzer die Ersetzungsmuster betrachten, die Vaam zur Erzeugung des hypothetischen Lexikons benutzt. Mit entsprechenden Rechten ist auch das Hinzufügen neuer Muster und das Löschen vorhandener Muster möglich.

### 5.7.1 Der Menüpunkt Ersetzungsmuster

Unter dem Menüpunkt Ersetzungsmuster werden dem Benutzer folgende Informationen präsentiert:

**Pattern** Das Ersetzungsmuster.

Alle Einträge folgen folgender Form:

*Zeichen(folge) des modernen Worts → Zeichen(folge) durch das das Zeichen des modernen Wortes ersetzt wird*

**User** Der Benutzer der das Muster angelegt hat

Unter **Show** kann sich der Nutzer zusätzliche Informationen anzeigen lassen. Mit ausreichenden Rechten hat er zudem die Möglichkeit Ersetzungsmuster mit **Destroy** zu entfernen.

### 5.7.2 Ersetzungsmuster hinzufügen

Die Möglichkeit Ersetzungsmuster hinzuzufügen steht nur Nutzern mit Administrationsrechten zur Verfügung.

Der Nutzer hat zwei Möglichkeiten neue Ersetzungsmuster bzw. Patterns zur Anwendung hinzuzufügen. Zum einen das präskriptive Vorgehen wie in 3.4.3 beschrieben als auch ein induktives Vorgehen wie in 3.4.3) beschrieben.

#### Ersetzungsmuster per Datenbank hinzufügen

Diese Möglichkeit entspricht dem präskriptiven Vorgehen. Der Benutzer gibt das Ersetzungsmuster, welches er zu verwenden gedenkt, direkt in die Tabelle Patterns ein. Dies geschieht unter dem Menüpunkt Patterns mit dem Befehl **New Pattern**.

Der Nutzer wird auf eine Formularseite geleitet auf der er ein Ersetzungsmuster eingeben kann. Dieses Muster wird auf seine Einzigartigkeit und korrekte Form geprüft. Sind beide Kriterien erfüllt wird es in Datenbank eingefügt.

#### Ersetzungsmuster beim Auffinden hinzufügen

Dies entspricht dem induktiven Vorgehen. In der Ansicht, die angezeigt wird nachdem der Nutzer sich entschieden hat ein Wort zu bearbeiten (siehe 5.3) befindet sich ein Formular in das der Benutzer ein Lemma und eine Folge von Ersetzungsmuster im Vaam-Format angeben kann.

Kann aus der Kombination von Lemma und Ersetzungsmuster ein Paradigma erzeugt und in die Datenbank **Historic** eingetragen werden, wird das Ersetzungsmuster automatisch der Tabelle **Patterns** hinzugefügt.

## 5.8 Auswertung

Unter diesem Menüpunkt sind verschiedene Tools versammelt die es dem Nutzer erlauben Verbindungen zwischen historischen und modernen Wortformen herzustellen, sich die Produktivität von Ersetzungsmuster anzeigen zu lassen oder Voraussagen über noch mögliche Einträge in die Lexika zu treffen.

### 5.8.1 Lemmata

Diese Ansicht erlaubt es, zu erfragen welche historischen Wortformen einem bestimmten Lemma zuzuordnen sind. Dazu kann der Nutzer über ein Dropdown-Menü das Lemma wählen und sich alle historischen Wörter anzeigen lassen die mit diesem Lemma in Verbindung stehen.

### 5.8.2 Muster

Hier kann man den Zusammenhang zwischen Ersetzungsmuster und Wort betrachten. Man kann sich die Produktivität aller Ersetzungsmuster anzeigen lassen. Die Darstellung entspricht der Form:

*Ersetzungsmuster : Anzahl der Wörter im Lexikon die durch diese Muster produziert werden konnten*

Beispiel:

i→y : 211  
 f→ff : 162  
 t→th : 161  
 k→ck : 94  
 k→c : 82  
 ie→i : 71  
 z→tz : 69  
 ä→e : 64  
 u→v : 53  
 s→ : 48

Als zweite Option kann man den Zusammenhang von Muster und Wortformen betrachten: Aus einem Dropdown-Mmenü wählt man das Ersetzungsmuster und bekommt dann alle Wörter des Lexikons angezeigt, die mit diesem Muster erzeugt wurden. Die Darstellung entspricht der Form:

*Historische Wortform : modernes Lemmma mit Flexion<sup>1</sup>*

Bespielsweise für das Ersetzungsmuster *au* → *aw*:

glawbens : glauben : neut(NS2,NPSG)#0:geN;1000000  
 glawbens : glaube : mask(NS6,NP4)#0:geM;2000000  
 glawbens : glauben : mask(NS2,NP0)#0:geM;1000000  
 fraw : frau : fem(NS0,NP3)#0:aeF:deF:geF:neF;0000000  
 ...

### 5.8.3 Korpora

In dieser Ansicht kann der Nutzer sich Daten zu Texten des Korpus über eine bestimmte Zeitspanne anzeigen lassen. Er erhält Information über die Anzahl der im Korpus vorhandenen Texte dieser Zeitspanne, der Types in dieser Spanne und der Anzahl der Tokens.

<sup>1</sup>Die Flexionsklassen sind in CISLEX-Notation angegeben

### 5.8.4 Lexikonabdeckung

Diese Ansicht erlaubt es die Lexikonabdeckung einer bestimmten Zeitspanne zu errechnen und auch Aussagen über die potentiell erreichbare Abdeckung zu treffen. Der Nutzer erhält folgende Informationen:

- Anzahl der Texte im gewählten Zeitraum
- Types im gewählten Zeitraum
- Tokens im gewählten Zeitraum
- Anzahl der Tokens in allen gesicherten Lexika und eine Prozentangabe zur damit erreichten Abdeckung.
- Abdeckung des modernen Lexikons: Types und Prozentangabe
- Abdeckung des Lexikons mit historischen Wortformen + Prozentangabe
- Abdeckung des Abkürzungslexikons + Prozentangabe
- Abdeckung des Lexikons mit historischen Wortformen ohne Nachfolger + Prozentangabe
- Abdeckung des Lexikons mit problematischen historischen Wortformen + Prozentangabe
- Abdeckung des Entitätenlexikons + Prozentangabe
- Abdeckung aller Lexika + Prozentangabe
- Abdeckung des hypothetischen Lexikons + Prozentangabe

Die Ausgabe kann beispielsweise so aussehen:

Ergebnis		
Zeitraum	Zwischen 1500 bis 1600	
Anzahl der Texte	13	
Tokens	31856	
Types	7949	
All Lex	24498	76.9023103967855%
Modern	18553	58.24020592667%
HistLex	5504	17.277749874435%
AbbrevLex	8	0.0251130085384229%
HistWosLex	66	0.207182320441989%
HistProbLex	367	1.15205926670015%
EntLex	233	0.731416373681567%
<b>Ergebnis mit hypothetischem Lexikon</b>		
Tokens	31856	
Types	7949	
All Lex + Hyp	28599	89.7758663987946%
Hypothetisches Lexikon	4101	12.873556002009%

## 5.9 Configuration

Diese nur dem Administrator zugängliche Ansicht erlaubt es, Einstellungen am GUI vorzunehmen. Dazu stehen folgende Menüpunkte zur Verfügung:

**Patterns wählen** Hier kann die Liste der Ersetzungsmuster gewählt werden, die Vaam zur Erzeugung des hypothetischen Lexikon verwendet.

**Datenbank reparieren** Entfernt fehlerhafte Einträge aus der Datenbank der historischen Wörter

**Neue Texte suchen** Nimmt neue Texte in den Korpus auf

## 5.10 Nutzerverwaltung

Hier können Nutzer angelegt, gelöscht und mit Rechten ausgestattet werden.

### 5.10.1 Hauptansicht Nutzerverwaltung

In der Hauptansicht wird eine Liste der Nutzer angezeigt die berechtigt sind mit dem GUI zu arbeiten. Durch das klicken auf [X] wird der Benutzer deaktiviert.

Unter dem Punkt **Neuen Nutzer anlegen** kann man neue Benutzer anlegen.



### 5.10.2 Neuen Nutzer anlegen

Mithilfe dieses Formulars kann man neue Benutzer anlegen. Folgende Felder sind auszufüllen:

**Name** Der Benutzername, mit dem sich der angelegte Benutzer authentifizieren soll.

**Passwort** Das Passwort, welches in Verbindung mit dem Benutzernamen ein Verwenden des Programms erlaubt.

**Bestätigung des Passworts** gleicher Inhalt wie das Feld Passwort, dient zur Überprüfung des Passworts

**Vollständiger Name** Der vollständige Name des Nutzers

**Rechte** Die Rechte die der Nutzer hat sind von der Gruppenzugehörigkeit abhängig. Zur Wahl stehen Administrator oder Benutzer.

Mit **Add User** wird der Nutzer hinzugefügt und kann sich dann im GUI einloggen.

## 5.11 Partial

Partial sind wiederkehrende Programmteile, die in mehreren Ansichten Verwendung finden.

### 5.11.1 Konkordanz

Die Konkordanzdarstellung dient in erster Linie dazu Belegstellen für ein Wort zu finden, dass einem Lexikon hinzugefügt werden soll. Die Konkordanzdarstellung hat folgenden Aufbau:

**Wort**

*Text*

[+] Belegstelle 1

[+] Belegstelle 2

Für die Darstellung im GUI siehe Grafik 5.11.1.

Durch klicken auf [+] wird die Belegstelle dem Belegstellenspeicher hinzugefügt. Durch klicken auf [**Text zeigen**] wird der Text aus dem die darunter stehenden Belege stammen in einem Popupfenster angezeigt, wobei jedes Vorkommen des Wortes farblich hervorgehoben wird.

### 5.11.2 Frequenzlisten

Die Frequenzlistendarstellung zeigt zwei frequenzgeordnete Listen an. Eine mit Einträgen aus dem hypothetischen Lexikon und eine zweite mit unbekannten Wörtern an.

### 5.11.3 State

Die Teildarstellung State zeigt welche Daten der Benutzer zu einem Wort gesammelt hat. Folgende Daten werden angezeigt:

**Wort** Das Wort, das aktuell bearbeitet wird

**Lesart** Die gewählte Lesart, mit *Ersetzungsmustern* und *Lemma*

**Paradigma** Das aktuelle Paradigma

**Belege** Eine Liste von Belegstellen, wobei Text und Position im Text angegeben werden

**Button “Wort hinzufügen”** Durch klicken dieses Knopfes wird das Wort dem Lexikon der historischen Wortformen hinzugefügt hinzugefügt. Er erscheint nur wenn alle zuvor genannten Datenpunkte erfüllt sind.

