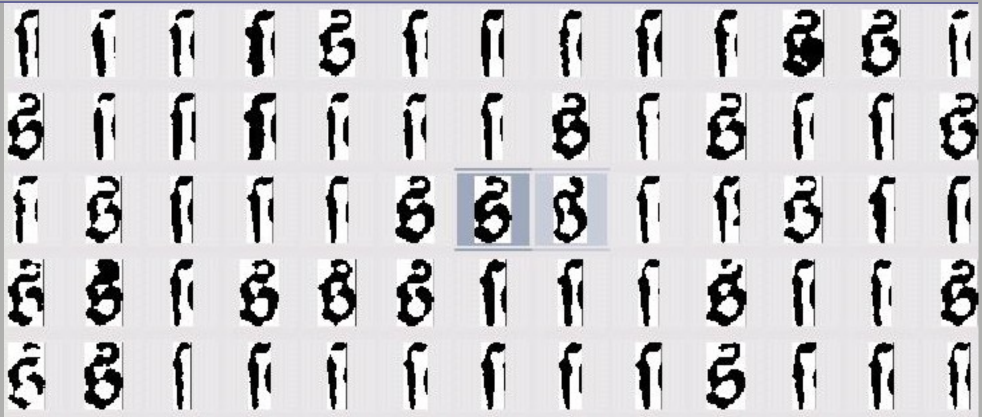


Constructing a Lexicon from a Historical Corpus

Annette Gotscharek, Andreas Neumann, Ulrich Reffle,

Christoph Ringlstetter, Klaus U. Schulz

CIS, University of Munich

AACL 2009, Edmonton, Canada

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Survey

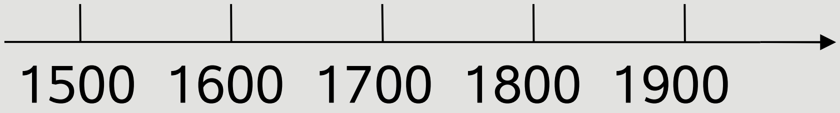
- 1. Motivate of historical lexica in digitization projects
- 2. A Preliminary historical corpus for German
- 3. Infrastructure for lexicon building
- 4. Evaluation and future work

0. Research Context IMPACT

- IMProve ACcess to Text, Improve mass digitisation of historical books, newspapers, legal documents...
- Four year EC project started in 2008
- 15 partners: national libraries, research institutions and commercial suppliers
- (1) Image Processing
- (2) Text Recognition: Adaption of Optical Character Recognition to historical documents
- (3) Lexicon Building: Enrichment of texts to Improve IR Access to historical documents
- (4) Capacity Building for Management of Digitization Projects

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

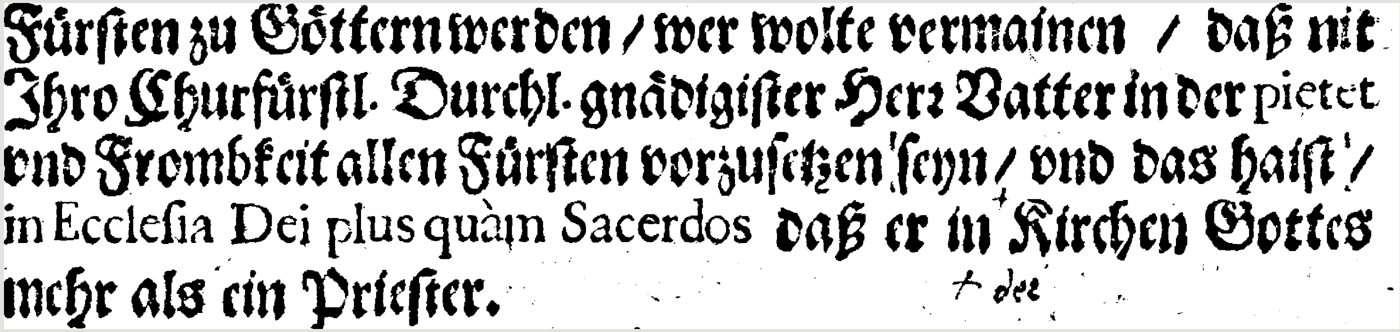
Special Challenges for the Digitization of Historical Materials



• Imaging	Damages on Originals
• **Optical Character Recognition**	**Rate of Recognition Errors**
• **Information Retrieval**	**Historical Variants**
• **Human Reading**	**Unknown Words**

Optical Character Recognition: medium quality

Fürsten zu Gstternwerden/wer wollte vermainen / daßwt
IhroKhurftrstl Durchl gnädiglsterHcttVatterinderpictcr
rndFrombkcltallmFürstenvorzusetzen!scyn/vnd das halst>
in^cclcQ^ vci pluz^uäzn 5accr6o5 daß tl iN KilchkN GottW
wehr als ein Priester.



Fürſten zu Göttern werden / wer wolte vermainen / daß nit
Jhro Churfürſtl. Durchl. gnädigiſter Herr Vatter in der pietet
vnd Frombkeit allen Fürſten vorzuſetzen ſeyn / vnd das haiſt /
in Eccleſia Dei plus quàm Sacerdos daß er in Kirchen Gottes
mehr als ein Prieſter. + dee

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

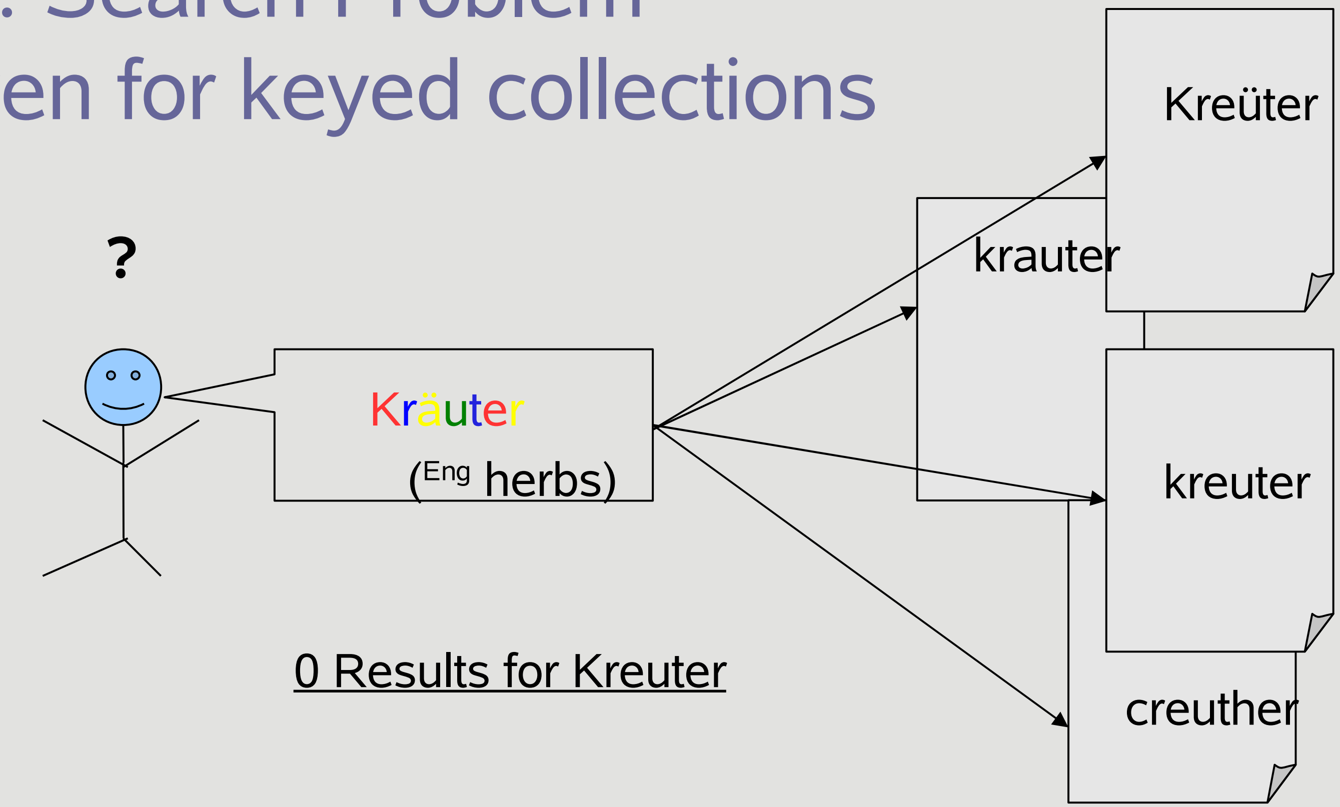
Why is it so bad

(1) The image quality is challenging, further processing needed

(2) The classifiers of the OCR disregard certain type faces used in historic print

(3) **The language resources of the OCR are inappropriate: historical language**

IR: Search Problem even for keyed collections


?

