

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 and Deployment in IMPACT

INL

## Types of variation (orthographical and other)

I  
uytterlicste uyerlijkste d'uyterlijke uiterlyke uyerlijcke uiterlijke uyerlijck uiterlyken  
uiterlijkste uiterlicke wterlicke wterlijcke ulterlijk uiterlyk uiterlijk uyerlick wterlicken  
d'uyterlijcke uiterlijken uiterlijks wterlijck uytterlicke uitterlijke ujterlijke uytterlijk uyerlycke  
uiterlicken uijterlicke d'uiterlijcke wtterlijcke wterlyke wtterlijk uuterlick uuterlic uyerlijke  
uiterlijcken uyerlicke d'uiterlyke wterlijke vuyterlijcke uuterlycke uuterlicke wterlijken  
uiterlijcksten uuyterlicke uuyterlick uuyterlycke uyterlijcke uyterlycke uyterlick vuytterlicke  
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 zwerlyts**  
**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 **weurlt wald weëled**

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

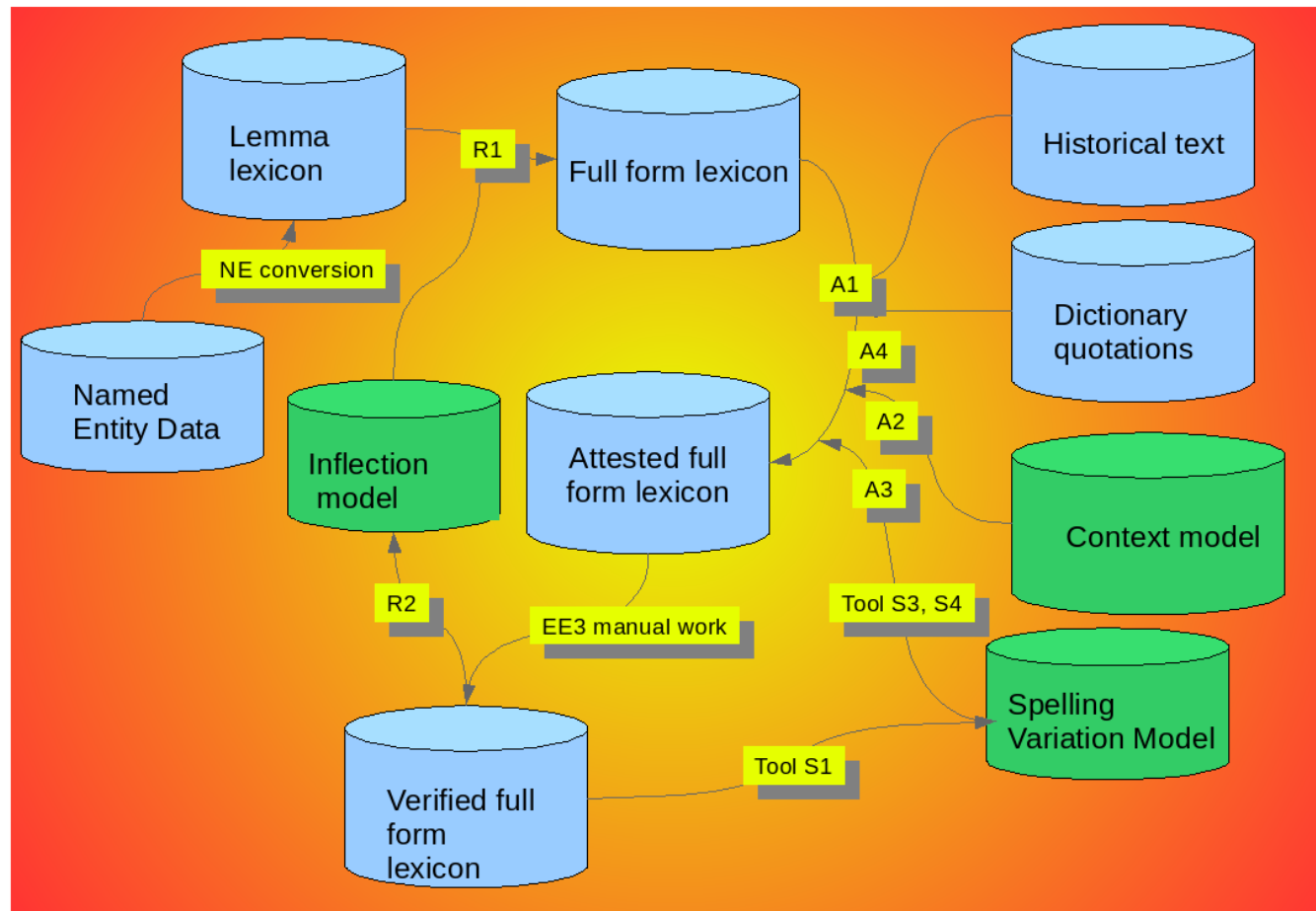


## What is needed for lexicon building

- Build models of linguistic variation (inflection, orthography)
- Collect variants

### *Approach*

- 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, containing both computational linguistic tools and tools supporting manual work



*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)

Language-independent approach:

- Supervised rule (pattern) induction from pairs (“modern” word, historical word), yielding patterns like *aa/ae*, *s/z*, ....
- Pattern weights are computed from example material

Additional approaches possible:

- Use of aligned data (parallel historical text and modern version)
- Unsupervised pattern weighting (=~ text profiling from TR5)

## Lemmatization and reverse lemmatization

We also need a lemmatization process for these situations

A typical lemmatizer assigns some standard form (infinitive, nominative, stem) to inflected forms. Usually based on patterns relating the inflected form to the standard form.

*But:*

Matching these patterns can be hard to combine with matching both spelling variation patterns and OCR errors (*bok/bokken/bokkeu*)

We adopt the solution of actually expanding the “hypothetical modern full form lexicon” containing the most plausible possible paradigmatic expansions of lemmata

This construction is carried out by means of a statistical reverse lemmatizer



## Attestation

- From hypothetical (non-witnessed) lexicon content to attested word forms in “real” text
- Automatic selection of candidate attestations
- Manual work: verification and correction