Für die Darstellung im GUI siehe Grafik 5.8 Hinter jedem Punkt, der in State angezeigt, wird befindet sich ein Link mit dem man dieses Feld zurücksetzen kann.

Grenze für Vorkommen wählen; Voreinstellung ist 5.

[neu berechnen] Time: 0.194946

<b>historische Wörter</b>	<b>unbekannte Wörter</b>
Anzahl: 9543	Anzahl: 9737
[+] 54: seind	[+] 103: hab
[+] 43: ime	[+] 62: ihme
[+] 25: wolt	[+] 33: het
[+] 23: vff	[+] 28: gehet
[+] 22: allmechtige	[+] 23: olgotzen
[+] 21: were	[+] 23: siessen
[+] 21: under	[+] 22: nimm
[+] 20: frey	[+] 19: seel
[+] 19: seye	[+] 18: adi
[+] 18: ausser	[+] 17: widrum
[+] 18: schuel	[+] 15: martinichen
[+] 17: guet	[+] 15: beschehen
[+] 17: ine	[+] 14: solchs
[+] 16: darbey	[+] 14: genennet

Abbildung 5.1: Wort nach Häufigkeit einfügen

Abbildung 5.2: Am Text arbeiten um ein Wort zu wählen

### Außzug etlicher Zeitungen von der Türcken Kriegshandlungt

Außzug etlicher Zeitungen von der Türcken Kriegshandlung

Außzug etlicher

Zeitungen / von der Türcken Kriegshandlung vor Zigeth / vnd andern orten im Königreich Hungern / auch auf dem Adriatischen Meer. 1566.

Gedruckt zu Nürnberg / durch Valentin Geyßler.

Erstlich *wirdet* vom XII. Augusti aus Wien geschriben / das der Ritterlich Graf von Serin/ den der Türckische Kayser *aigner* person / mit grossem gewalt / in dem Schloß zu *Sigeth* / *belogert* / einen *starcken außfall* gethan / vnnnd auf dem Berg / so der Türck mit einer *vnzabl* vnd menge der *Schantzknecht* / etlich viel tag vnnnd nacht *aufgeworffen* / vnd das Schloß *vberhöhen* wollen / biß in 600. *Janitzschern* / so die *Schantzknecht* *verwaren* sollen / erstochen / vnd der *Schantzknecht* auch ein guten *theil vmbgebracht* haben soll.

Also sol auch aus gedachtem Schloß / aus einer *Notschlange*/ ein Schuß durch des *Türckischen Keyzers* selbst *leybszeiten* geschehen sein.

Item / Es hat auch gedachter Graf aus *benelb* der *Röm. Kay. May*: mit seinem *untergebenem treflichem Kriegßvolck* / so zu Roß vnd fuß biß in 4000. *starck* / alle *Propbant* in derselbigen *gantzen gegendt* / hinein *inn* die Besatzung gebracht / vnd das *vberig* rings weiß *vmb jhn herum* / *dermassen verbrent* / das der Türck mangels halben der *Propbant aigner* person / mit viel *Volcks* wider abgezogen ist / Ob aber der Rest desselbigen *Volcks* auch abziehen / oder die *belegerung* noch *lenger* beharren werde / das gibt die zeit zuerkennen / Sonst ist der Graf vnd sein *Kriegßvolck* / vor solchem gewalt *gantz vnerschrocken* / *Got verleyhe* *weiter* gnad.

Ferner so ist ein *hauff Tarter*/ *inn* einem streif / *nechtlicher* weil / *nabendt* auf *Erla* kommen / vnd *inn* den *vmbliegenden Dörffern* *vil* armer *vnschuldiger* Christen / Mann vnd Weibs personen / auch *vil* junger Kinder *erbermlich* erschlagen vnd *ermordt* / auch hinter *jnen* / alles was sie gefunden / verbrennet.

Gleicher gestalt hat ein ander *hauff Tarter* in *Zipß*/ *nit* weit von des Herrn *Schwendiläger* / ein *einfaßl* gethan / vnd etlich viel armer Christen / aus *ein Dörff* hinweg *geführt* / Des hat ein *guthertziger Baußman* / gedachtem Herrn *Schwendi* so eilends angezeigt / das er gleich alsbald den *Balasij Melchior* / mit seinen *Vngerischen* ringen Pferden / *hinnaß* geschickt / das er ein guten *theil* derselben *Tartern* erlegt / vnd die armen Christen / so hungers halben schier halb *todt* gewesen / *Gotseliglich* wider erlediget hat.

### Interpretationen für den String "zeyt"

☑ **zeyt** läßt sich auf **seid** zurückführen mit:

Ersetzungsmuster: s→z an Position 1 i→y an Position 3 d→t an Position 4

☑ **zeyt** läßt sich auf **seiet** zurückführen mit:

Ersetzungsmuster: s→z an Position 1 ie→y an Position 3

☑ **zeyt** läßt sich auf **seiht** zurückführen mit:

Ersetzungsmuster: s→z an Position 1 ih→y an Position 3

☑ **zeyt** läßt sich auf **seit** zurückführen mit:

Ersetzungsmuster: s→z an Position 1 i→y an Position 3

☑ **zeyt** läßt sich auf **zeiht** zurückführen mit:

Ersetzungsmuster: ih→y an Position 3

☑ **zeyt** läßt sich auf **zeit** zurückführen mit:

Ersetzungsmuster: i→y an Position 3

☑ **zeyt** läßt sich auf **zäheit** zurückführen mit:

Ersetzungsmuster: äh→e an Position 2 ei→y an Position 4 [Diese Lesarten betrachten](#)

Bei zeyt handelt es sich um ein historisches Wort das auf

Abbildung 5.3: Auswahl der Lesarten

Lesarten

**teilen** lässt sich auf **tailen** zurückführen mit:  
Ersetzungsmuster: t→th an Position 1 ai→ei an  
Position 2 ll→l an Position 4

**teilen** lässt sich auf **teilen** zurückführen mit:  
Ersetzungsmuster: t→th an Position 1

[\[Eine Lesart zurück\]](#)

Abbildung 5.4: Anzeige der aktuellen Lesart

Paradigmen

teilen interpretiert als **Nomen (neut)**  
teilen interpretiert als **Nomen (mask)**  
**teilen interpretiert als Nomen (neut)**  
teilen interpretiert als **schwaches Verb**

[\[Ein Paradigma zurück\]](#) [\[Es fehlt eine Interpretation\]](#)

Abbildung 5.5: Anzeige des aktuell in Bearbeitung befindlichen Paradigmas

theilen als Nomen (neut)		
<b>PraeKontext</b>	<b>Wortform</b>	<b>Vorkommen</b>
das	theil	60
des	theils	24
des	theiles	0
dem	theil	60
dem	theile	22
das	theil	60
die	theile	22
der	theile	22
den	theilen	20
die	theile	22
<b>Score: 126</b>		

Abbildung 5.6: Darstellung des Paradigmas mit Score

## Konkordanz

### Teglich

**1601: Tagebuch des Hans Conrad Lang**

[Text zeigen]

[+] 3568: (und Theils bis Ich als hernach volget von Yßnj weck gezogen bin) **Teglich** vier praeceptores gehabt, und das volgender gestalt. Morgens umb 7, biswei

### teglich

**1553: Verhaltens Regeln Pestilenz 1553**

[Text zeigen]

[+] 11709: Ich habe dir derhalben so viel stuecke benuehmet / auff das du nit **teglich** einerley brauchen darffest / sondern abzuwechseln habest / das sein die nat

[+] 27901: sind / der ich droben gnugsam gedacht habe / vnnd gib dem krancken **teglich** mit ein / ein wenig guttes Theriacs / oder der Latwerge vom eyge / in Sawer

**1566: Außzug etlicher Zeitungen von der Türcken Kriegshandlungt**

[Text zeigen]

[+] 2970: uldigen Christen / sonder auch an seinen aigen trewisten Dienern / **teglich** mit grosser vnuernunft / viehisch vnd vmenschlich blutdurstiger weiß erze

[+] 6680: ay: noch auf dato zu Vngerischen Aldenburg [g]ewest / des vorhabens **teglich** zu dem hellen hauffen zuziehen / Ir May: ist mit sehr gutem Kriegßvolck /

**1601: Tagebuch des Hans Conrad Lang**

[Text zeigen]

[+] 50964: Ordnung helffen zu halten als möglich, mir monatlich 120 fl. oder **teglich** 4 fl. zu geben versprochen, dabei aber kein weiter serviz oder nichts zu ha

Abbildung 5.7: Darstellung der Konkordanz

## Derzeitige Auswahl

**Wort:** vortheil [-]

**Lesart:** vortheil läßt sich auf **vorteil** zurückführen mit:

Ersetzungsmuster: t→th an Position 4 [←]

**Paradigma:** Nomen (mask) [←]

**Belege:** korpusÜber das Marionettentheaterposition8064

[-]

[Wort hinzufügen]

Abbildung 5.8: Die Teildarstellung State





# 6 Testen des GUIs

In diesem Kapitel wird die Funktionalität des GUIs an einem kleinen Beispiel getestet.

## 6.1 Das Testkorpus

### 6.1.1 Die Ausgangsdaten

Das Korpus setzt sich aus Auszügen aus Texten des Wikisource Projekts zusammen. Es wurden pro Jahrhundert zwei Texte gewählt, jeweils einer aus der ersten und einer aus der zweiten Hälfte und mit dem script *100words.rb* (siehe Anhang C.3 ) die ersten hundert Wörter extrahiert. (Für die Texte siehe Sektion A.2 im Anhang.)