Kräuter

(Eng herbs)

Kreüter

krauter

kreuter

creuther

0 Results for Kreuter

– kräuter (= Engherbs) as kra‚uter, kreuter, kreüter, kreuter, creuther

Special challenge for IR and OCR on historical Texts: Spelling variation

- Missing normalization of orthography leads to many variants in historical documents, e.g. in German texts (1500-1850):
  - teil (= $^{\text{Eng}}$part) as theil, teyl, theyl
  - fragte (= $^{\text{Eng}}$asked) as frug, fruk
  - statt (= $^{\text{Eng}}$instead) as statt (=$^{\text{Eng}}$town)
  - ringer (= $^{\text{Eng}}$wrestler) as ringer (=$^{\text{Eng}}$ less)
- Resource: a lexicon mapping from historical variants to modern lemma

Language Resources need a Corpus Base

Possible sources:

- Keyed Materials on the Web
- Non Public Electronic Corpora
- Keying/corrected OCR of Image Corpora
- Noisy OCR Corpora

Status of German historical Corpora

1. First development corpus

- Proofread texts from 1400 to 1900: 4 institutional sources
- Medium size: 2.7 Mill. tokens
- For lexicon construction
- For preliminary diachronic analysis

2. IR annotated corpus for strategy testing

   Texts from 16th, 17th, 18th, 19th century 31080 tokens

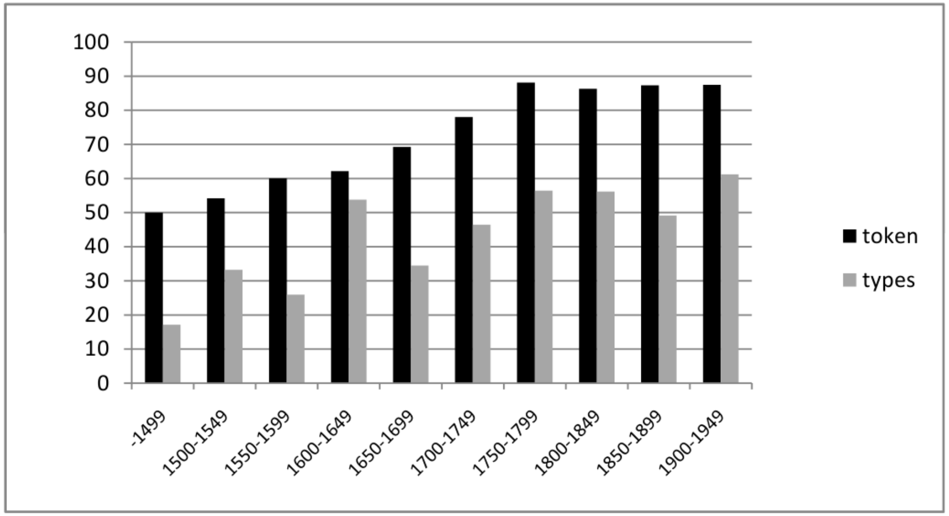
3. Project for corpus building 16th century started

4. Agreements on additional background materials from 1750s with 30 million tokens literature and 10 million tokens other

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Language in a historical corpus for German

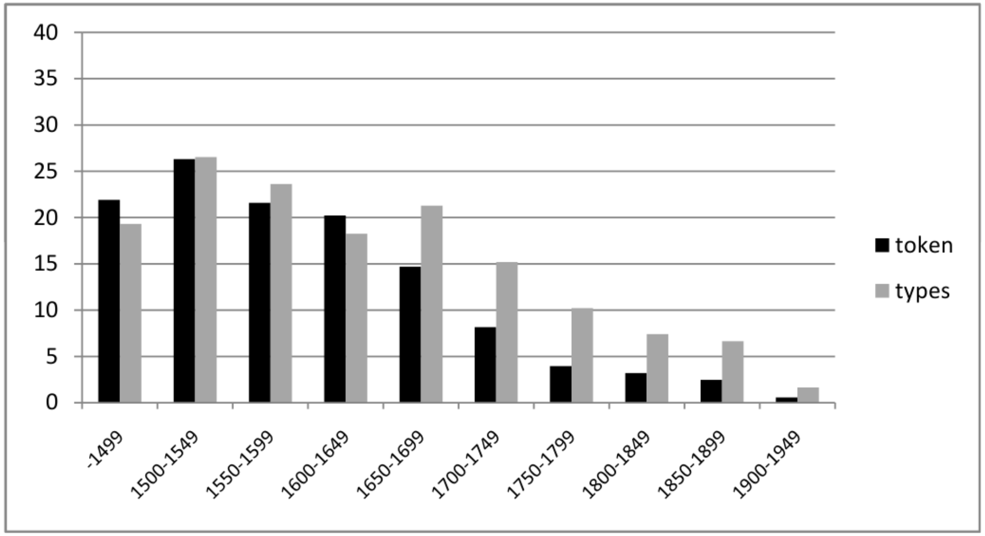
Modern lexicon (CISLEX): coverage for 10 time periods



IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Language in a historical corpus for German

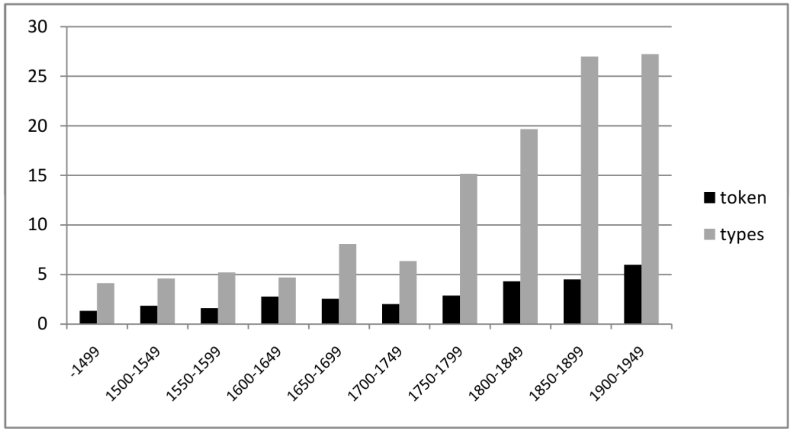
Pattern based lexicon; coverage for 10 time periods



IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

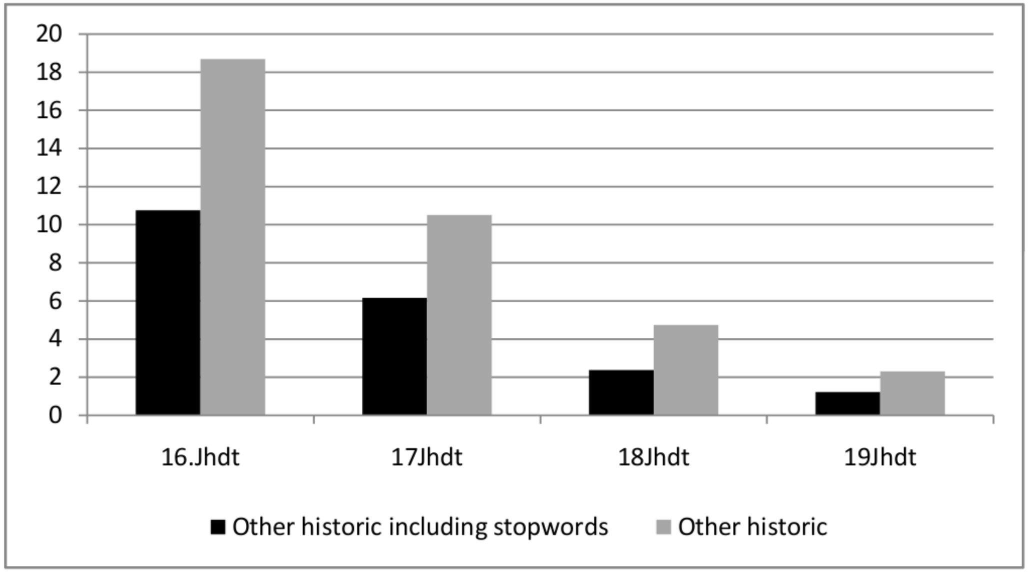
Language in a historical corpus for German

Compounds (modern components); coverage for 10 time periods



Test corpus:

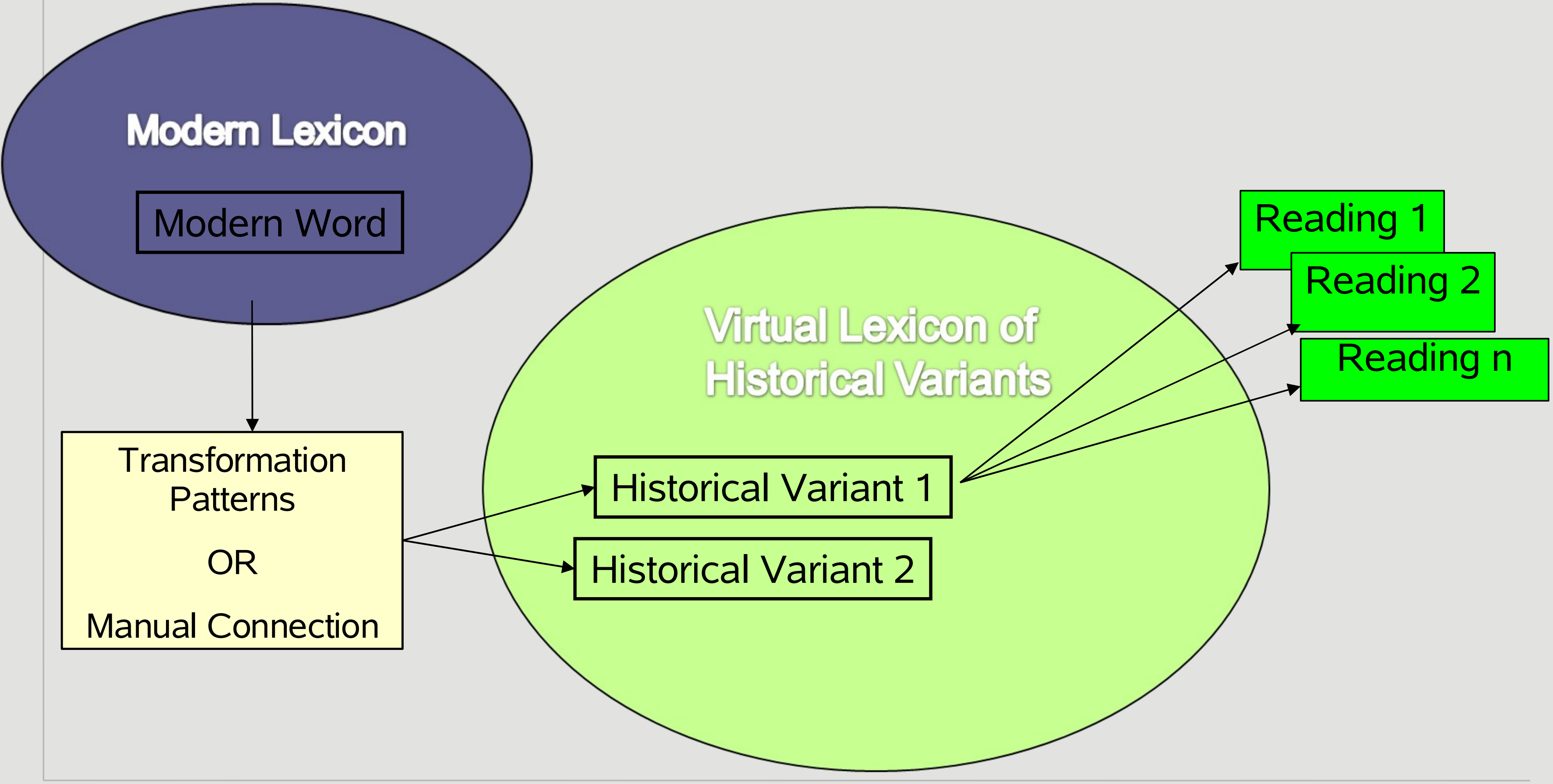
potential of additional non-pattern based historical lexicon



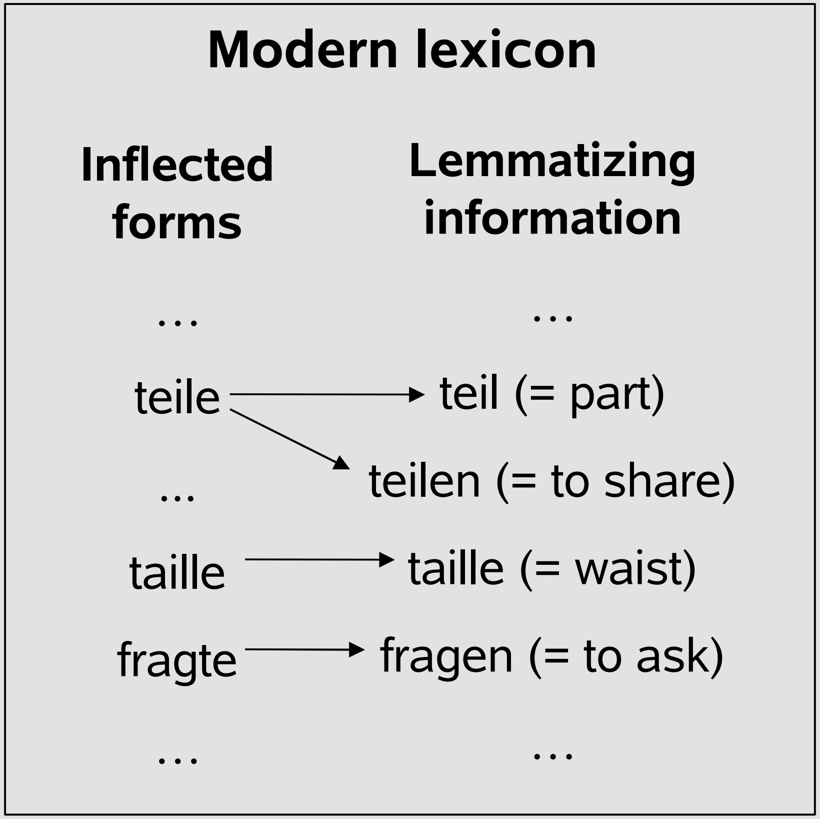
Tool for collaborative web based lexicon building

- Goal: lexicon that assigns modern lemmas to historical full forms
- Corpus management component
- Two kinds of historical words: pattern based and other
- Modern lemmas and POS information for historical full forms automatically suggested by system
- Collaborative handling of difficult cases
- Workflows: corpus frequencies or on single documents

Main Background Ressource



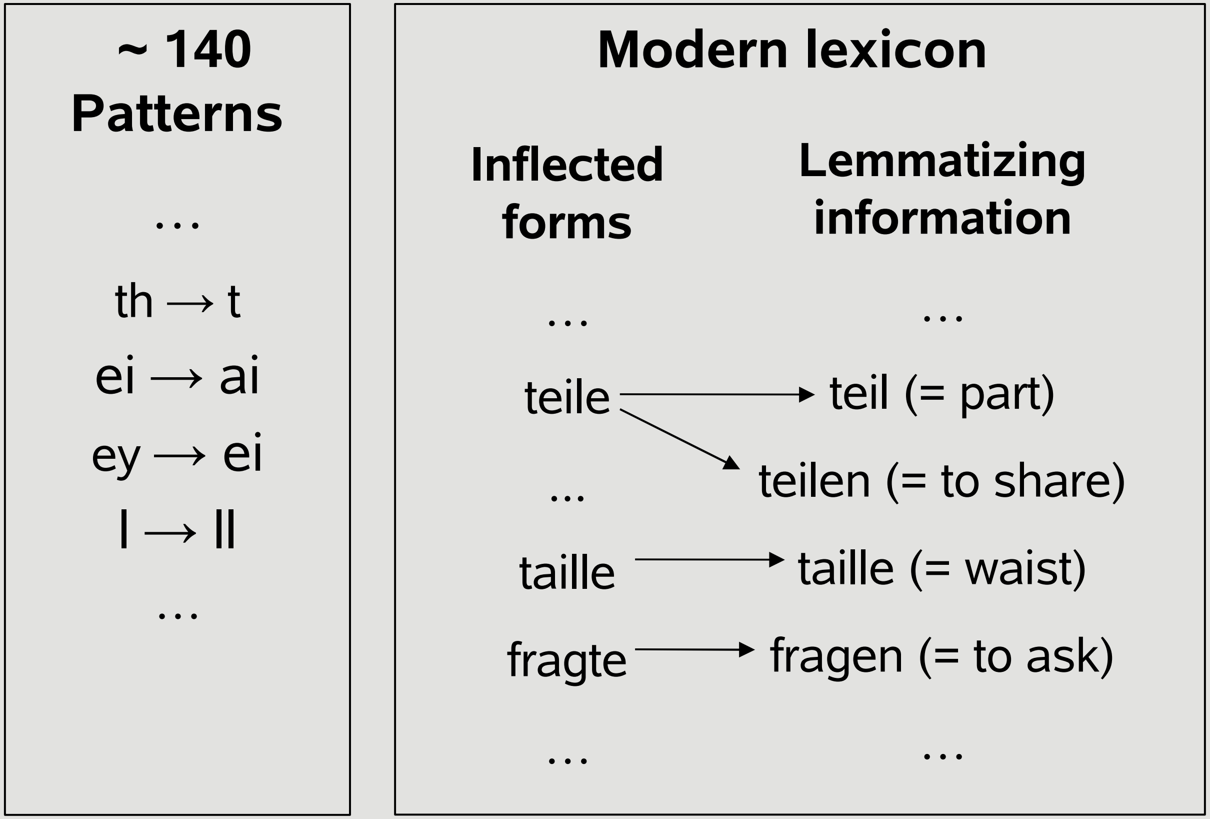
Approximate matching procedure



Modern lexicon

Inflected forms	**Lemmatizing information**
…	…
teile	teil (= part)
	teilen (= to share)
…	
taille	taille (= waist)
fragte	fragen (= to ask)
…	…

