

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

Historical Lexicon Building & How It Improves Access to Text

Katrien Depuydt, INL IMPACT Team

LMU IMPACT Team

Can we handle 'de wereld'?

A. My daarentegen verſchaft het een onbedenkelyk
genoegen, als ik de voortgang van deeze zo wel als
alle andere Weetenſchappen door de verſcheiden eu-
wen der **wereld** naargaa, en zie, hoe veel juifter onze
begrippen, hoe veel bondiger onze beweegredenen
in de Zedekunde zyn, en welk eene heugelyke ver-
andering inzonderheid de Chriſtelyke Openbaaring,
ten deezen opzigte, heeft uitgewerkt.

OCR:

Abby Finereader SDK 9.0 with built in standard Dutch dictionary

werreid

Abby Finereader SDK 9.0 with lexicon of historical Dutch

werreld

RETRIEVAL:

Key in 'wereld' and find 'werreld'.



Main theme for impact EE2/3: overcome historical language barrier

- providing lexica for better OCR (→ Klaus yesterday)
- Try to equal/approximate results of basic IR (word lookup) on modern text for historical text

Language barrier (more than spelling variation)

- Orthography (wereld/werreld)
- Inflection (wereld/werelden)
- Other problems like words written together or split (swerelds)

Types of variation (orthographical and other)

I
uytterlijcste uyerlijkste d'uyterlijke uiterlyke uyerlijcke uiterlijke uyerlijck uiterlyken
uiterlijkste uiterlicke wterlicke wterlijcke ulterlijk uiterlyk uiterlijk uyerlick wterlicken
d'uyterlijcke uiterlijken uiterlijks wterlijck uyterlicke uitterlijke ujerlijke uyterlijk uyerlycke
uyterlicken uijterlicke d'uiterlijcke wtterlijcke wterlyke wtterlijk uuterlick uuterlic uyerlijke
uyterlijcken uyerlicke d'uiterlyke wterlijke vuyterlijcke uuterlycke uuterlicke wterlijken
uyterlijcksten uuyterlicke uuyterlick uuyterlycke uyterlijcke uyterlycke uyterlick vuyterlicke
uiterlijker uyerlyck uterliek wterlijcken uiterlijkst uitterlijk uyterlijcken uyerlyk wterlick
uutterlijck uuyterlicken uyttelijck uijterlijk uyterlijck uuterlijck uiterlick uitterlyk uuyterlic
uuyterlyck uuyterlijck uiterlijck uyterlyck uterlyc wterlijk

(most of these can be dealt with by means of patterns)

II
werelt weerelt wereld weerelds wereldt werelden weereld werrelts waerelds **weerlyt**
wereldts vveerelts waereld weerelden waerelden weerlt werlt werelds **sweerels zwerlys**
swarels swerelts werelts **swerrels** weirelts tsweerelds **werret** vverelt werlts werrelt
worreld werlden **wareld weirelt weireld** waerelt werreld wereld vvereld weerelts werlde
tswerels werreldts weereldt wereldje waereldje **weurt wald weëled**

(some of these can be dealt with by patterns and/or fuzzy matching,
others can only be handled by explicit listing)

Summary: why build historical lexica

- Words no longer in use
- “Unpredictable” (other than purely orthographical) variants
- Morphological variants

EE 2/3 Approach

- Develop historical lexica combining scholarly precision with broad coverage for use in mass digitization (for both text recognition (TR5) and enhanced retrieval (EE2))
- Deliver a set of tools for the efficient production of such lexica + guidelines (“lexicon cookbook”)
- The lexica:
 - able to specialize to periods / subject matter by keeping attestation information (*wereld is not used for OCR of modern texts*)
 - suitable for retrieval in applications for the general public by providing ‘modern’ query terms to search for historical variants (*use “wereld” to search for all variants*)
- The tools:
 - part of a workflow involving both automatic and manual processing
 - Language-independent (generic) whenever possible
 - Fit to quickly process large quantities of data

Lexicon structure (1)

The core objects in the lexicon structure developed for IMPACT:
word forms, lemmata and documents.

To enable the extraction of *relevant* lexicon data for OCR, it is not sufficient to convert existing lexica and dictionaries into a large word list.

We also need to

- Keep track of the sources from which we took the words.
- List the actually encountered words in the language and record occurrences in actual texts, with frequency information (attestation).
- Record in what kind of texts these words occur (document properties).

Lexicon structure (2)

We need mechanisms to extend the lexicon and to be able to assess the plausibility of “*hypothetical*” words without previous attestations, i.e. words we have not seen before. Supporting data for these mechanisms have to be present in the database, such as:

- Unknown inflected forms of lemmata which already are in the database can be dealt with by means of the automatic expansion from the lemma to the full paradigm of word forms (paradigmatic expansion).
- New spellings of known words can be dealt with by developing a good model of the spelling conventions of the period at hand. The database structure provides for the storage of orthographic variant patterns.
- Previously unseen compounds can be dealt with by means of a good model of word formation.