Wie das Linuxtool **wc** zeigt, hat jeder Beispieltext 100 Wörter:

```
wc *
 1    100    643 Brief_1723.txt_hundred.txt
 1    100    614 Comet_1532.txt_hundred.txt
 1    100    624 Hand_1579.txt_hundred.txt
 1    100   720 Krieg_dreissigjaehriger_1791.txt_hundred.txt
 1    100    676 Mordt_That_1606.txt_hundred.txt
 1    100    743 Vertrag_1871.txt_hundred.txt
 1    100    711 Wundermensch_1689.txt_hundred.txt
 1    100    740 padagogik_1803.txt_hundred.txt
 8    800   5471 total
```

### 6.1.2 Ausgangsdaten im GUI

Die Sichtweise des GUIs auf die Ausgangsmenge kann man Tabelle 6.1 entnehmen. Da nicht jede Zeichenfolge als Token interpretiert wird (zum Beispiel Zahlen gelten nicht als Token) weicht die Tokenmenge vom Ergebnis von **wc** ab. Das einzige befüllte Lexikon ist das der modernen Worformen, das CISLEX. Für alle Testtexte erreicht die Abdeckung der Lexika (hier: nur mit dem modernen Lexikon!) ungefähr 78,5 %. Betrachtet man das Ergebnis nach Jahrhunderten aufgegliedert, steigt die Abdeckung mit der (zeitlichen) Nähe zur modernen Sprache. (siehe dazu Tabelle 6.2)

Ergebnis: gesamt		
Zeitraum	Zwischen 1500 bis 2000	
Anzahl der Texte	8	
Types	489	
Tokens	781	
All Lex	613	78.4891165172855%
Modern	613	78.4891165172855%
HistLex	0	0.0%
AbbrevLex	0	0.0%
HistWosLex	0	0.0%
HistProbLex	0	0.0%
EntLex	0	0.0%

Tabelle 6.1: Testkorpus: Übersicht Ausgangsmenge

Zeitraum	Abdeckung durch modernes Lexikon
16. Jahrhundert	63.9593908629442%
17. Jahrhundert	73.1578947368421%
18. Jahrhundert	81.7258883248731%
19. Jahrhundert	94.9238578680203%

Tabelle 6.2: Testkorpus: Übersicht nach Jahrhunderten

## 6.2 Beschreibung des Ziellexikons

Das Ziel des Testlaufs ist es das “perfekte historische Lexikon”<sup>1</sup> für den Testkorpus zu erstellen. Aus diesem Grund wird mit einer leeren Menge von Ersetzungsmustern begonnen und die Ersetzungsmuster werden über die induktive Methode gewonnen und in einem zweiten Schritt klassifiziert.

## 6.3 Vorgehen

Die Texte werden jahrhundertweise abgearbeitet. Dabei wird beim 19. Jahrhundert begonnen und sich bis zum 16. Jahrhundert durchgearbeitet. Dies geschieht aus Effizienzgründen. Denn Ersetzungsmuster die in einem späteren Jahrhundert auftreten treten mit Sicherheit in einem früheren Jahrhundert auf. Der Umkehrschluss gilt aber nicht.

Die gefundenen Wortformen werden dann in die entsprechenden Lexika eingetragen.

Nach jedem abgearbeiteten Jahrhundert folgt eine Betrachtung der Einträge und der theoretisch erkennbaren Wörter der vorangegangenen Jahrhunderte.

Die e-Löschung und fehlende Kompositazerlegungen werden für diesen Testlauf mithilfe von Patterns und den Zusatzlexika simuliert.

### 6.3.1 Das 19. Jahrhundert

#### Vertrag1871

Der Text besteht aus 66 Tokens, davon werden 61 erkannt. Drei fehlen im modernen Lexikon (*Postvertrag*, *Dampfschiffsslinie*, *resp.*) und werden hinzugefügt. Zwei sind dem historischen Wortschatz zuzurechnen (*gesamnten*, *Postvertrage*).

HistorischeForm	moderne Form	Ersetzungsmuster
gesamnten	gesamnten	m→mm
Postvertrage	Postvertrag	e→nil

#### padagogik\_1803

Der Text besteht aus 86 Tokens, davon werden 81 vom modernen Lexikon erkannt ( 94%). Unbekannt sind fünf Tokens. Davon fehlen zwei im modernen Lexikon (D und ehedessen). Drei lassen sich auf Vorgängerformen zurückführen:

HistorischeForm	moderne Form	Ersetzungsmuster
Consistorialrath	konsistorialrat	k→c,t→th
Collegen	kollegen	k→c
studirenden	studierenden	ie→i

<sup>1</sup>Perfekt bedeutet hier dass das Lexikon alle (historischen) Wörter des Korpus erklärt und sonst keine Daten enthält die zu Fehlklassifikationen führen könnten. Es ist also nur für diese Textmenge perfekt.

Ergebnis: 19. Jahrhundert		
Zeitraum	Zwischen 1800 bis 1900	
Anzahl der Texte	2	
Types	138	
Tokens	197	
All Lex	190	100%
Modern	192	97.4619289340102%
HistLex	5	2.53807106598985%
AbbrevLex	0	0.0%
HistWosLex	0	0.0%
HistProbLex	0	0.0%
EntLex	0	0.0%

Tabelle 6.3: Testkorpus: 19. Jahrhundert bearbeitet

### Ergebnis nach 19. Jahrhundert

Das Ergebnis in Tabelle 6.3 zeigt, dass die Texte des 19. Jahrhunderts komplett erschlossen werden konnten. Welche Auswirkungen dies auf die gesamte Textmenge hat, zeigt Tabelle 6.4. Man sieht, dass das hypothetische Lexikon das Ergebnis um fast 2 Prozentpunkte verbessert, da die gefundenen Ersetzungsmuster auf Texte der vorigen Zeitalter angewandt werden können. Die Verbesserung durch die aufgefundenen und eingetragenen historischen Formen beläuft sich aber nur knapp auf 0,6 Prozentpunkte.

Gliedert man das Ergebnis genauer auf, zeigt Tabelle 6.5, dass im 16. Jahrhundert das hypothetische Lexikon einen Gewinn von 2,5 Prozentpunkten bringt, die anderen Lexika aber keine Steigerung der Abdeckung. Im 17. Jahrhundert bringt das hypothetische Lexikon eine Steigerung von 1,5 Prozentpunkten. Im 18. eine Steigerung von um die 2 Prozentpunkte.

Ergebnis: gesamt		
Zeitraum	Zwischen 1500 bis 2000	
Anzahl der Texte	8	
Types	488	
Tokens	780	
All Lex	623	79.7695262483995%
Modern	618	79.1293213828425%
HistLex	5	0.640204865556978%
AbbrevLex	0	0.0%
HistWosLex	0	0.0%
HistProbLex	0	0.0%
EntLex	0	0.0%
All Lex + Hyp	635	81.3060179257362%
Hypothetisches Lexikon	12	1.53649167733675 %

Tabelle 6.4: Testkorpus: Übersicht Ausgangsmenge nach Bearbeitung des 19. Jahrhunderts

Zeitraum	Abdeckung aller Lexika	alle Lexika + hypotetisches Lexikon
16. Jahrhundert	63.9593908629442%	66.497461928934%
17. Jahrhundert	73.1578947368421%	74.7368421052632%
18. Jahrhundert	81.7258883248731%	83.756345177665%
19. Jahrhundert	100%	100%

Tabelle 6.5: Testkorpus: Übersicht nach Jahrhunderten nach Bearbeitung des 19. Jahrhunderts

### 6.3.2 Das 18. Jahrhundert

#### Krieg1791

Der Text besteht aus 82 Tokens, davon deckt das moderne Lexikon 75 ab. Vier sind unbekannt, zwei können über das hypothetische Lexikon erkannt werden. Von den Unbekannten ist eines ein nicht erkanntes modernes Wort (*Glaubensverbesserung*), eines eine historische Schreibweise der Entität Europa (*Europens*). Das Wort *dreyssigjährigen* wird dem Problemlexikon hinzugefügt da, der Nachfolger nicht im modernen Lexikon steht und die hypothetische Wortform daher nicht erzeugt werden konnte.

HistorischeForm	moderne Form	Ersetzungsmuster
Beynahe	beinahe	i→y

#### Brief1723

Der Text besteht aus 80 Tokens, 45 davon wurden als Teil des modernen Wortschatzes erkannt. 22 unbekannte Formen und sechs, die sich über das hypothetische Lexikon hinzufügen lassen. Unter den 22 unbekannte finden sich zwei Wörter, die dem modernen Lexikon fehlen (*etc, Hochgeehrtester*), vier Entitäten (es handelt sich um die Kopfzeile eines Briefes, also Absender und Empfänger), zwei historische Abkürzungen (*ew, hochedelgeb*), zwei historische Wörter ohne Nachfolger (*insonders, wesfalls*) und ein Wort bei dem es sich wahrscheinlich um einen Tipfehler handelt (*hochgelalhrter*). Das Token *hochedelgebohrner* bereitet Probleme, da es sich um einen Titel handelt der nicht im modernen Lexikon vorhanden ist und daher die hypothetische Wortform nicht erzeugt werden kann. Das Pattern(o→oh) wird händisch zur Datenbank hinzugefügt, das Wort zum Lexikon der problematischen historischen Formen. Elf Wörter können dem historischen Wortschatz hinzugefügt werden. (einige kamen mehrmals im Text vor: *kan, wol*).

HistorischeForm	moderne Form	Ersetzungsmuster
wol	wohl	oh→o
kan	kann	nn→n
damahls	damals	a→ah
zumahl	zumal	a→ah
maszen	maßen	ß→sz
miener	meiner	ei→ie
freundt	freund	d→dt
beschleinigen	beschleunigen	eu→ei
gelauffen	gelaufen	f→ff

#### Ergebnis nach 18. Jahrhundert

Das Ergebnis in Tabelle 6.6 zeigt, dass das die Texte des 18. Jahrhunderts komplett erschlossen werden konnten. Es hat sich gezeigt, dass die fehlende Abdeckung des modernen Lexikons durch die historischen Lexika kompensiert werden kann. Welche Auswirkungen

Ergebnis: 18. Jahrhundert		
Zeitraum	Zwischen 1700 bis 1800	
Anzahl der Texte	2	
Types	151	
Tokens	196	
All Lex	196	100%
Modern	164	83.6734693877551%
HistLex	20	10.2040816326531%
AbbrevLex	2	1.02040816326531%
HistWosLex	2	1.02040816326531%
HistProbLex	3	1.53061224489796%
EntLex	5	2.55102040816327%

Tabelle 6.6: Testkorpus: 18. Jahrhundert bearbeitet

dies auf die gesamte Textmenge hat, zeigt Tabelle 6.7. Man sieht, dass die Abdeckung durch das hypothetische Lexikon um fast 3 Prozentpunkte verbessert werden konnte. Gliedert man das Ergebnis genauer auf zeigt Tabelle 6.8, dass im 16. Jahrhundert das hypothetische Lexikon einen Gewinn von 11 Prozentpunkten bringt, die anderen Lexika aber keine Steigerung der Abdeckung bewirken. Im siebzehnten Jahrhundert bringt zumindest das Lexikon der historischen Wortformen eine leichte Steigerung (um einen Prozentpunkt), das hypothetische Lexikon verbessert das Ergebnis noch mal um 6 Prozentpunkte.