Approximate matching procedure



~ 140 Patterns

…

th → t

ei → ai

ey → ei

l → ll

…

Modern lexicon

Inflected forms

Lemmatizing information

…

…

teile → teil (= part)

teile → teilen (= to share)

…

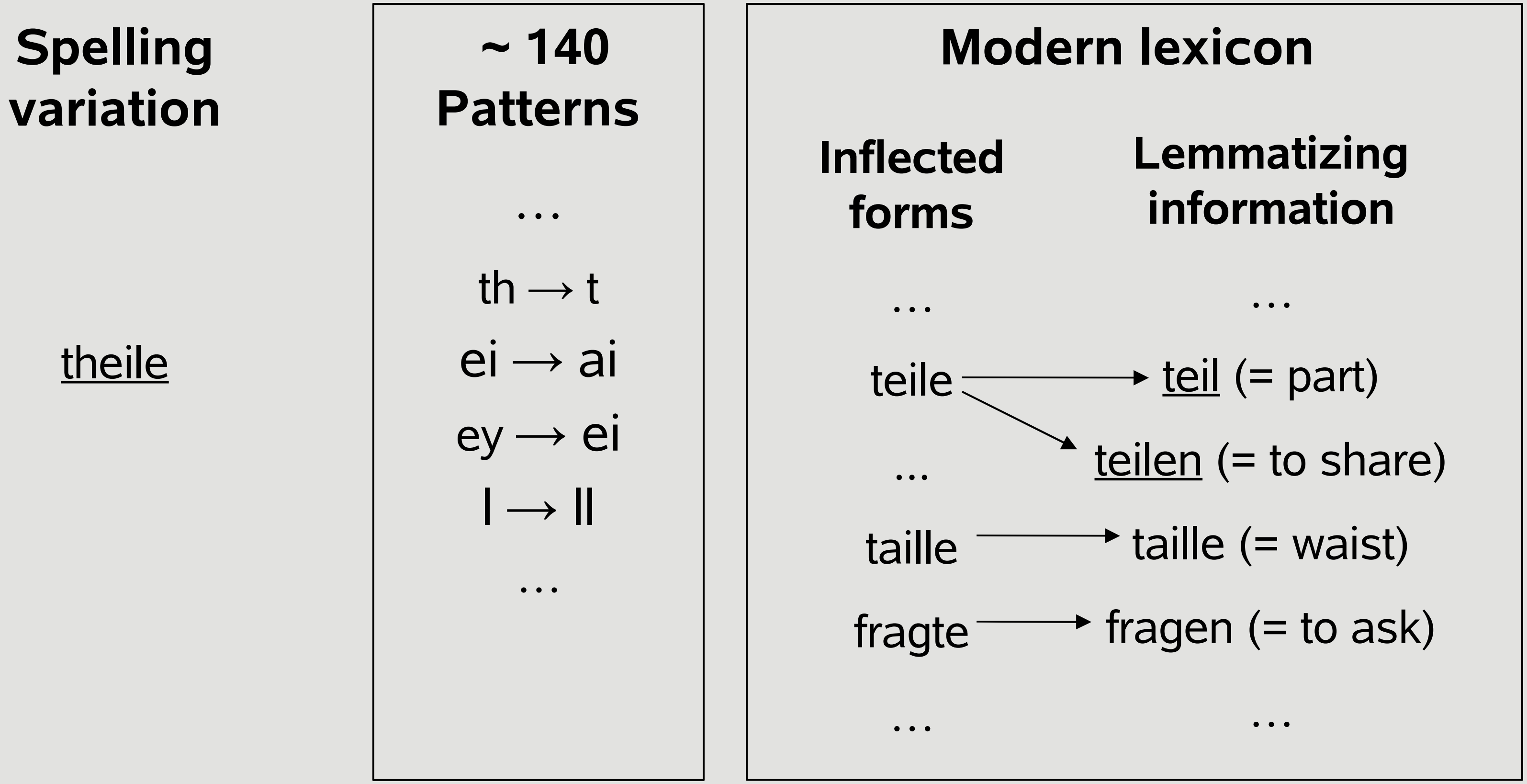
taille → taille (= waist)

fragte → fragen (= to ask)

…

…

Approximate matching procedure



Spelling variation

theile

~ 140 Patterns

…

th → t

ei → ai

ey → ei

l → ll

…

Modern lexicon

Inflected forms

Lemmatizing information

…

teile → teil (= part)

teile → teilen (= to share)

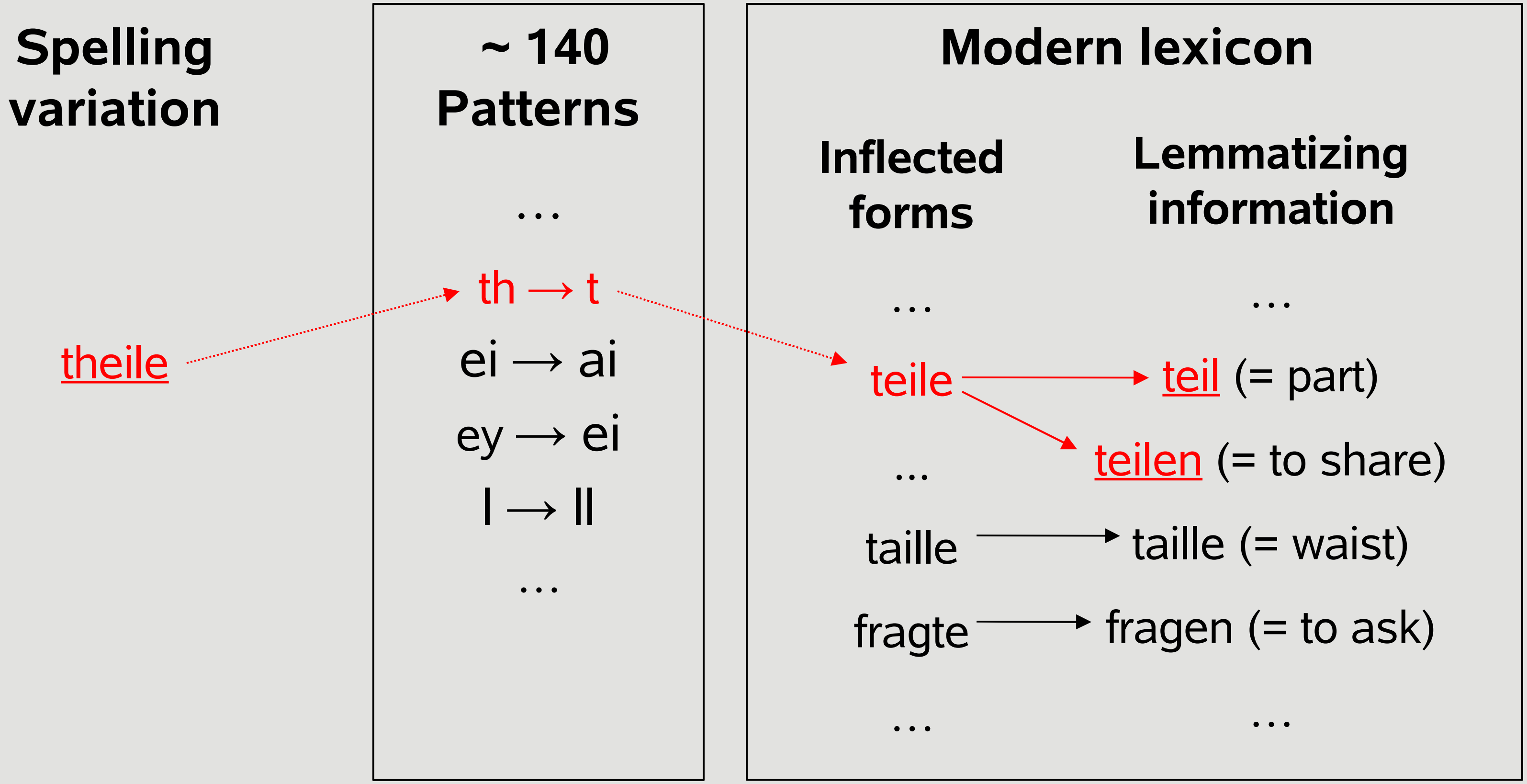
…

taille → taille (= waist)

fragte → fragen (= to ask)