Lexicon structure (3)

In order to effectuate word searches without having to worry about inflection and variation of wordforms, enrichment will use “modern lemmata” as variation-independent retrieval keys for the full spectrum of inflectional and orthographical variation.

The database structure is divided into a few main blocks:

- Information attached to word forms, either **unlabeled** (i.e. not yet lemmatized or labeled with Part of Speech) or **labeled** (i.e. with lemma and possibly PoS)
- Information attached to lemmata
- Information about documents, parts of documents, document collections
- Auxiliary information needed for expansion and for plausibility-of-unseen-words prediction

Lexicon structure (4)

To each labeled or unlabeled word form, we link *attestation* objects, verified occurrences of the words in documents.

The attestation objects contain the following information:

- Link to the relevant word form and a location in a document
- Verification (yes/no): Is the occurrence of a labeled word form checked manually by an expert?
- Frequency in a document or document collection.

Attestation information enables us to

- Always recheck the sources for a certain analysis of a word form
- Derive the the relevant information about the domain of applicability of word forms is from the properties of the documents they occur in

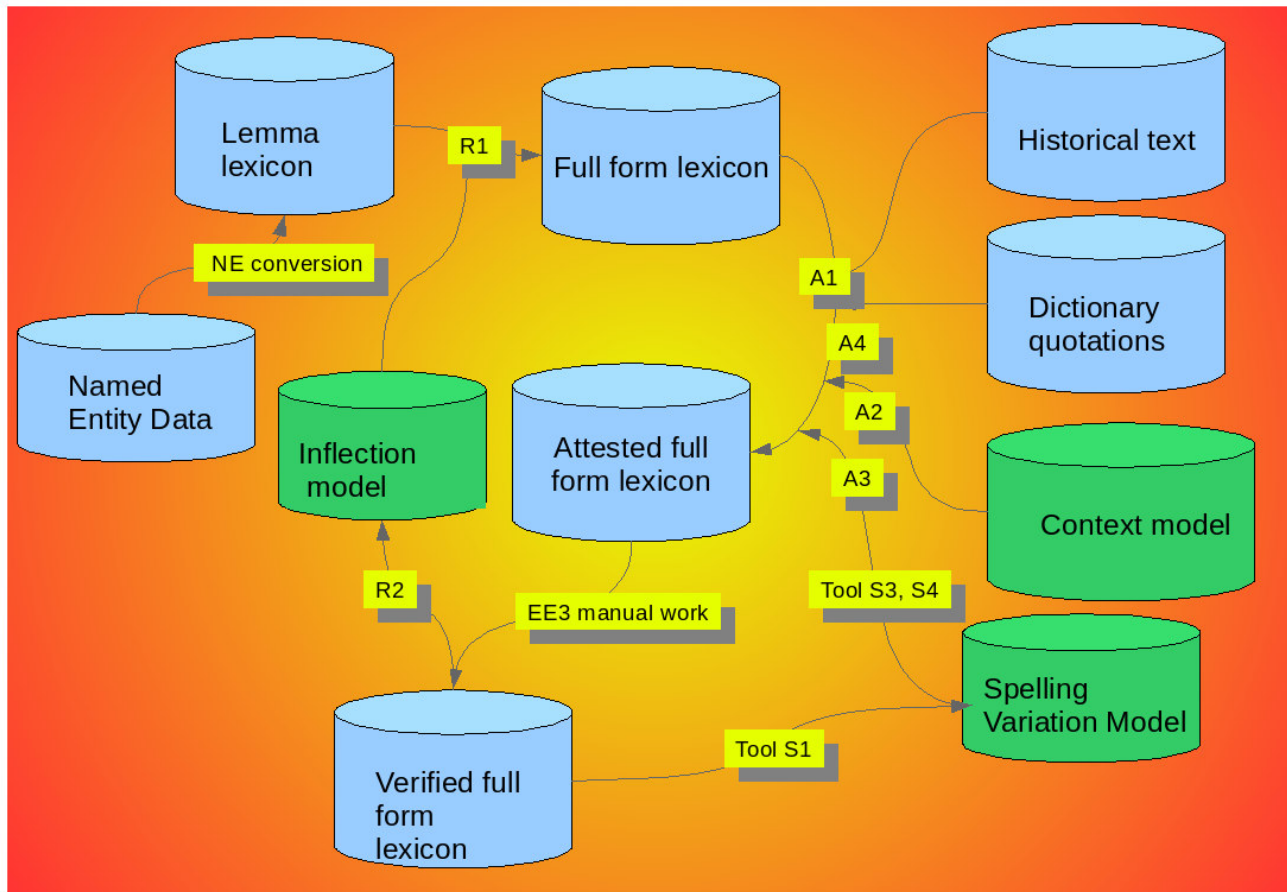
Lexicon building

Aim:

- Build a lexicon of variants
- Build models of variation (inflection, orthography)

Procedure:

- Cycle: model helps to construct lexicon, and vice versa (induction of rules/patterns)
- Combination of manual work and computational linguistics
- Lexicon building toolkit to support development



Cf. Computational Tools and Lexica to Improve Access to Text, Jesse de Does, Katrien Depuydt, on the IMPACT website from june.