Ergebnis: gesamt nach 18. Jahrhundert		
Zeitraum	Zwischen 1500 bis 2000	
Anzahl der Texte	8	
Types	488	
Tokens	780	
All Lex	660	84.6153846153846%
Modern	621	79.6153846153846%
HistLex	27	3.46153846153846%
AbbrevLex	2	0.256410256410256%
HistWosLex	2	0.256410256410256%
HistProbLex	3	0.641025641025641%
EntLex	5	0.0%
All Lex + Hyp	690	88.4615384615385%
Hypothetisches Lexikon	35	4.48717948717949 %

Tabelle 6.7: Testkorpus: Übersicht Ausgangsmenge nach Bearbeitung des 18. Jahrhunderts

Zeitraum	Abdeckung aller Lexika	alle Lexika + hypotetisches Lexikon
16. Jahrhundert	63.9593908629442%	75.6345177664975%
17. Jahrhundert	74.2105263157895%	80.5263157894737%

Tabelle 6.8: Testkorpus: Übersicht nach Jahrhunderten nach Bearbeitung des 19. Jahrhunderts



### 6.3.3 Das 17. Jahrhundert

#### Wundermensch1689

Der Text besteht aus 90 Tokens, von denen 67 im modernen Lexikon und eines im Lexikon der historischen Wortformen zu finden sind. Das hypothetische Lexikon deckt zu Beginn vier Wörter ab. Unter den unbekannten Tokens finden sich acht Entitäten (*Elisabetha, Rosina, Petri, Antonii, Consiglio, Biglia, Apuglia, Elisabeth*), ein Wort das im modernen Lexikon fehlt (*Leuten*), zwei nicht mehr in Gebrauch befindliche Wortformen (*alldorten, obbesagten*), drei Wörter die eine historische Flexion aufweisen (*dero, gebüset, gemeldte*) und drei Wörter die man durch neue Ersetzungsmuster erklären kann. Die Flexion wird in diesem Fall durch Patterns emuliert. Am Wort *wunderseltzame* scheitert die Kompositionserlegung sonst wäre es im hypothetischen Lexikon aufgetaucht; es wird dem Lexikon der problematischen historischen Wörter hinzugefügt.

HistorischeForm	moderne Form	Ersetzungsmuster
zeugniß	Zeugnis	s→ß
haußfrau	Hausfrau	s→ß
gebüset	gebüßt	ß→ss
beeden	beiden	ei→ee

#### Mordthat1606

Von den 66 Tokens aus denen der Text besteht sind 46 im modernen Lexikon. Sechs Tokens können durch das hypothetische Lexikon erklärt werden. Zwei werden zwar mit dem hypothetischen Lexikon erkannt, bei einem handelt es sich jedoch um eine Entität (*Reusischen*). Für das andere fehlt noch das Pattern, aber es würde eine andere Lesart geben (*Jah*). Von den elf unbekannten Tokens sind fünf Entitäten (*voigtlande, plawaischen, Leipzig, reusischen, plawischen*). Zum historischen Lexikon lassen sich noch sieben Wortformen hinzufügen. Das Wort *Aprilis* ist ein Sonderfall, da es noch in der lateinischen Form gebraucht wird. Es wird in das Lexikon der problematischen historischen Wörter eingetragen.

HistorischeForm	moderne Form	Ersetzungsmuster
Jah	Jahr	h→hr
maij	Mai	i→ij
vnd	und	u→v
vnerhörte	unerhorte	u→v
vmbgebracht	umgebracht	u→v , m→mb
sampt	samt	m→mp
gantzes	ganzes	z→tz

#### Ergebnis nach dem 17. Jahrhundert

Wie Tabelle 6.9 zeigt konnte das 17. Jahrhundert im Testkorpus abgeschlossen werden. Der Anteil der anderen Lexika an Abdeckung steigt. Betrachtet man die Auswirkung auf

Ergebnis: 17. Jahrhundert		
Zeitraum	Zwischen 1600 bis 1700	
Anzahl der Texte	2	
Types	150	
Tokens	190	
All Lex	190	100%
Modern	140	73.6842105263158%
HistLex	31	116.3157894736842%
AbbrevLex	0	0%
HistWosLex	2	1.05263157894737%
HistProbLex	3	1.57894736842105%
EntLex	14	7.36842105263158%

Tabelle 6.9: Testkorpus: 17. Jahrhundert bearbeitet

das gesamte Korpus (siehe Tabelle 6.10) sieht man, dass nun ein Großteil über die Zusatzlexika abgedeckt werden kann wobei die vom hypothetischen Lexikon abgedeckte Menge schrumpft. Dies liegt daran, dass die hypothetischen Wortformen bereits in einem vorherigen Schritt in die Lexika eingetragen werden konnten. Betrachtet man nun das verbleibende 16. Jahrhundert, erkennt man, dass die Abdeckung durch das historische Lexikon schon bei 6 Prozentpunkten liegt und das hypothetische Lexikon die Erkennungsrate noch einmal 15 Prozentpunkte erhöht.

Ergebnis: gesamt nach 17. Jahrhundert		
Zeitraum	Zwischen 1500 bis 2000	
Anzahl der Texte	8	
Types	488	
Tokens	780	
All Lex	721	92.4358974358974%
Modern	622	79.7435897435898%
HistLex	68	8.71794871794872%
AbbrevLex	2	0.256410256410256%
HistWosLex	4	0.512820512820513%
HistProbLex	3	0.641025641025641%
EntLex	19	2.43589743589744%
All Lex + Hyp	732	93.8461538461538%
Hypothetisches Lexikon	30	3.84615384615385 %

Tabelle 6.10: Testkorpus: Übersicht Ausgangsmenge nach Bearbeitung des 17. Jahrhunderts

### 6.3.4 Das 16. Jahrhundert

#### Die Hand mit auffgerekten dreyen schwartzen Fingern / So der falsche Eyd bedeut

Der text enthält 69 Tokens, 49 davon im modernen Lexikon. Zwei Tokens sind bekannte historische Wörter. Elf Tokens lassen sich durch das hypothetische Lexikon erklären. Zwei Tokens sind Entitäten (*Bresburck, Vngerlandt*). Das latinisierte *Septembris* wird dem Lexikon der problematischen historischen Wortformen hinzugefügt. *Fürgestellet* wird dem Lexikon der Wortformen ohne Nachfolger hinzugefügt. Die restlichen Tokens konnten dem historischen Lexikon hinzugefügt werden.

HistorischeForm	moderne Form	Ersetzungsmuster
wil	will	ll→l
jar	Jahr	ah→a
neue	neue	eu→ew
zeitung	Zeitung	tt→t

#### Was ein Comet sey: woher er komme / vnd seinen vrsprung habe / [...] (1532)

Der aus 79 Tokens bestehende Text enthält 45 Tokens die im modernen Lexikon verzeichnet sind. In den historischen Lexika findet sich ein Token. Das hypothetische Lexikon erkennt 14 Kandidaten, wovon alle richtig interpretiert worden sind. Von den verbleibenden 20 Tokens können dem Lexikon der historischen Wörter ohne Nachfolger (*beschehen, yenyemant*) zwei zugeordnet werden, eines dem modernen Zusatzlexikon (*weinmonat*), zwei dem Lexikon der problematischen historischen Wortformen (*Anzeygung, xxxij*) und drei dem Entitätenlexikon (*Nico, Laum, Prucknerum*). Die restlichen zwölf können in das historische Lexikon eingetragen werden.

HistorischeForm	moderne Form	Ersetzungsmuster
jhrer	ihrer	i→j
jrrigen	irrgien	i→j
himel	Himmel	mm→m
bedeutung	Bedeutung	eu→eü
zuo	zu	u→uo
geuolgt	gefolgt	f→u
taeglich	täglich	ä→ae
vnderscheidung	unterscheidung	u→v, t→d
mer	mehr	eh→e
erschinnen	erschienen	ie→i, n→nn
gwonlich	gewöhnlich	gew→gw, öh→o

#### Ergebnisse nach dem 16. Jahrhundert

Betrachtet man die Tabelle 6.11 sieht man dass ein großer Teil der Wörter aus den historischen Lexika stammt. Wenig überraschend zeigt sich in Tabelle 6.12 dass die komplette

Ergebnis: 17. Jahrhundert		
Zeitraum	Zwischen 1500 bis 1600	
Anzahl der Texte	2	
Types	135	
Tokens	196	
All Lex	196	100%
Modern	126	64.2857142857143%
HistLex	59	30.1020408163265%
AbbrevLex	0	0%
HistWosLex	3	1.53061224489796%
HistProbLex	3	1.53061224489796%
EntLex	5	2.55102040816327%

Tabelle 6.11: Testkorpus: 16. Jahrhundert bearbeitet

Abdeckung nur mit den historischen Lexika zu erreichen ist.

Ergebnis: gesamt nach 16. Jahrhundert		
Zeitraum	Zwischen 1500 bis 2000	
Anzahl der Texte	8	
Types	488	
Tokens	780	
All Lex	721	92.4358974358974%
Modern	622	79.7435897435898%
HistLex	68	8.71794871794872%
AbbrevLex	2	0.256410256410256%
HistWosLex	4	0.512820512820513%
HistProbLex	3	0.641025641025641%
EntLex	19	2.43589743589744%
All Lex + Hyp	779	100%
Hypothetisches Lexikon	0	0%

Tabelle 6.12: Testkorpus: Übersicht Ausgangsmenge nach Bearbeitung des 16. Jahrhunderts

## 6.4 Ergebnisse

Das Ziel, das “perfekte Lexikon” für den Testkorpus zu erstellen, konnte erreicht werden. Folgende Abschnitte beschreiben das Ergebnis genauer.

### 6.4.1 Zeitaufwand und schwierige Situationen

Mit dem Alter des Textes steigt der Zeitaufwand. Je älter der Text, um so mehr unbekannte Wortformen treten auf. Die hypothetischen Lexika leisten gute Dienste, können aber keine Sonderbildungen erkennen. Dies kann durch die Zusatzlexika kompensiert werden, ist aber zeitaufwendig.

Die Flexion in historischen Texten komplett zu bändigen wird noch eine Aufgabe für sich sein. Auch hier trifft es zu, dass mit steigendem Alter des Textes mehr unbekannte/ungewohne Flexionsmechanismen auftreten. Auch dialektale Einschläge sind in älteren Texten wesentlich stärker vertreten.

Der Zeitaufwand pro hundert Wörter liegt zwischen einer Minute (19. Jahrhundert), bis zu einer halben Stunde (16. Jahrhundert).