…

Approximate matching procedure



Spelling variation

theile

~ 140 Patterns

…

th → t

ei → ai

ey → ei

l → ll

…

Modern lexicon

Inflected forms

Lemmatizing information

…

teile → teil (= part)

teile → teilen (= to share)

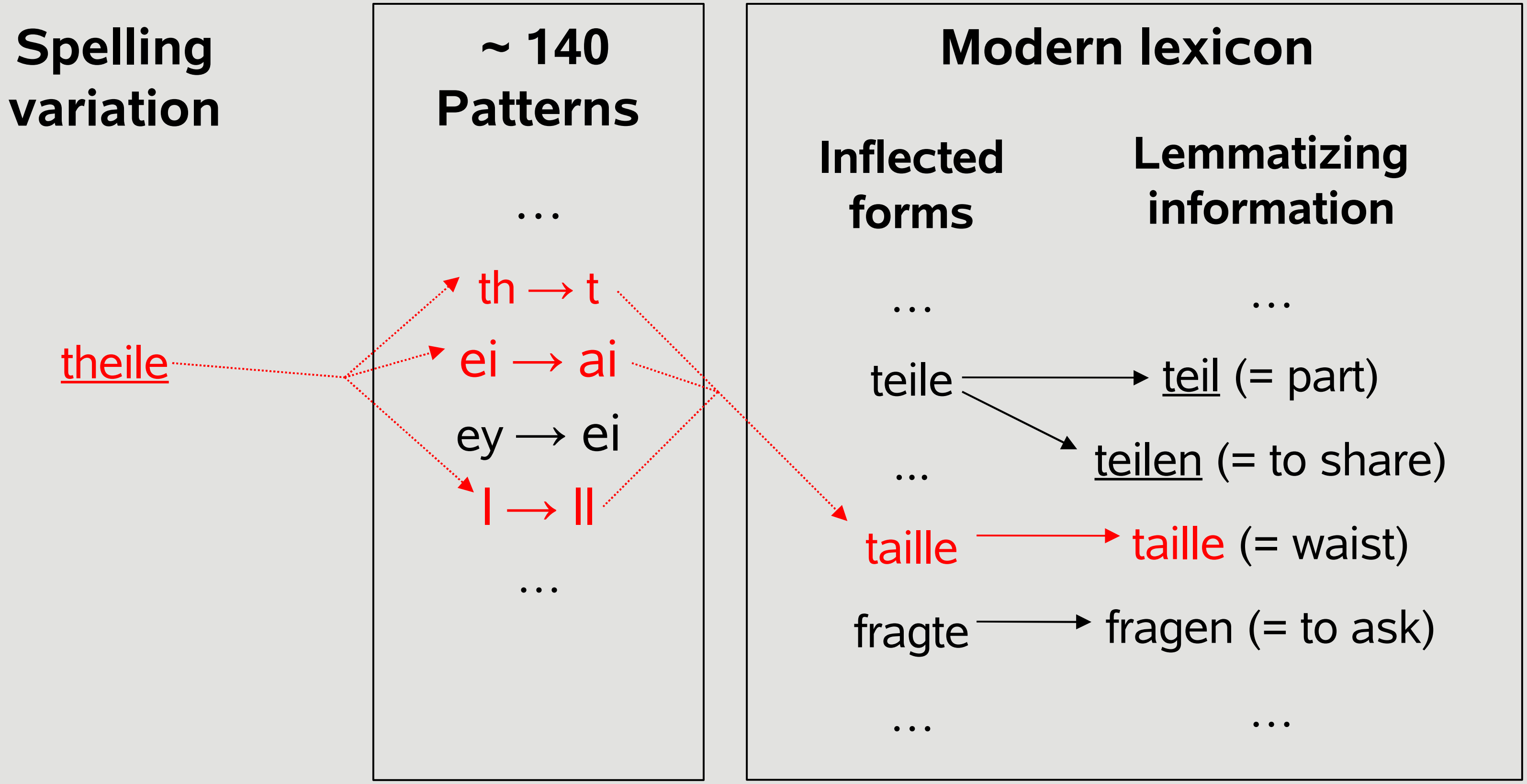
…

taille → taille (= waist)

fragte → fragen (= to ask)

…

Approximate matching procedure



Spelling variation

theile

~ 140 Patterns

…

th → t

ei → ai

ey → ei

l → ll

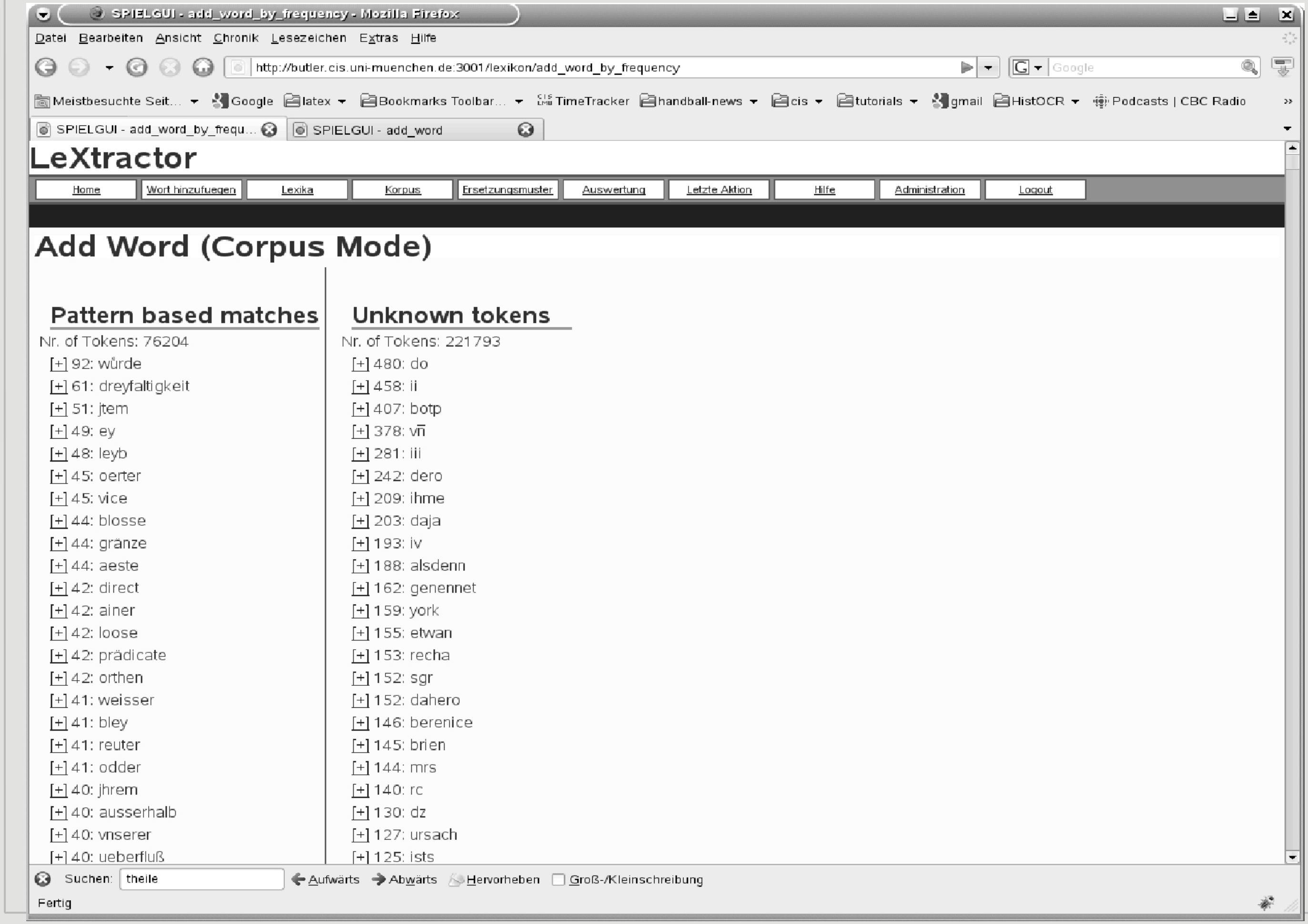
…

Modern lexicon

Inflected forms	**Lemmatizing information**
…	…
teile	teil (= part)
	teilen (= to share)
…	
taille	taille (= waist)
fragte	fragen (= to ask)
…	…

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Lexicon Tool Corpus Mode: User selects a word