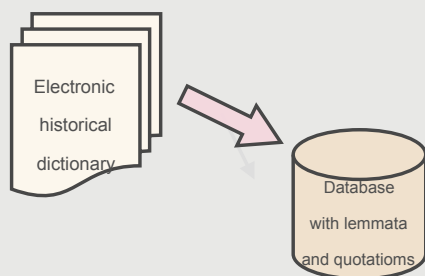
### Two approaches

- Dictionary based
- Corpus based

# IMPACT Dictionary 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, wrowght*

# IMPACT Attestation Tool

## Tool

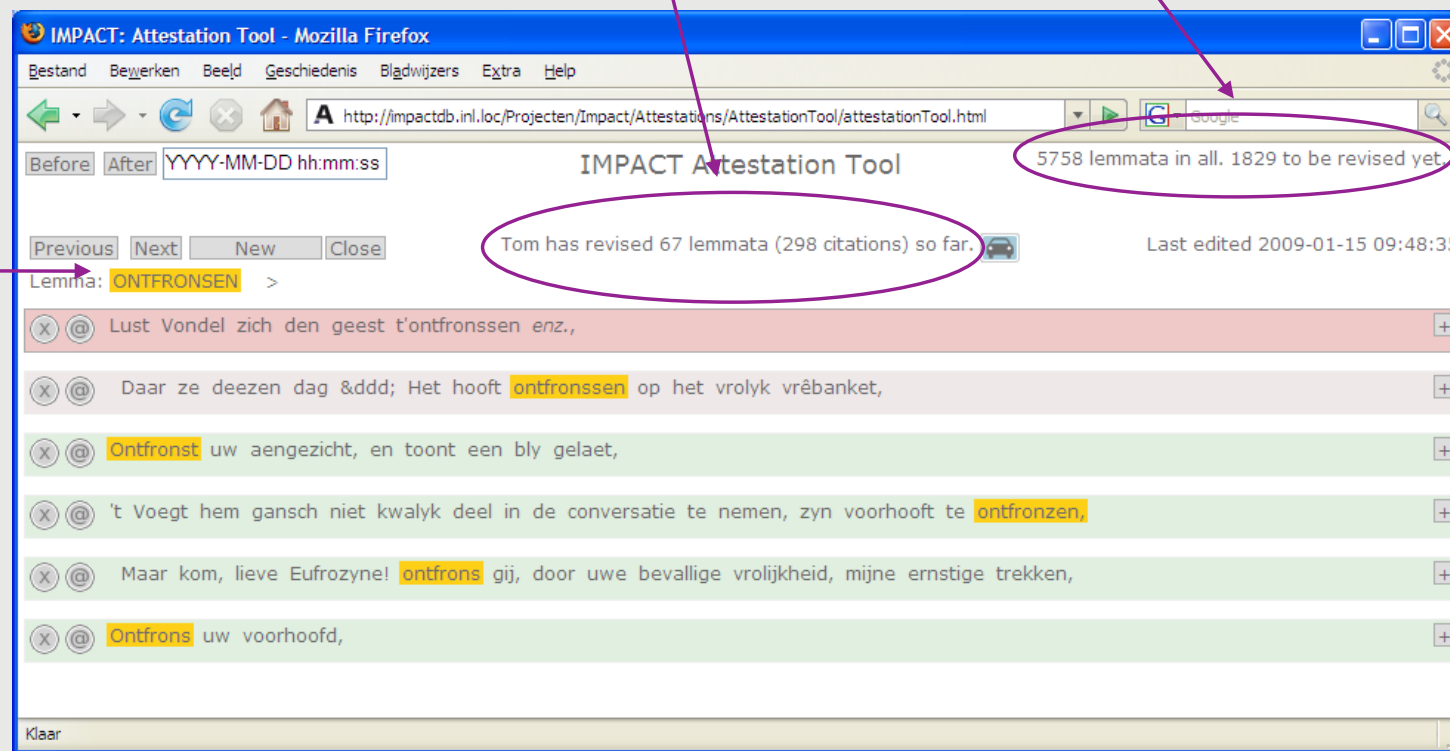
Up-to-date overview of what is done and needs to be done

Done by this user so far

Lemma headword

Quotations

Sorted by uncertainty



# IMPACT Lexicon Tool

## Task

Find and verify attestations in a historical corpus

### ***Automatically (preprocessing = apply lemmatizer)***

- match literally  
e.g: *work* → *work, Work*
- match using existing lexica and lists  
e.g: *work* → *works, worked, wrought*
- matching using spelling variation module  
e.g: *uiterlijk* → *uyterlick*

### ***By hand (using the tool)***

- assign correct lemma  
e.g: *was (N)* → *zijn (V)*
- group tokens belonging together  
e.g: *konings zoon* → *koningszoon*
- select attestations

# Corpus-based lexicon building: Impact Lexicon Tool

IMPACT - Lexicon Tool

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

Sort by:   User tom, working on corpus 'Vondel' << Start page

Lemma	Frequency	Occurrences
<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"/> afghemaeyde	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 hoog te stijghen in **al** het ghene wat loflijck ende eerlijck by hun mochte  
 k by hun mochte ghenaeamt 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  
 lfs 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



## Results on

- Efficiency of lexicon building
- Lexicon coverage
- Reverse lemmatization accuracy
- Lemmatization



## Efficiency of lexicon building

- Dictionary-based lexicon building using historical dictionary:  
*Woordenboek der Nederlandsche Taal*
- Lemmata: 220211, quotations: 1524366
- Tempo: 1725 quotations/hour; 231 lemmata/hour

## Measuring results for Dutch

We use the ground truth data developed in the project

- Evaluation of lexicon coverage
- Evaluation of modern lemma assignment
- Evaluation of OCR and lexicon usage in OCR (2010)

OCR performance with modern lexicon on DPO35 book:	89%
With corpus-based historical word list	94%