### 6.4.2 Die gewonnenen Ersetzungsmuster

#### Ersetzungsmuster im Überblick

Für das “perfekte Lexikon” des Testkorpus waren 42 Ersetzungsmuster nötig:

ä→ae, öh→o ß→ss, ß→sz, a→ah, ah→a, d→de, d→dt, dete→dte, eh→e, ei→ee, ei→ie, ester→ster, eu→eü, eu→ei, eu→ew, f→ff, f→u, gew→gw, hr→h, i→ij, i→j, i→y, ie→i, k→c, ll→l, m→mb, m→mm, m→mp, mm→m, n→nn, nn→n, o→oh, oh→o, r→ro, s→ß, t→d, t→et, t→th, t→tt, u→uo, u→v, z→tz

#### Bewertung der Ersetzungsmuster

Die gewonnenen Ersetzungsmuster sind nicht alle optimal. Um die besten für eine weitere Verwendung zu finden werden sie nach deren Ursprung/Ursache (für einen Überblick über die zur Klassifizierung herangezogenen Kriterien siehe Abschnitte 2.3-2.5) klassifiziert:

- Rechtschreibreform 1903: t→th, k→c
- Vokallängung durch Zusatzzeichen nicht angewandt: ie→i, oh→o ah→a, eh→e
- Vokallängung durch Zusatzzeichen, in der heutigen Orthographie nicht mehr üblich: a→ah
- Vokalkürzung durch Doppelung, in der heutigen Orthographie bei diesen Wörtern nicht mehr üblich: m→mm, nn→n, f→ff, n→nn, t→tt
- Vokalkürzung durch Doppelung nicht angewandt: ll→l, mm→m

Lexikon	Anzahl der Einträge
<b>Historische Wortformen (durch Ersetzungsmuster)</b>	148
<b>Historische Wortformen ohne Nachfahren</b>	7
<b>problematische Historische Wortformen</b>	8
<b>historische Abkürzungen</b>	2
<b>Entitätenlexikon</b>	23
<b>modernes Zusatzlexikon</b>	10

Tabelle 6.13: Lexikon des Testkorpus

- Laut durch Zeichenfolge ausdrücken:  $\beta \rightarrow sz, \beta \rightarrow ss, \ddot{a} \rightarrow ae$
- “Schreiben, wie man spricht” :  $d \rightarrow dt, m \rightarrow mb, m \rightarrow mp, ,i \rightarrow ij$
- Zeichen noch nicht getrennt, Position hat Sinn bestimmt :  $u \rightarrow v, f \rightarrow u$
- Vertrauen auf Sprachkenntnis des Lesers :  $h \rightarrow o, \rightarrow o, s \rightarrow \beta$
- Dialektaler Einfluss, Sprachwandel :  $t \rightarrow d, u \rightarrow uo, ei \rightarrow ie, eu \rightarrow ei, ei \rightarrow ee, eu \rightarrow ew$
- unbekannt :  $i \rightarrow j, eu \rightarrow e, hr \rightarrow h, i \rightarrow y, z \rightarrow tz$
- Flexion emulieren :  $ester \rightarrow ster, t \rightarrow et, r \rightarrow ro, dete \rightarrow dte, gew \rightarrow gw, d \rightarrow de$

Die meisten Ersetzungsmuster scheinen gut anwendbar zu sein. Die Klasse der Muster, die die Flexion simulieren, führt aber zu Fehleinordnungen und sollte deshalb im Produktiveinsatz entfernt werden. Betrachtet man die Reihen fällt auf, dass sie Löcher enthalten: zum Beispiel in der Reihe *Vokallängung durch Zusatzzeichen nicht angewandt* fehlt in der Vokalreihe nur noch das u. Diese Muster sollten bei einem grösseren Korpus auftreten. Diese Vermutung bestätigte sich im Produktiveinsatz.

### 6.4.3 Das gewonnene Lexikon

Das aus dem Testkorpus gewonnene historische Lexikon umfasst 165 Einträge. Dass die Zahl der Einträge wesentlich höher ist, als die Zahl der bearbeiteten Tokens liegt an der Besonderheit des Verfahrens auch nicht im Text vorhandene Lesarten zu betrachten und somit keine Interpretationsmöglichkeit eines Strings zu übersehen.

Für eine genaue Zusammensetzung des Lexikons siehe Tabelle 6.13, das gesamte Lexikon findet sich im Anhang D.



## 7 Schlußbetrachtung

Im vorigen Kapitel wurde gezeigt, dass es mithilfe von LeXtractor einfach möglich ist ein historisches Wörterbuch für einen kleinen Testkorpus zu erstellen.

Ein weiterer Testlauf mit 52 Texten die aus dem 16. Jahrhundert bis zum frühen 20. Jahrhundert stammen, wurde gestartet. Hierbei hat sich gezeigt, dass sich die im sechsten Kapitel gewonnenen Erkenntnisse auf eine größere Textmenge übertragen lassen.

Somit konnten die in der Einleitung gestellten Ansprüche erfüllt werden ein flexibles, robustes und leicht erweiterbares Graphical-User-Interface für die Erstellung historischer Lexika zu schaffen.

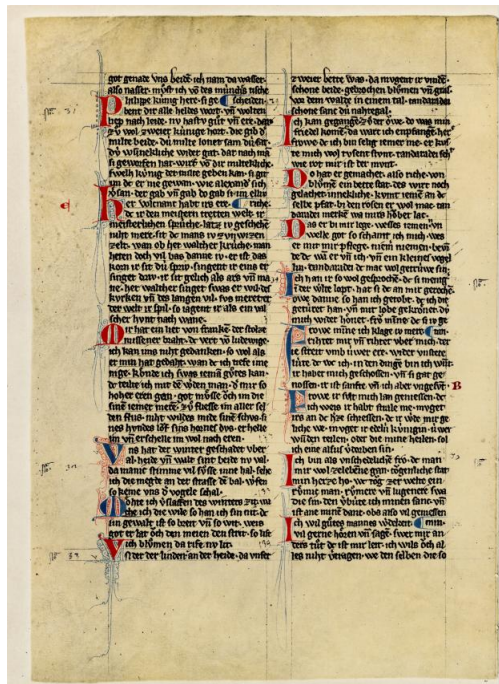


# A Textbeispiele

Hier befinden sich die Texte auf die aus der Arbeit heraus verwiesen wurde.

## A.1 Textbeispiel 13. Jahrhundert – Under der Linden (Codex Manesse)

### A.1.1 Scan des Originaltextes



[22, Blatt 130v]

Ein Auszug aus der "Manessischen Liederhandschrift". Diese Handschrift entstand um 1300 und enthält mittelhochdeutsche Dichtung. Das Textbeispiel beginnt in der linken Spalte in der letzten Zeile. Es handelt sich um ein Lied von Walther von der Vogelweide.

### A.1.2 Textinhalt

Vnder der linden  
an der heide  
da vnser zweier bette was  
da mugent ir vinden

schone beide  
 gebrochen bluomen vnd gras  
 vor dem walte in einem tal  
 tandaradei schone sanc dui nahtegal.

Ich kan gegangen  
 zvo der ovwe  
 do was min friedel komene  
 da wart ich enpfangen  
 here frowe  
 dc ich bin selig iemer me.  
 er kuste mich wol tusent stunt.  
 tandaradei seht wie rot mir ist der munt.

Do hat er gemachet  
 also riche  
 von bluomen ein bette stat  
 des wirt noch gelachet innekliche  
 kvmt iemen an dc selbe pfat  
 bi den rosen er wol mac  
 tandaradei merken wa mirs hovbet lac.

Das er bi mir lege  
 wesses iemen  
 nvn welle got so schamt ich mich  
 wes er mit mir pflege  
 niemer niemen  
 bevinde dc wan er vnd ich  
 vnd ein kleines vogellin  
 tandaradei dc mac wol getruiwe sin.[24]

## A.2 Der Testkorpus

Alle Texte stammen aus dem Wikisource-Projekt und sind auf die ersten hundert Wörter gekürzt.

### A.2.1 Was ein Comet sey: woher er komme / vnd seinen vrsprung habe / [...]

Was ein Comet sey wo her er komme vnd seinen vrsprung habe vnd vnder-scheidung vnd in was form vnd gestalt sye erscheynen Auch von jrer bedeütung mit anzeygung etlicher historien vnd geschichten so denen Cometen nach geu-olgt vnd sonderlich von dem Cometen erschinnen im Weinmonat des xxxij jars

Durch Nico laum Prucknerum beschriben Ein vorred zuo dem Leser ES ist gwonlich das sich die welt vil mer von kleinen doch vngewonten dingen verwundert dann von grossen so doch taeglich beschehen Dann yenyemant wundert der herzlich vnd gewaltig lauff des himels vnd der Sonnen vnd anderer beyder jrrigen so man Planeten

[14]

### **A.2.2 ie Hand mit auffgerekten dreyen schwartzen Fingern / So der falsche Eyd bedeut**

Die Hand mit auffgerekten dreyen schwartzen Fingern So der falsche Eyd bedeut Ein wahrhafftige vnd Erschreckliche Newe zeit tung Von einem Falschen Eyd welcher zu Bresburck im Vngerlandt den 24 Septembris im 1579 Jar von einem Messerschmidt vmb 4 gülden halben geschworen ist worden wie ihn auch Gott endlich gestraffet hat Als baldt er geschworen hat ist ihm die halbe Hand schwartz blieben vnd am dritten tag gestorben allen frommen Christen zu einem Exempel fürgestellt Die Hand mit auffgerekten dreyen schwartzen Fingern So der falsche Eyd bedeut Eln jeglich Mensch das ein Eyd schweren wil der sol auff heben drey

[19]

### **A.2.3 Erschreckliche / vnerhörte Mordt That / So sich den 28. Aprilis / dieses 1606. Jahres [] zugetragen / []**

Erschreckliche vnerhörte Mordt That So sich den 28 Aprilis dieses 1606 Jahres in der Reussischen Plawischen Herrschafft Lobenstein im Voigtlande zugetragen Da einer sein hochschwangers Eheweib seine sechs lebendige Kinder so wol auch die Magd vnd also sein gantzes Haußgesinde bis auff einen einzigen Knecht auff einmal jämmerlich ermordet vnd vmbgebracht Sampt einem bericht welcher gestalt der Mörder den 22 Maij dieses 1606 Jah res gerechtfertiget worden Gedruckt zu Leipzig im Jahr 1 6 0 6 Erschreckliche vnerhörte vnd erbärmliche Mordthat so sich den 28 Aprilis dieses 1606 Jahres in der Reusischen Plawischen Herrschafft Lobenstein zugetragen vnd wie solche allda

[20]

### **A.2.4 Beschreibung eines Wunder-Menschen / zu diesen unsern Zeiten entsprungen in der Neapolitanischen**

Beschreibung eines Wunder Menschen zu diesen unsern Zeiten entsprungen in der Neapolitanischen Landschafft Daß der gar zu grosse Weibische Fürwitz jederzeit seinen Frevel gebüßet hat gibt dessen klare Zeugniß Elisabetha Rosina