Spelling variation tools (pattern-based)

- Supervised Rule (pattern) induction from pairs (“modern” word, historical word), yielding patterns like *aa/ae*, *s/z*, Pattern weights are computed from counts in example material
- Language-independent approach

Other approaches under investigation:

- Use of aligned data (parallel historical text and modern version)
- Semi-supervised pattern weighting for text profiling purposes

Lexicon building: IMPACT Attestation Tool

Lexicon building at work: Verifying attestations in historical dictionaries

Task

Find the variants of a headword as they occur in the quotations

headword

work

- We are working on what works.
- Depart from me, ye that worke iniquity.
- She worcketh knittinge of stockings.

Quotations

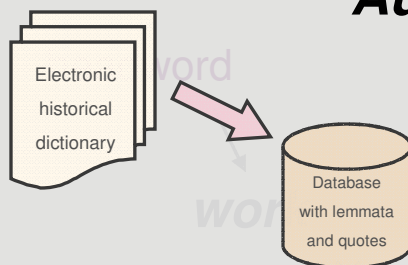
variants

IMPACT Attestation Tool

Task

Find the variants of a headword as they occur in the quotations

Automatically (preprocessing)



- match literally
e.g: work → work, Work
- match using existing lexica and lists
e.g: work → works, worked, wrought

- approximate matching
e.g: work → worke

By hand (using the tool)



- correct automatic mismatches
e.g: works → words, worms
- find missed matches
e.g: work → worketh, wrought

Quotations

IMPACT Attestation Tool

Tool

Up-to-date overview of what is done and needs to be done

Done by this user so far

Lemma headword

Quotations

Sorted by uncertainty

Some figures

- Used historical dictionary: *Woordenboek der Nederlandsche Taal*
- Expected lemmata: 220211, quotations: 1524366
- Progress: 155012 lemmata, 1200140 quotes, about 400.000 distinct word forms [6 April 2009]
- Tempo: 1725 quotes/hour 231 lemmata/hour



Lexicon coverage (1)

Example ground truth text: “de Denker I” (journal volume, 1763)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	56%	81%
WNTQuotationLexicon	74%	94%
M+Q	79%	95%

(both M and Q have lemma information)

Lexicon coverage (2)

OCR'd text: Dutch Parliamentary Papers: 1891-1892


	Type coverage	Token coverage
Modern lexicon (e-Lex)	46%	91%
WNTQuotationLexicon	44%	90%
M+Q	54%	94%
Corpus-based word list (no lemma information)	63%	97,6%

Aside:

OCR performance with modern lexicon on sample SG page: 90%

With corpus-based word list: 94.6%

Corpus-based lexicon building



IMPACT - Lexicon Tool

http://localhost/~LexiconTool/lexiconTool.php

Sort by: lemma frequency

User tom, working on corpus "Vondel"

<< Start page

<input checked="" type="checkbox"/> aenghewesen,	1	aanwijzen, V
<input checked="" type="checkbox"/> aenmerckende	1	aanmerken, V
<input checked="" type="checkbox"/> aerdighe	1	aardig, BIJW NW
<input checked="" type="checkbox"/> aert	1	aard, N
<input type="checkbox"/> aghemaeyde	1	
<input checked="" type="checkbox"/> al	5	al, TELW al, ONBEP VNW al, BIJW
<input checked="" type="checkbox"/> alle	4	alle, BIJW
<input type="checkbox"/> als	3	
<input type="checkbox"/> alzoo	2	
<input type="checkbox"/> andere	2	
<input type="checkbox"/> beelden	1	

ghden op te klimmen, ende om hoogh te stijghen in **al** het ghene wat loflijck ende eerlijck by hun mochte
 k by hun mochte ghenaemt worden, als zijnde eenen **al** te steylen bergh; zoo hebben sy in alle manieren g
 heschiedenissen wederom te ververschen, ende voor **al** de Werelt op't Toneel te stellen: om alzoo door ze
 ifs plumpe, rouwe- ende ongheleerde menschen, die **al** hoorende doof, ende al ziende blindt waren, zonder
 ongheleerde menschen, die al hoorende doof, ende **al** ziende blindt waren, zonder bril mochten hun feyle

al, TELW
 al, ONBEP VNW
 al, BIJW

Done

Lexicon deployment

For example:

- Match modern query term against historical words from text
- Match historical word form against modern lemma list (“lemmatization”)

Example: retrieval in a historical corpus of about 150 million tokens, using only the modern lexicon for **wereld** yields **23396** hits, using the WNT Quotation Lexicon we get **34339** hits.

In retrieval applications, the lexicon content will be combined with linguistic tools