## Dutch ground truth data

<i>Type and genre</i>	<i># words</i>
Gold Standard Book	300k
Random Set Book	340k
Random Set Staten Generaal	2.5M
Gold Standard Staten Generaal	500k
Gold Standard Newspapers 1	3.4M
Gold Standard Newspapers 2	170k
Random Set Newspapers	3.2M
<i>total</i>	<i>13.1M</i>

## Lexicon coverage (1: ground truth books)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	46%	76%
EE3.3	56%	84%
1 + 2	63%	89%
Type frequency list historical corpus, top 200K (freq $\geq$ 19)	70%	93%
Type frequency list historical corpus, top 500K (freq $\geq$ 5)	78%	95%

## Lexicon coverage (2: ground truth newspapers)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	40%	83%
EE3.3	41%	84%
1 + 2	51%	89%
Type frequency list historical corpus, top 200K	52%	93%
Type frequency list historical corpus, top 500K	62%	95%

## Lexicon coverage (3: ground Parl. Papers 19<sup>th</sup> century)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	51%	89%
EE3.3	47%	88%
1 + 2	58%	93%
Type frequency historical corpus, top 200K	59%	96%
Type frequency historical corpus, top 500K	68%	97%

## Lexicon coverage (4: ground Parl. Papers 20<sup>th</sup> century)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	70%	93%
EE3.3	66%	93%
1 + 2	76%	96%
Type frequency historical corpus, top 200K	74%	97%
Type frequency historical corpus, top 500K	81%	98%

## Lexicon coverage (5: Genesis, 1637 bible)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	31%	61%
EE3.3	62%	83%
1 + 2	65%	89%
Type frequency historical corpus, top 200K	76%	97%
Type frequency historical corpus, top 500K	87%	98.6%

## Lexicon coverage (6: Hooft, historiën)

	Type coverage	Token coverage
Modern lexicon (e-Lex)	26%	67%
EE3.3	47%	88%
1 + 2	50%	90%
Type frequency historical corpus, top 200K	44%	93%
Type frequency historical corpus, top 500K	58%	96%

## Reverse lemmatization (lc 5.5)

- Reminder: build hypothetical (non-attested) word forms in a “quick and dirty” way to use in lemmatization and corpus-based lexicon building
- Using simple statistical algorithms and a simple approach to inflection
- *Results:*

	Accuracy
Small Dutch lexicon (JVKlex)	96.6%
French lexicon (Morphalou)	99.4%
Polish lexicon, verbs (Morfologik)	98.7%

## Lemmatization

- Combination of lookup, matching of spelling variation, reverse lemmatization
- As yet no good evaluation set for IMPACT (current work)
- Evaluation on “type” level

We will use other material here (1637 Genesis, 97144 tokens)

### *Approach*

- *Restrict to “ordinary words” (no names, numbers, clitic combinations)*
- *Ambiguous lemmatization (context is not used) (avg. 5 suggestions per word)*
- *Ranking based on frequency and pattern weights*

## Result

- *6265 distinct types. 5991 (95.7%) had at least one correct suggestion*
- Average rank of correct suggestions: **1.23**
  - 5222 types found in current EE3.3 (**83%**)
  - 65 additional types in modern lexicon
  - 49 types without any match
  - 969 types (**15%**) had some “approximate” matching using ~500 weighted patterns and returning at most 2 suggestions

## Real and hypothetical lexicon coverage (Hooft, historiën)

*Result (again restricting to ‘ordinary’ words)*

- **36332** distinct types. Avg rank of correct suggestions: 1.23
  - **20087** types found in current EE3.3 (**55%**)
  - **1061** additional types in modern lexicon
  - **2411** types without any match (**7%**)
  - **12773** types (**35%**) identified with “approximate” matching using ~500 weighted patterns and returning at most 2 suggestions (Probably about **75%** of the highest-ranking approximate matches are correct)



## General vocabulary vs. Named entities

- Tools for lexicon building described so far: applicable to general lexicon
- Tools for NE recognition, classification and variant matching
  - library requirement
  - distinguish general vocabulary from NE's
  - avoid unpleasant mixups like Abimelech ↔ apemelk!