Petri Antonii Consiglio Eheliche Haußfrau wohnhafft in der Stadt Biglia in Apuglia Nach dero dann selbst eigener Aussage ist von diesen beeden Leuten dieses wunderseltzame Kind auf diese Welt herfür kommen Seiner Monstrosität oder Abscheulichkeit soll ein Ursprung gewesen seyn folgendes Die obbesagte arme Elisabeth begabe sich zum öfftern hinaus an das Gestade des Meers all-dort nothwendiger Lebens Mittel Auffenthalt zu erheischen Diese aber gemeldte Gegend oder Ende des

[21]

### A.2.5 Brief an Christian Wolff

Brief von Orffyraeus an Chr Wolff HochEdelgebohrner Vest und Hochgelahrter etc Insonders Hochgeehrtester Herr Hoff Rath werthster Freundt und Gönner Ew HochEdelgeb geehrtes und abends den 18 dieses richtig wieder wol erhaltenes setzt mich in Verwunderung dass wie und wann meine Antwort an Sie sonst wohin etwa gelauffen und verwechselt seyn solle weszfalls auch um die Rücksendung derselben freundlichst bitte um sehen zu mögen wo doch die antwort hinkommen und welcherley gattung Sie erhalten kan wol seyn maszen damahls 13 Briefe auf die Post zu beschleinigern hatte wie wol mir ein solches miener Tage nicht passirt zumahl in einer

[25]

### A.2.6 Geschichte des 30jährigen Kriegs

Historischer Calender für Damen für das Jahr 1791 GESCHICHTE DES DREYSIGJÄHRIGEN KRIEGS ERSTES BUCH Seit dem Anfang des Religionskriegs in Deutschland bis zum Münsterischen Frieden ist in der politischen Welt Europens kaum etwas großes und merkwürdiges geschehen woran die Reformation nicht den vornehmsten Antheil gehabt hätte Alle Weltbegebenheiten welche sich in diesem Zeitraum ereignen schließen sich an die Glaubensverbesserung an wo sie nicht ursprünglich daraus herfließen und jeder noch so große und noch so kleine Staat hat mehr oder weniger mittelbarer oder unmittelbarer den Einfluß derselben empfunden Beynahe der ganze Gebrauch den das Spanische Haus von seinen ungeheuern politischen Kräften

[17]

### A.2.7 über Pädagogik

Vorrede des Herausgebers fertig Nach einer älteren Verordnung musste ehedessen fortwährend auf der Universität Königsberg und zwar abwechselnd jedesmal von einem Professor der Philosophie den Studirenden die Pädagogik

vorgetragen werden So traf denn zuweilen auch die Reihe dieser Vorlesungen den Herrn Professor Kant welcher dabei das von seinem ehemaligen Kollegen dem Consistorialrath D Bock herausgegebene Lehrbuch der Erziehungskunst zum Grunde legte ohne sich indessen weder im Gange der Untersuchung noch in den Grundsätzen genau daran zu halten Diesem Umstande verdanken folgende Anmerkungen über die Pädagogik ihr Entstehen Sie würden wahrscheinlich interessanter noch und in mancher Hinsicht ausführlicher sein wenn der

[16]

**A.2.8 Additional-Artikel zu dem am 21. Oktober 1867 zwischen der Postverwaltung des Norddeutschen Bundes und der Postverwaltung der Vereinigten Staaten von Amerika abgeschlossenen Verträge für die Verbesserungen des Postdienstes zwischen den beiden Ländern, sowie zu dem Additional-Vertrage vom 7./23. April 1870.**

Additional Artikel zum Postvertrag zwischen dem Norddeutschen Bund und den Vereinigten Staaten von Amerika Wenn eine regelmäßige Dampfschiffslinie zwischen einem Hafen Deutschlands und einem Hafen der Vereinigten Staaten von Amerika zum Transport der Deutsch Amerikanischen Posten gegen eine solche Vergütung benutzt werden kann daß die gesammten Beförderungskosten zwischen den Grenzen der beiden Gebiete für jeden einfachen Brief Silbergroschen nicht übersteigen So haben die Unterzeichneten mit gehöriger Vollmacht von ihren Auftraggebern resp dem Deutschen Reiche und den Vereinigten Staaten von Amerika versehen sich über folgenden Additional Artikel zu dem Postvertrage vom 21 Oktober 1867 und zu dem Additional Vertrage vom

[1]





# B Aufbau der Tabellen

## B.1 Tabelle Historic

### B.1.1 Basistabelle – Historic

Das Lexikon der historischen Wortformen enthält pro Zeile folgende Felder:

**h\_string** Die Zeichenfolge, die ein historisches Wort darstellt

**lemma\_id** Id des Eintrags in der Tabelle Lemma, dort finden sich Grundform und die Flexionklasse des Eintrags

**beleg\_id** Id des Eintrags in der Tabelle Beleg, dort finden sich die Belegstellen für den Eintrag

**used\_patterns\_id** Id des Eintrags UsedPattern, stellt eine Verbindung zwischen der Tabelle Pattern und Historic her

**sure** Konfidenzwert, ist der Eintrag wahrscheinlich eine 1, ist er unwahrscheinlich eine 0

**user\_id** Zeigt wer den Eintrag angelegt hat

### B.1.2 Tabelle – Lemma

Die Tabelle Lemma enthält die von Historic benutzten Lemmata und über die Tabelle FlexionClass die verwandeten Flexionsklassen.

Aufbau:

**lemma** Das Lemma

**flexion\_classid** die dazugehörige Flexionsklasse aus der Tabelle FlexionClass

**user\_id** Zeigt wer den Eintrag angelegt hat

### B.1.3 Tabelle – FlexionClass

In dieser Tabelle sind die Flexionsklassen enthalten.

Aufbau:

**flexion** Die Flexionsklasse

**user\_id** Zeigt wer den Eintrag angelegt hat

### B.1.4 Tabelle – UsedPattern

Die Tabelle UsedPattern zeigt an an welcher Stelle im Lemma aus der Tabelle Lemma ein Ersetzungsmuster aus der Tabelle Pattern angewandt werden muss um die historische Form zu erzeugen.

### B.1.5 Tabelle – Beleg

In dieser Tabelle wird der Byteoffset eines Belegs und die ID des Korpus in dem das Wort aufgetreten ist, für ein Wort aus der Tabelle Historic gespeichert.

## B.2 Tabelle – Abbreviation

Umgesetzt wird das Lexikon in der Tabelle Abbreviation, die folgende Felder enthält:

**a\_string** Die Zeichenfolge, die eine historische Abkürzung darstellt

**long\_form** Ausgeschriebene Form der Abkürzung

**beschreibung** ein Feld für Anmerkungen

**beleg** Einen Verweis auf die Belegtablette

**user.id** Zeigt wer den Eintrag angelegt hat

## B.3 Tabelle – HistoricWithoutAncestor

Umgesetzt wird das Lexikon in der Tabelle HistoricWithoutAncestor, die folgende Felder enthält:

**h\_string** Die Zeichenfolge, der Eintrag

**description** hier wird die Wortbedeutung umschrieben oder es werden Synonyme angegeben

**beleg** Einen Verweis auf die Belegtablette

**user.id** Zeigt wer den Eintrag angelegt hat

## B.4 Tabelle – HistoricProblem

Umgesetzt wird das Lexikon in der Tabelle AbbHistoricProblemreviation, die folgende Felder enthält:

**h\_string** Die Zeichenfolge, der Eintrag

**m\_string** Die moderne Entsprechung

**beschreibung** ein Feld für Anmerkungen

**beleg** Einen Verweis auf die Belegtablelle

**user\_id** Zeigt wer den Eintrag angelegt hat

## B.5 Tabelle – Entity

Folgende Einträge sind möglich die folgende Felder enthält:

**e\_string** Die Zeichenfolge der Entität

**type** Art der Entität

**beleg** Einen Verweis auf die Belegtablelle

**user\_id** Zeigt wer den Eintrag angelegt hat

## B.6 Tabelle – Modern

Folgende Einträge sind möglich:

**m\_string** Die Zeichenfolge des modernen Worts

**user\_id** Zeigt wer den Eintrag angelegt hat

## B.7 Tabelle – Korpus

Folgende Einträge sind möglich:

**Name** Der Name des Texts

**Author** Der Autor

**Year** Entstehungsjahr

**Place** Ort der Veröffentlichung

**Genre** Die Textsorte

**Source** Die Quelle aus der der Text stammt, Internetadresse

**File** Die Textdatei, in der der Text auf dem Speichermedium liegt

**Genre** Die Textsorte

**Quality** Dient zur Bewertung, wie gut der Text geprüft und korrigiert wurde.

**Active** Aktiviert (1) oder deaktiviert (0) den Text.

**Complete** Zeigt an ob der Text abgerabeitet wurde (1) oder bicht (0).

**User** Id des Benutzers der den Eintrag angelegt bzw. ihn zuletzt bearbeitet hat.

# C Programmcode

## C.1 VamInterface

```
#!/usr/bin/ruby -w
#==Author: Andi Neumann
#==Date: 23.03.2008
#==Synopsis: Ruby-Schnittstelle zu Vaam f"ur Rails

$KCODE="u"

class VamInterface
  #Stellt eine Verbindung zu Vaam her
  def initialize( patterns )
    # Parameter zum Aufruf von vam
    # F"ur Freeze, einheitliche Patterns
    # patterns="/mounts/Users/student/andi/Lextractor/tmp/patterns"
    #####

    path_to_vam= #<Pfad zu Vaam>
    programm="#{path_to_vam}#<Ausf"uhrbare Datei>"
    levenshtein_distance="0"
    lexikon="#{path_to_vam}#<Lexikon>"
    aufruf="#{programm} #{levenshtein_distance} #{lexikon} #{patterns}"
    @vam=IO.popen(aufruf,"r+")
  end

  # Stellt eine Anfrage an Vam
  # Es wird eine Liste mit Antworten zur"uckgegeben
  def ask(anfrage)
    @vam << anfrage.downcase+"\n"
    @vam.gets.split("|")
  end

  # Gibt auf eine Anfrage ein Hash mit geordnet nach Typ der Antwort zur"uck
  # answers -> [:modern] --> [Liste mit modernen Formen]
  # answers -> [:cand_hist] --> [Liste mit historischen Formen]
```

```

# answers -> [:cand_entity] --> [Liste mit Entit"aten]
def ask_and_return_sorted(anfrage)
  sorted_answers = Hash.new()

  answers=self.ask(anfrage)

  sorted_answers[:modern]=  answers.select {|w| w =~ /\[\]\}/
  sorted_answers[:cand_entity]  =  answers.select {|w| w =~ /NONE/}
  sorted_answers[:cand_hist]=  answers.select {|w| w =~ /\[.+?\]\}/
  self.close()
  sorted_answers
end

#Hier wird ein Array erwartet, jede Antwort in einer Zeile
def ask_many(anfrage)
  antwort=[]
  for a in anfrage do
    @vam << a.downcase+"\n"
    antwort << @vam.gets
  end
  antwort
end

# Gibt ein Hash mit zwei Arrays zur"uck
# in h[:cand_hist] liegen die historischen Kandidaten
# in h[:cand_entity] liegen unerkannte W"orter
def ask_many_and_return_candidates(anfrage)
  candidates=Hash.new()
  antwort=[]
  for a in anfrage do
    @vam << a.downcase+"\n"
    antwort << @vam.gets
  end
  antwort.flatten!
  candidates[:cand_hist]=
  antwort.select {|w| w =~ /\[[^\]]+\]\}/.map {|w| w.sub!(/:.*$/,"").chomp!}.uniq
  candidates[:cand_entity]=
  antwort.select {|w| w =~ /NONE/}.map {|w| w.sub!(/:.*$/,"").chomp!}.uniq
  self.close()
  candidates
end