SPIELGUI - add_word_by_frequency - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://butler.cis.uni-muenchen.de:3001/lexikon/add_word_by_frequency
Google
Meistbesuchte Seit... Google latex Bookmarks Toolbar... TimeTracker handball-news cis tutorials gmail HistOCR Podcasts | CBC Radio
SPIELGUI - add_word_by_frequ... SPIELGUI - add_word
Home Wort hinzufuegen Lexika Korpus Ersetzungsmuster Auswertung Letzte Aktion Hilfe Administration Logout
Suchen: theile Aufwärts Abwärts Hervorheben □ Groß-/Kleinschreibung
Fertig

LeXtractor

Add Word (Corpus Mode)

Pattern based matches

Nr. of Tokens: 76204

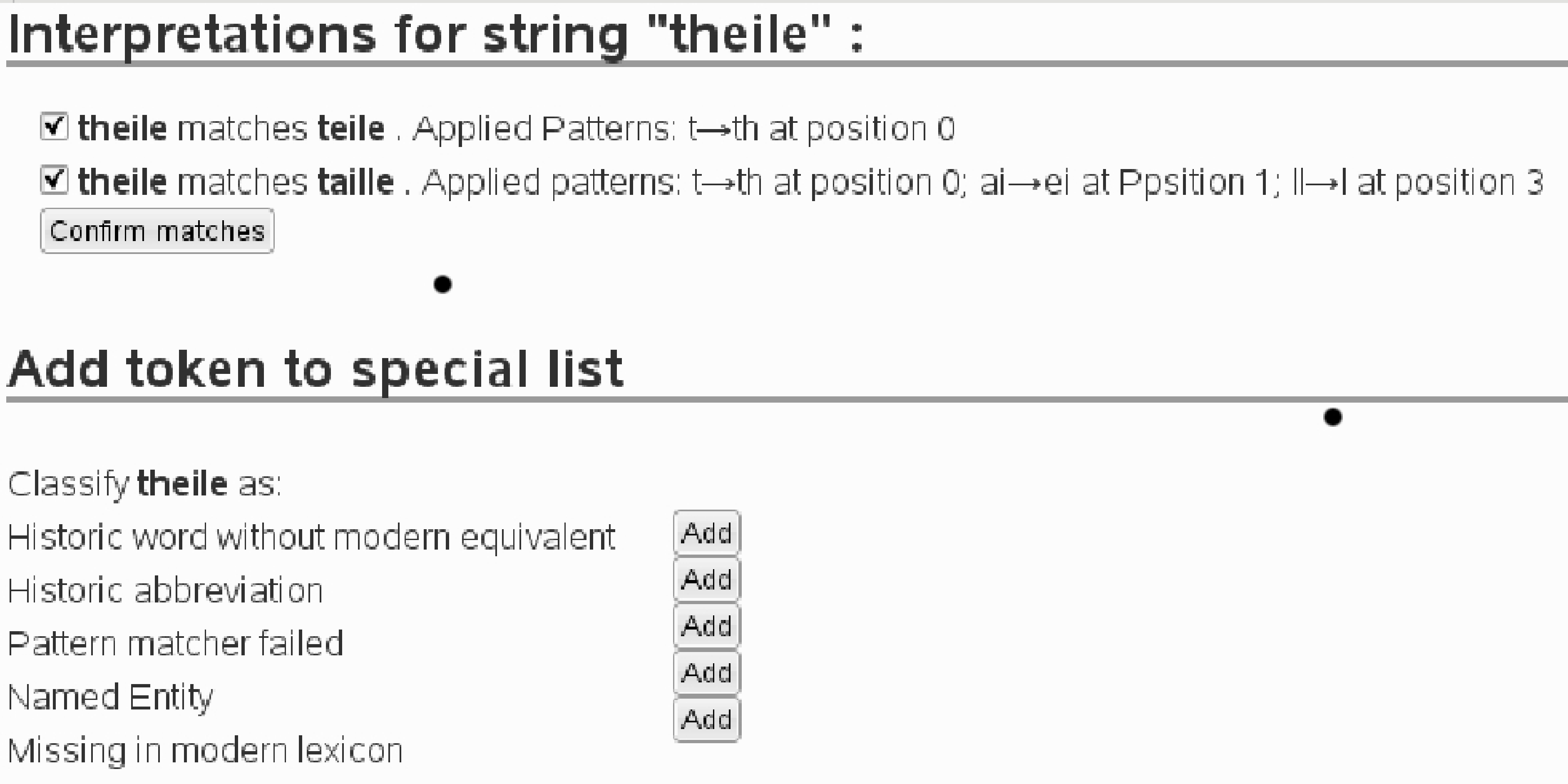
- [+] 92: wůrde
- [+] 61: dreyfaltigkeit
- [+] 51: jtem
- [+] 49: ey
- [+] 48: leyb
- [+] 45: oerter
- [+] 45: vice
- [+] 44: blosse
- [+] 44: granze
- [+] 44: aeste
- [+] 42: direct
- [+] 42: ainer
- [+] 42: loose
- [+] 42: prädicate
- [+] 42: orthen
- [+] 41: weisser
- [+] 41: bley
- [+] 41: reuter
- [+] 41: odder
- [+] 40: jhrem
- [+] 40: ausserhalb
- [+] 40: vnserer
- [+] 40: ueberfluß

Unknown tokens

Nr. of Tokens: 221793

- [+] 480: do
- [+] 458: ii
- [+] 407: botp
- [+] 378: vñ
- [+] 281: iii
- [+] 242: dero
- [+] 209: ihme
- [+] 203: daja
- [+] 193: iv
- [+] 188: alsdenn
- [+] 162: genennet
- [+] 159: york
- [+] 155: etwan
- [+] 153: recha
- [+] 152: sgr
- [+] 152: dahero
- [+] 146: berenice
- [+] 145: brien
- [+] 144: mrs
- [+] 140: rc
- [+] 130: dz
- [+] 127: ursach
- [+] 125: ists

Pattern Variants created automatically: User selects



Interpretations for string "theile" :

☑ **theile** matches **teile** . Applied Patterns: t→th at position 0

☑ **theile** matches **taille** . Applied patterns: t→th at position 0; ai→ei at Ppsition 1; ll→l at position 3

Add token to special list

Classify **theile** as:

Historic word without modern equivalent

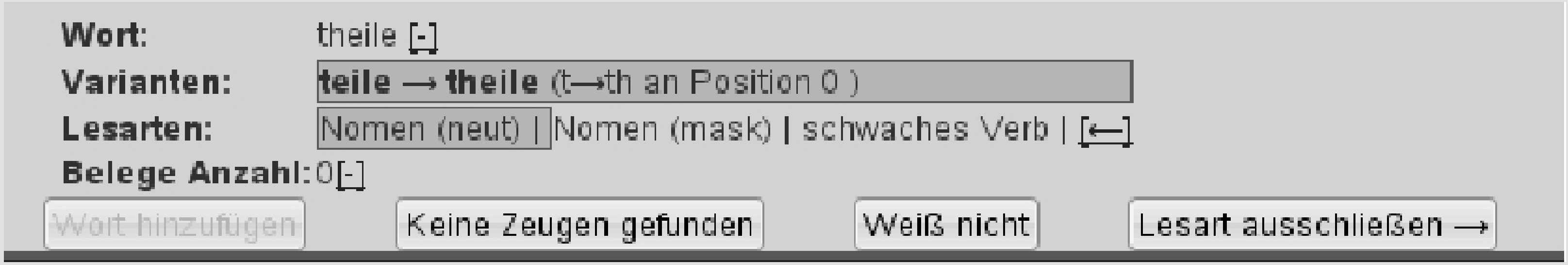
Historic abbreviation

Pattern matcher failed

Named Entity

Missing in modern lexicon

Lemma Readings created automatically: User confirms



Wort: theile [-]