```

```

# Gibt ein Hash mit zwei Arrays zur"uk
# in h[:cand_hist] liegen die historischen Kandidaten
# in h[:cand_entity] liegen unerkannte W"orter
# in h[:modern] liegen moderne W"oter
def ask_many_and_return_all(anfrage)
  candidates=Hash.new()
  antwort=[]
  for a in anfrage do
    @vam << a.downcase+"\n"
    antwort << @vam.gets
  end
  antwort.flatten!
  candidates[:modern]=
  antwort.select {|w| w =~ /\[\]\}/.map {|w| w.sub!(/:.*$/,"").chomp!}.uniq
  candidates[:cand_hist]=
  antwort.select {|w| w =~ /\[[^\]]+\]/.map {|w| w.sub!(/:.*$/,"").chomp!}.uniq
  candidates[:cand_entity]=
  antwort.select {|w| w =~ /NONE/}.map {|w| w.sub!(/:.*$/,"").chomp!}.uniq
  self.close()
  candidates
end

def close()
  @vam.close()
end

end

__END__
#Test#
test=
VamInterface.new("<Datei mit Ersetzungsmustern>")
puts test.ask("da\ss{}")
#puts test.ask_many(["uns","da\ss{"},"that"])
puts test.ask_many_and_return_sorted(["uns","da\ss{"},"that"])
test.close

```

## C.2 FlexionInterface

```

#!/usr/bin/ruby -w
#==Author: Andi Neumann

```

```

#==Date: 23.03.2008
#==Synopsis: Ruby-Schnittstelle zu Flector f"ur Rails

$KCODE="u"

class FlexionInterface
  require "rexml/document"

  def initialize(mod_wort,patterns)
    @paradigmen=Hash.new()
    @wort=mod_wort

    #Parameter zum Aufruf von Annetes Tool
    programm="#<Pfad zum Programm+Programm>"
    aufruf="#{programm} --modStr=#{@wort} '--vs=#{patterns}' "

    @flexion=IO.popen(aufruf,"r+")
  end

  def ausgabe
    @formen=@flexion.gets(nil)
    puts @formen
  end

  # Parst die Ausgabe, gibt ein Hash zur"uck,
  # in dem unter jedem key ein Paradigma liegt,
  # jeder Eintrag im Paradigma besteht aus einem array
  # mit einem Prae Kontext und dem Wert
  def parse_ausgabe
    xml=@flexion.gets(nil)
    parsed_xml=REXML::Document.new(xml)

    #normaler Eintrag
    parsed_xml.elements.each("*/simplex/paradigma") do |paradigma|
      key,para_forms=scan_paradigma(paradigma)
      @paradigmen[key]=para_forms
    end

    #Kompositum
    parsed_xml.elements.each("*/kompositum/zerlegung/paradigma") do |paradigma|
      key,para_forms=scan_paradigma(paradigma)
      @paradigmen[key]=para_forms
    end
  end
end

```



```

end

def get_paradigm
  parse_ausgabe()
  @flexion.close()
  @paradigmen
end

def scan_paradigma(paradigma)
  p=[]
  paradigma.elements.each("form") do |form|
    # Zu jeder Form werden Merkmale der Klasse und des Kontexts gespeichert
    p << [
      form.text ||= "",
      form.attributes["merkmale"] ||= "",
      form.attributes["praeKontext"] ||= "",
    ]

    end
    key=paradigma.attributes["gf"]+"=>"+paradigma.attributes["lex"]+
    ""+paradigma.attributes["gut"]
    [key,p]
  end
end

end

__END__
#Test#
test=FlexionInterface.new("teil","(t_th,0)")
puts test.parse_ausgabe.inspect
test=FlexionInterface.new("hilfsmittel","(i_u,1)")
puts test.parse_ausgabe.inspect

```

## C.3 100words

```

#!/usr/bin/env ruby -wKU

$KCODE="u"

for f in ARGV do
  next unless f =~ /txt/

```

```
text=File.open(f).read(nil)
first_x_words=text.scan(/\w+/)[0..99]

target=File.open("./hundred/"+f+"_hundred.txt","w")
target.puts first_x_words.join(" ")
target.close()
end
```

# D Lexikon des Testkorpus

## D.1 Testkorpus – Historische Lexika

### D.1.1 Testkorpus – Lexikon der historischen Formen

String	Lemma	
antheil	anteil	Nomen (mask)
auff	auf	Präposition
auffenthalt	aufenthalt	Nomen (mask)
auffgereckten	aufgereckt	Adjektiv
auffgereckten	aufgereckten	NA
baldt	bald	Adverb
bedeutung	bedeutung	Nomen (fem)
beeden	beiden	Determinator
beschleunigen	beschleunigen	Nomen (neut)
beschleunigen	beschleunigen	schwaches Verb
beschriben	beschreiben	starkes Verb
beschriben	beschrieben	Adjektiv
beyder	beider	Determinator
beyder	beider	Pron
beynahe	beineinah	ADJu
calender	kalender	Nomen (mask)
collegen	kollege	Nomen (mask)
comet	komet	Nomen (mask)
cometen	komet	Nomen (mask)
consistorialrath	konsistorialrat	Nomen (mask)
damahls	damals	Adverb
dero	der	Determinator
drey	drei	Adjektiv
drey	drei	Nomen (fem)
dreyen	drei	Adjektiv
dreyen	drei	Nomen (fem)
erscheynen	erscheinen	Nomen (neut)
erscheynen	erscheinen	starkes Verb
erschinnen	erscheinen	starkes Verb
erschinnen	erschienen	Adjektiv

erschinnen	erschiene	Nomen (neut)
erschinnen	erschiene	schwaches Verb
eyd	eid	Nomen (mask)
freundt	freund	Nomen (mask)
gantzes	ganz	Adjektiv
gantzes	ganz	Adjektiv
gantzes	ganzes	NA
gebüsset	büßen	schwaches Verb
gebüsset	gebüßt	Adjektiv
gelauffen	laufen	starkes Verb
gelauffen	gelaufen	Adjektiv
gemeldte	gemeldet	Adjektiv
gemeldte	gemeldete	NA
gesammten	gesamt	Adjektiv
gesammten	gesamten	NA
gestraffet	straffen	schwaches Verb
gestraffet	gestrafft	Adjektiv
gestraffet	strafen	schwaches Verb
gestraffet	gestraft	Adjektiv
geuolgt	folgen	schwaches Verb
geuolgt	gefolgt	Adjektiv
grossen	groß	Adjektiv
grossen	großen	NA
gwonlich	gewöhnlich	Adjektiv
haußfrau	hausfrau	Nomen (fem)
haußgesinde	hausgesinde	Nomen (neut)
herrschaft	herrschaft	Nomen (fem)
himels	himmel	Nomen (mask)
jah	jahr	Nomen (neut)
jar	jahr	Nomen (neut)
jars	jahr	Nomen (neut)
jrer	ihrer	Pron
jrer	ihrer	Determinator
jrer	ihrer	Pron
jrrigen	irrig	Adjektiv
jrrigen	irrigen	NA
kan	können	unregelmäßiges Verb
lauff	laufen	starkes Verb
lauff	lauf	Nomen (mask)
maij	mai	Nomen (mask)
maij	mai	Nomen (mask)
maij	mai	Nomen (mask)
maszen	messen	starkes Verb

maszen	maß	Nomen (neut)
maszen	maß	Nomen (fem)
maszen	maße	Nomen (fem)
maszen	maßen	ASUFF
mer	mehren	schwaches Verb
mer	mehr	Adverb
mer	mehr	Nomen (neut)
messerschmidt	messerschmied	Nomen (mask)
miener	meiner	Determinator
miener	meiner	Pron
mordt	morden	schwaches Verb
mordt	mord	Nomen (mask)
mordthat	mordtat	Nomen (fem)
neue	neu	Adjektiv
neue	neue	NA
nothwendiger	notwendig	Adjektiv
passirt	passieren	schwaches Verb
passirt	passiert	Adjektiv
postvertrage	postvertrag	Nomen (mask)
rath	raten	starkes Verb
rath	rat	Nomen (mask)
sampt	samt	Präposition
sampt	samt	Nomen (mask)
schwartz	schwarz	Adjektiv
schwartz	schwarz	Nomen (neut)
schwartzzen	schwarz	Adjektiv
schwartzzen	schwarzen	NA
sey	sein	unregelmäßiges Verb
seyn	sein	Determinator
seyn	sein	Nomen (neut)
seyn	sein	unregelmäßiges Verb
studirenden	studierend	Adjektiv
studirenden	studierenden	NA
sy	sie	Pron
sy	sie	Pron
taeglich	täglich	Adjektiv
that	tun	unregelmäßiges Verb
that	tat	Nomen (fem)
vil	viel	PART
vil	viel	ADJu
vil	viel	ADJu
vil	viel	Determinator
vmb	um	Adverb

vmb	um	Konjunktion
vmb	um	Präposition
vmbgebracht	umgebracht	Adjektiv
vnd	und	Konjunktion
vnderscheidung	unterscheidung	Nomen (fem)
vnerhörte	unerhört	Adjektiv
vnerhörte	unerhörte	NA
vngewonten	ungewohnt	Adjektiv
vngewonten	ungewohnten	NA
vorred	vorreden	VSWT
vorred	vorrede	Nomen (fem)
vrprung	ursprung	Nomen (mask)
wahrhaftige	wahrhaftig	Adjektiv
wahrhaftige	wahrhaftige	NA
welcherley	welcherlei	Adverb
werthster	wert	Adjektiv
werthster	wertester	NA
wil	wollen	unregelmäßiges Verb
wohnhafft	wohnhafft	Adjektiv
wol	wohl	Adverb
wol	wohl	PART
wol	wohl	Nomen (neut)
zeitung	zeitung	Nomen (fem)
zeugniß	zeugnis	Nomen (neut)
zumahl	zumal	Adverb
zumahl	zumal	PART
zumahl	zumal	Konjunktion
zuo	zu	PART
zuo	zu	Konjunktion
zuo	zu	Präposition
öfftern	öfftern	Adverb
öfftern	öfftern	XINC