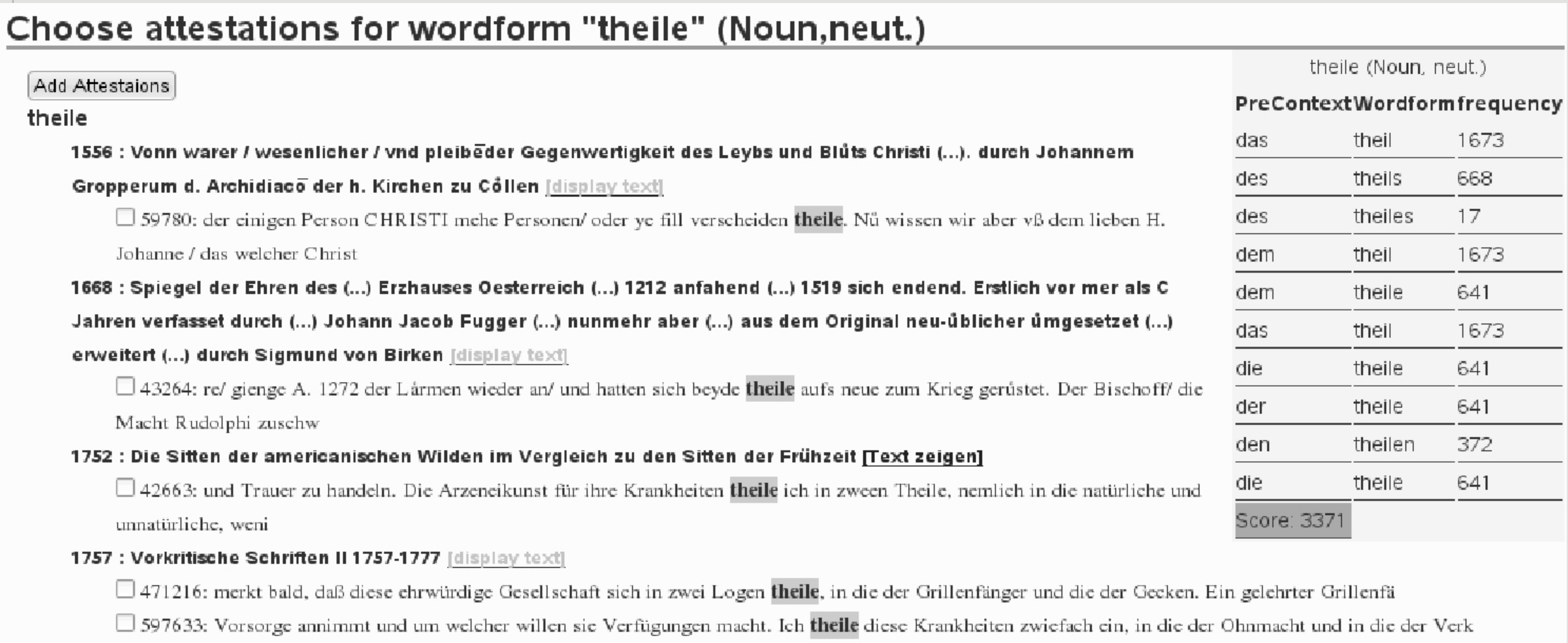
Varianten: **teile → theile** (t→th an Position 0)

Lesarten: Nomen (neut) | Nomen (mask) | schwaches Verb | [⟵]

Belege Anzahl: 0[-]

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Lexicon Tool Lemma Attestations: User selects



Choose attestations for wordform "theile" (Noun,neut.)

theile

1556 : Vonn warer / wesenlicher / vnd pleibēder Gegenwertigkeit des Leybs und Blůts Christi (...). durch Johannem Gropperum d. Archidiacō der h. Kirchen zu Cőllen [display text]

□ 59780: der einigen Person CHRISTI mehe Personen/ oder ye fill verscheiden **theile**. Nũ wissen wir aber vß dem lieben H. Johanne / das welcher Christ

1668 : Spiegel der Ehren des (...) Erzhauses Oesterreich (...) 1212 anfahend (...) 1519 sich endend. Erstlich vor mer als C Jahren verfasset durch (...) Johann Jacob Fugger (...) nunmehr aber (...) aus dem Original neu-ůblicher ůmgesetzet (...) erweitert (...) durch Sigmund von Birken [display text]

□ 43264: re/ gienge A. 1272 der Lármen wieder an/ und hatten sich beyde **theile** aufs neue zum Krieg gerústet. Der Bischoff/ die Macht Rudolphi zuschw

1752 : Die Sitten der americanischen Wilden im Vergleich zu den Sitten der Frühzeit [Text zeigen]

□ 42663: und Trauer zu handeln. Die Arzeneikunst für ihre Krankheiten **theile** ich in zween Theile, nemlich in die natürliche und unnatürliche, weni

1757 : Vorkritische Schriften II 1757-1777 [display text]

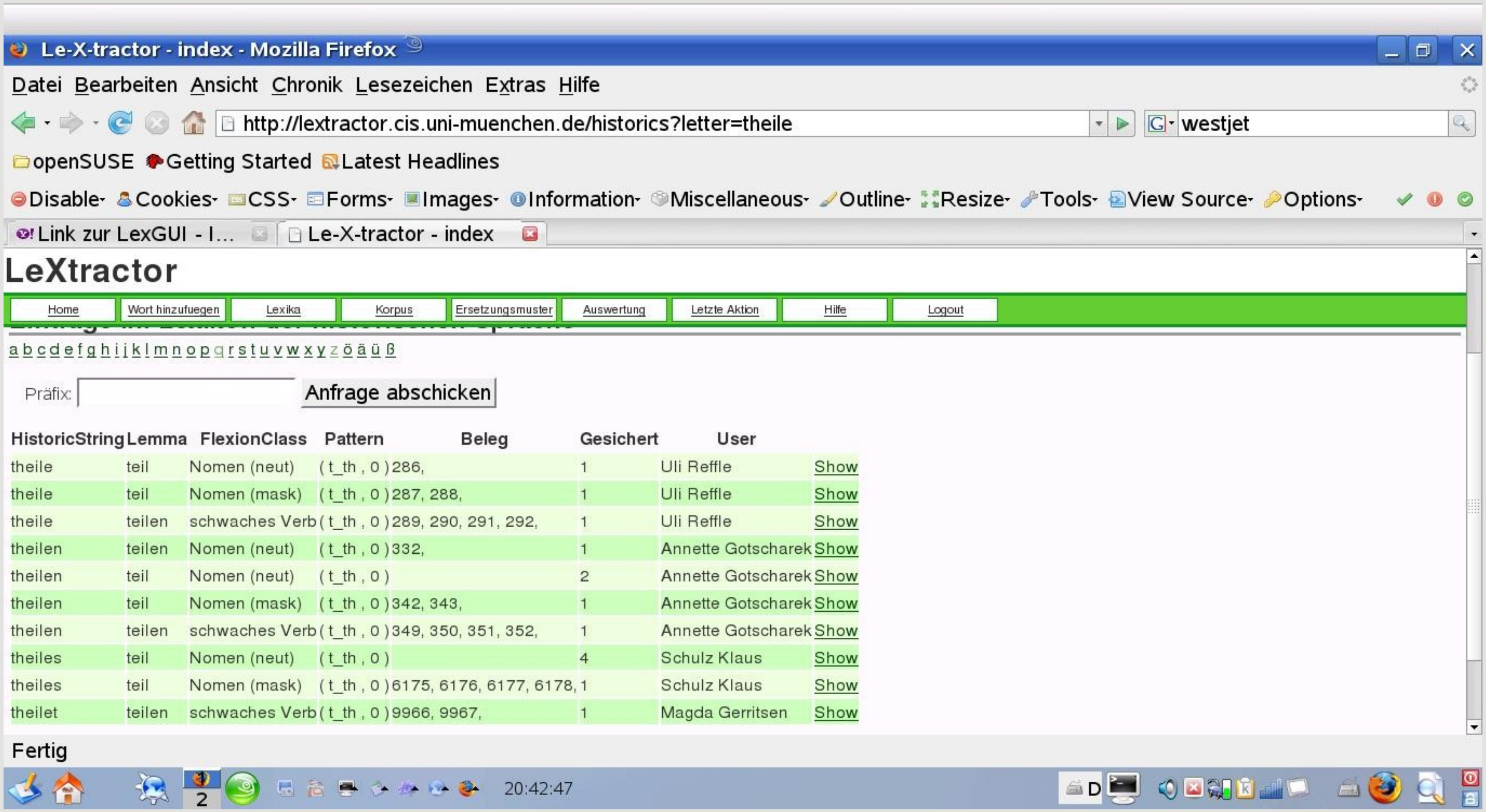
□ 471216: merkt bald, daß diese ehrwürdige Gesellschaft sich in zwei Logen **theile**, in die der Grillenfänger und die der Gecken. Ein gelehrter Grillenfä

□ 597633: Vorsorge annimmt und um welcher willen sie Verfügungen macht. Ich **theile** diese Krankheiten zwiefach ein, in die der Ohnmacht und in die der Verk

theile (Noun, neut.)

PreContext	Wordform	frequency
das	theil	1673
des	theils	668
des	theiles	17
dem	theil	1673
dem	theile	641
das	theil	1673
die	theile	641
der	theile	641
den	theilen	372
die	theile	641
Score: 3371		

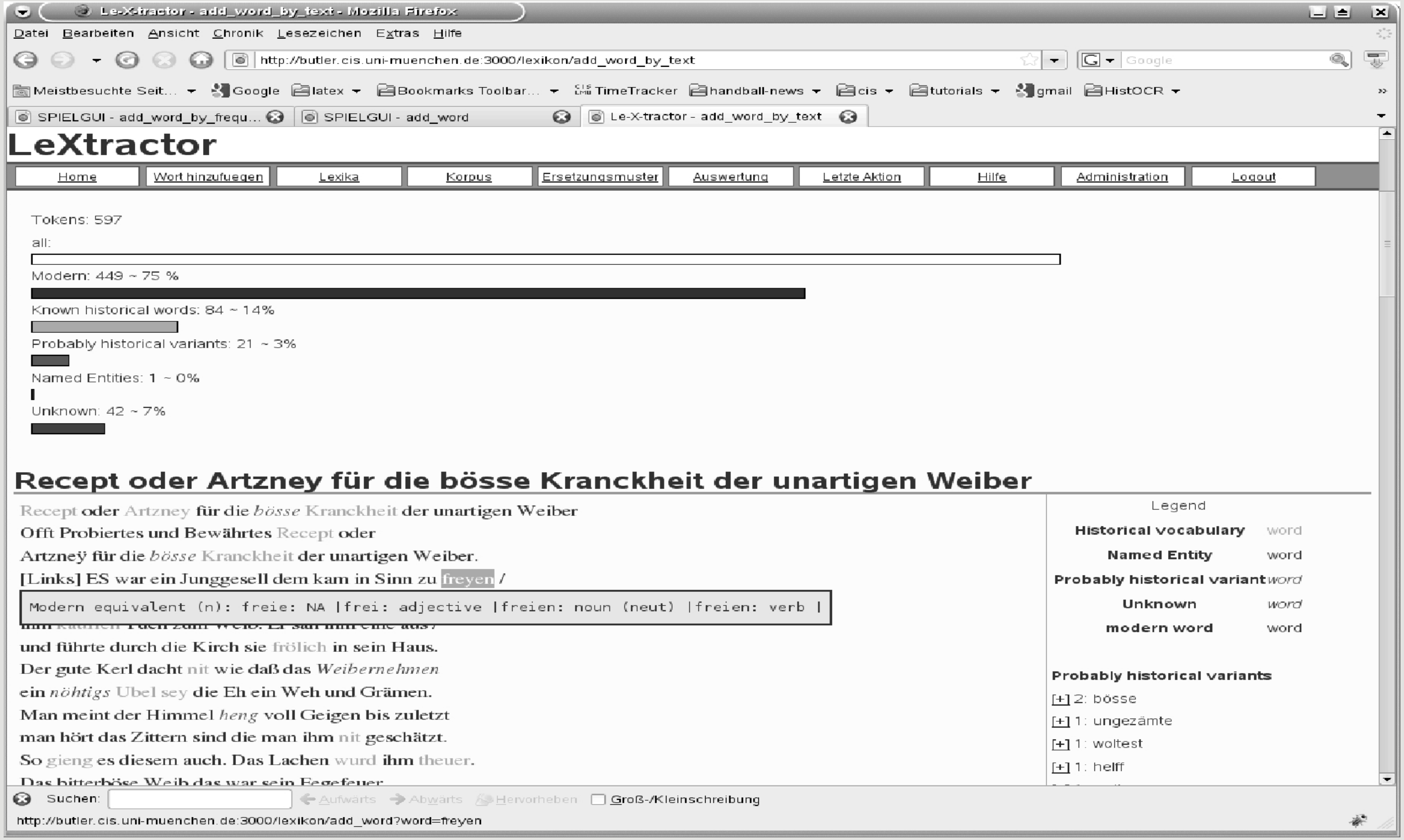
Resulting Lexicon Structure



LeXtractor

HistoricString	Lemma	FlexionClass	Pattern	Beleg	Gesichert	User	
theile	teil	Nomen (neut)	(t_th , 0)	286,	1	Uli Reffle	Show
theile	teil	Nomen (mask)	(t_th , 0)	287, 288,	1	Uli Reffle	Show
theile	teilen	schwaches Verb	(t_th , 0)	289, 290, 291, 292,	1	Uli Reffle	Show
theilen	teilen	Nomen (neut)	(t_th , 0)	332,	1	Annette Gotscharek	Show
theilen	teil	Nomen (neut)	(t_th , 0)		2	Annette Gotscharek	Show
theilen	teil	Nomen (mask)	(t_th , 0)	342, 343,	1	Annette Gotscharek	Show
theilen	teilen	schwaches Verb	(t_th , 0)	349, 350, 351, 352,	1	Annette Gotscharek	Show
theiles	teil	Nomen (neut)	(t_th , 0)		4	Schulz Klaus	Show
theiles	teil	Nomen (mask)	(t_th , 0)	6175, 6176, 6177, 6178,	1	Schulz Klaus	Show
theilet	teilen	schwaches Verb	(t_th , 0)	9966, 9967,	1	Magda Gerritsen	Show

Additional View: Lexicon Tool Document Mode


Le-X-tractor - add_word_by_text - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://butler.cis.uni-muenchen.de:3000/lexikon/add_word_by_text
Google
Meistbesuchte Seit... Google latex Bookmarks Toolbar... TimeTracker handball-news cis tutorials gmail HistOCR
SPIELGUI - add_word_by_frequ... SPIELGUI - add_word Le-X-tractor - add_word_by_text
Home Wort hinzufuegen Lexika Korpus Ersetzungsmuster Auswertung Letzte Aktion Hilfe Administration Logout
Suchen: Aufwärts Abwärts Hervorheben □ Groß-/Kleinschreibung
http://butler.cis.uni-muenchen.de:3000/lexikon/add_word?word=freyen

LeXtractor

Tokens: 597

all:

Modern: 449 ~ 75 %

Known historical words: 84 ~ 14%

Probably historical variants: 21 ~ 3%

Named Entities: 1 ~ 0%

Unknown: 42 ~ 7%

Recept oder Artzney für die bösse Kranckheit der unartigen Weiber

Recept oder Artzney für die *bösse* Kranckheit der unartigen Weiber
Offt Probiertes und Bewährtes Recept oder
Artzneÿ für die *bösse* Kranckheit der unartigen Weiber.
[Links] ES war ein Junggesell dem kam in Sinn zu freyen /

```
Modern equivalent (n): freie: NA |frei: adjective |freien: noun (neut) |freien: verb |
```

und führte durch die Kirch sie frölich in sein Haus.
Der gute Kerl dacht nit wie daß das *Weibernehmen*
ein *nöhtigs* Ubel sey die Eh ein Weh und Grämen.
Man meint der Himmel *heng* voll Geigen bis zuletzt
man hört das Zittern sind die man ihm nit geschätzt.
So gieng es diesem auch. Das Lachen wurd ihm theuer.
Das bitterböse Weib das war sein Fegefeuer

Legend

Historical vocabulary	word
Named Entity	word
Probably historical variant	*word*
Unknown	*word*
modern word	word

Probably historical variants

[+]	2: bösse
[+]	1: ungezämte
[+]	1: woltest
[+]	1: helff

Resources built

- **Validated lexicon**
  15,000 entries (“kernel lexicon for hist. German”)
- **Improved hypothetical lexicon**
  Set of observed transition patterns with time stamps and frequency information from automatic attestations enables a targeted hypothetical lexicon and quantitative investigations on language change.

Pattern based IR lexicon: precision and recall

Results on IR corpus, manual evaluation of 31,745 tokens

	16th century	17th century	18th century	19th century
Prec	71.84%	84.22%	94.76%	97.68%
Prec(-stop)	**62.09%**	**76.49%**	**91.09%**	**95.78%**
Rec	84.70%	86.60%	94.25%	94.95%
Rec(-stop)	**76.16%**	**78.69%**	**89.69%**	**90.26%**
FF	6.07%	4.80%	1.57%	0.29%
FF(-stop)	**8.47%**	**6.92%**	**1.52%**	**0.55%**

Project Status

- Lexicon work continues
- Benefits for IR and OCR already proven
- New corpus especially 16th century integrated by end of 2009
- CL: study on historical transformation patterns
- CL: study on German compound words
- Work on English starts 2010

Corpus Extension

Iterative Process with Bavarian State Library to Create More Resources for Early High German

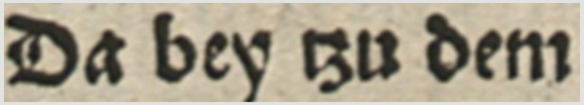
(1) A random selection of 200 pages from 100 sources

(2) Selection of usable sources

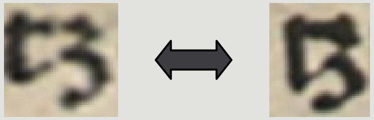
(3) Specification of keying by BSB/CIS for 70 complete books usable for both presentation and linguistic resources building: 500,000 tokens

(4) Contract with service providers

Next Step Corpus Extension: difficulties



Da bey tzu dem



From the service provider:
„Our question is whether there exists a difference between tz that looks normal and the form that looks more similar to ß.“