### D.1.2 Testkorpus – Lexikon der historischen Abkürzungen

String	moderne Langform
ew	euer
hochedelgeb	hochedelgeborener

### D.1.3 Testkorpus – Lexikon der historischen Wörter ohne Nachfolger

String	Synonym oder Erklärung
insonders	besonders
weszfalls	weshalb
alldort	dort
obbesagte	zuvor genannte
fürgestellt	gezeigt?
beschehen	geschehen
yenyemant	jemand

### D.1.4 Testkorpus – Lexikon der problematischen historischen Wortformen

String	moderne Entsprechung	Anmerkung
hochgelahrter	hochgelehrter	Tippfehler?
hochedelgebohrner	Hochedelgeborener	Titel, nicht mehr im Gebrauch
dreyssigjährigen	dreissigjährigen	Kein Eintrag im Lexikon
wunderseltzame	wunderseltsame	Kompositazerlegung
aprilis	April	
septembris	Sptember	latinisierte Form
anzeygung	Anzeige	heute andere Bildung
xxxij	32	römische Ziffer mit Druckfehler

## D.2 Testkorpus – Entitätenlexikon

String	moderne Entsprechun	Entitätentyp
antonii	Antonius	name
apuglia	Apuglia	geo
biglia	Biglia	geo
bresburck	Bratislava (Pressburg)	geo
chr	Christian	name
consiglio	Coniglio	name
elisabeth	Elisabeth	geo
elisabetha	Elisabeth	name
europens	Europas	geo
laum	Laum	geo
leipzig	Leipzig	geo
nico	Niko	name
orffyraeus	Orffyraeus	person
petri	Petrus	name
plawischen	plawischen	geo
prucknerum	Pruckner	geo
reusischen	reusischen	geo
reussischen	reussischen	geo
rosina	Rosina	name
vest	Vest	name
vngerlandt	Ungerland	geo
voigtlande	Voigtland	geo
wolff	Wolff	name

## D.3 Testkorpus – modernes Zusatzlexikon

String
d
ehedessen
postvertrag
dampfschiffslinie
resp
glaubensverbesserung
etc
hochgeehrtester
leuten
weinmonat



# Literaturverzeichnis

- [1] *Additional-Artikel zu dem am 21. Oktober 1867 zwischen der Postverwaltung des Norddeutschen Bundes und der Postverwaltung der Vereinigten Staaten von Amerika abgeschlossenen Verträge für die Verbesserungen des Postdienstes zwischen den beiden Ländern, sowie zu dem Additional-Vertrage vom 7./23. April 1870.* 1871 [http://de.wikisource.org/wiki/Additional-Artikel\\_zum\\_Postvertrag\\_zwischen\\_dem\\_Norddeutschen\\_Bund\\_und\\_den\\_Vereinigten\\_Staaten\\_von\\_Amerika](http://de.wikisource.org/wiki/Additional-Artikel_zum_Postvertrag_zwischen_dem_Norddeutschen_Bund_und_den_Vereinigten_Staaten_von_Amerika)
- [2] DAVE THOMAS, David Heinemeier H.: *Agile Web Development with Rails, Second Edition*. Raleigh-North Carolina, Dallas-Texas : The Pragmatic Programmers, 2007
- [3] DILBA, Eberhard: *Typographie Lexikon und Lesebuch für alle*. Nordstett : Books on Demand GmbH, 2005
- [4] DUDENREAKTION, Wissenschaftlicher R.: *Duden Band 4: Die Grammatik*. Mannheim/Leipzig/Wien/Zürich : Dudenverlag, 2006
- [5] DUDEN/WÜLSING/SCHMIDT: *Duden, Rechtschreibung der deutschen Sprache und der Fremdwörter*. Leipzig : Bibliographisches Institut Leipzig, 1926
- [6] FAULMANN, Carl: *Schriftzeichen und Alphabete aller Zeiten und Völker*. Wien : Kaiserlich-königliche Hof- und Staatsdruckerei, 1880
- [7] FRANZ GUENTHNER, Petra M.: *Das CISELX - Wörterbuchsystem*. 1994
- [8] GLÜCK, Helmut: *Metzler – Lexikon Sprache*. Stuttgart; Weimar : Metzler, 2000
- [9] GOOSSENS, Benoit: *Europe's cultural and scientific heritage at a click of a mouse*. Europe : European Commission Information Society and Media, 2007
- [10] GOTSCHAREK, Annette: *Ein System zur Erstellung eines Kompositalexikons für das Deutsche*. 2005
- [11] HAUSER, Andreas W.: *OCR Postcorrection of Historical Texts*. 2007
- [12] PROJECT, Impact: <http://www.impact-project.eu/about-the-project/main-goals/>
- [13] PROJECT, Impact: <http://www.impact-project.eu/about-the-project/objectives/>

- [14] PRUCKNER, Nicolaus: *Was ein Comet sey: woher er komme / vnd seinen vrsprung habe / [...]*. Straßburg : Johan Albrecht, 1532 [http://de.wikisource.org/wiki/Die\\_Hand\\_mit\\_auffgerekten\\_dreyen\\_schwartzten\\_Fingern/\\_So\\_der\\_falsche\\_Eyd\\_bedeut](http://de.wikisource.org/wiki/Die_Hand_mit_auffgerekten_dreyen_schwartzten_Fingern/_So_der_falsche_Eyd_bedeut)
- [15] REFFLE, Ullrich: *Aus einer privaten Email zu Vaam*. 2008
- [16] RINK, D. Friedrich T.: *Über Pädagogik*. Königsberg : D. Friedrich Theodor Rink, 1803 [http://de.wikisource.org/wiki/%C3%9Cber\\_P%C3%A4dagogik](http://de.wikisource.org/wiki/%C3%9Cber_P%C3%A4dagogik)
- [17] SCHILLER, Friedrich: *Geschichte des 30jährigen Kriegs*. 1791 - 1793 [http://de.wikisource.org/wiki/Geschichte\\_des\\_30j%C3%A4hrigen\\_Kriegs](http://de.wikisource.org/wiki/Geschichte_des_30j%C3%A4hrigen_Kriegs)
- [18] THOMA, Ludwig: *Briefwechsel eines bayrischen Landtagsabgeordnete*. München : Albert Langen, Verlag für Litteratur und Kunst, 1909
- [19] UNBEKANNT: *Die Hand mit auffgerekten dreyen schwartzten Fingern / So der falsche Eyd bedeut*. Wien : Steffan Kreutzer, 1579 [http://de.wikisource.org/wiki/Was\\_ein\\_Comet\\_sey:\\_woher\\_er\\_komme\\_vnd\\_seinen\\_vrsprung\\_habe](http://de.wikisource.org/wiki/Was_ein_Comet_sey:_woher_er_komme_vnd_seinen_vrsprung_habe)
- [20] UNBEKANNT: *Erschreckliche / vnerhrte Mordt That / So sich den 28. Aprilis / dieses 1606. Jahres [] zugetragen / []*. Leipzig, 1606 [http://de.wikisource.org/wiki/Erschreckliche/\\_vnerh%C3%B6rte\\_Mordt%\\_That](http://de.wikisource.org/wiki/Erschreckliche/_vnerh%C3%B6rte_Mordt%_That)
- [21] UNBEKANNT: *Beschreibung eines Wunder-Menschen / zu diesen unsern Zeiten entsprungen in der Neapolitanischen*. Wien, 1689 [http://de.wikisource.org/wiki/Beschreibung\\_eines\\_Wunder-Menschen\\_-\\_entsprungen\\_in\\_der\\_Neapolitanischen\\_Landschafft](http://de.wikisource.org/wiki/Beschreibung_eines_Wunder-Menschen_-_entsprungen_in_der_Neapolitanischen_Landschafft)
- [22] VOGELWEIDE, Walther von d.: In: *Under der linden in Codex Manesse*. 1300, S. 131r. – (Digitalisat der Uni Heidelberg)
- [23] WIKISOURCE: *Briefwechsel eines bayrischen Landtagsabgeordneten*. [http://de.wikisource.org/wiki/Briefwechsel\\_eines\\_bayrischen\\_Landtagsabgeordneten](http://de.wikisource.org/wiki/Briefwechsel_eines_bayrischen_Landtagsabgeordneten),
- [24] WIKISOURCE: *Under der linden (Codex Manesse)*. [http://de.wikisource.org/wiki/Under\\_der\\_linden](http://de.wikisource.org/wiki/Under_der_linden)
- [25] (ORFFYREUS), Johann Ernst Elias B.: *Brief an Christian Wolff*. St. Petersburg, Leipzig : Eggers et Comp., Leopold Voss, 1723 [http://de.wikisource.org/wiki/Orffyraeus\\_an\\_Christian\\_Wolff\\_11.\\_Mai\\_1723](http://de.wikisource.org/wiki/Orffyraeus_an_Christian_Wolff_11._Mai_1723)

# Lebenslauf

Andreas Neumann

- 27.11.1982    Geburt in München
- 1988 – 1992    Besuch der Grundschule in Höhenkirchen-Siegersbrunn
- 1992 – 2002    Besuch des Gymnasiums Neubiberg
- 2002            Abitur: Leistungskurse Englisch/Wirtschaft
- 2002 – 2003    Grundwehrdienst Roth / LMK-1
- 2003 – 2004    Studium an der Ludwigs-Maximiliansuniversität-München (LMU)  
Magister: Deutsch als Fremdsprache, Ethnologie, Politik
- 2004 –           Studium an der Ludwigs-Maximiliansuniversität-München (LMU)  
Magister: Computerlinguistik, Deutsch als Fremdsprache, Ethnologie
- 2006 – 2008    Tutor für C++, Perl, Prolog

**ERKLÄRUNG**

Hiermit versichere ich, dass ich diese Magisterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht. Dasselbe gilt sinngemä für Tabellen, Karten und Abbildungen. Diese Arbeit hat in dieser oder einer ähnlichen Form noch nicht im Rahmen einer anderen Prüfung vorgelegen.

---

(Ort, Datum)    (Unterschrift)