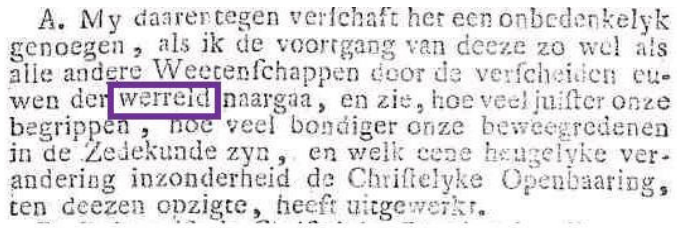
Computational Tools and Lexica to Improve Access to Text

Jesse de Does, Katrien Depuydt[1] (Instituut voor Nederlandse Lexicology)

1. Introduction: improving access to historical documents

IMPACT is a project funded by the European Commission. It aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitization of the European cultural heritage. For that IMPACT wants to improve the quality of OCR (Optical Character Recognition) for historical documents and to enhance their accessibility. There are many aspects involved in dealing with this problem which are addressed by IMPACT. Image processing, which tries to remedy typical problems like skewed, warped or otherwise noisy data; better segmentation procedures and adaptive OCR aim to overcome the irregularities of historical typography.

Full-text accessibility for historical text documents is also hindered by the historical language barrier. The language is not only a problem for text recognition, but also for users wanting to access the texts. How are they to find the necessary information, without having to take into account all possible spellings and inflections of words?
The following picture exemplifies the problem:[2]


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 werreid naargaa, en zie, hoe veel juiſter 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.

The variant form '*werreld*' poses a problem for text recognition (the recognition process will have to recognize this as a valid word; in fact Abbyy FineReader Engine 9.0 recognizes 'werreid') and retrieval: the user should be able to key in 'wereld' and find 'werreld' and other variants of this word.

Part of the solution for this type of problem is using a computational historical lexicon, supplemented by computational tools and linguistic models of variation. The lexicon lists historical variants (orthographical variants, inflected forms) and links them to a corresponding dictionary form in modern spelling ('*modern lemma*').

Work package EE2 in IMPACT will provide guidelines and general tools for lexical data development from historical source material and tools to deploy the lexicon in enrichment. Work package EE3 will deliver lexicon content. Work package TR-5 will provide suitable language models and algorithms to deal with historical language in text recognition[3].

This paper describes the lexicon building process. It does not cover the treatment of Named Entity data.

---

[1] With many thanks to Adrienne Bruyn, Bob Boelhouwer, Michel Boekestein, Tom Kenter, Mika Poss and Tilly Ruitenberg, also members of the INL IMPACT-team and Annette Gotscharek, Uli Refle, Christoph Ringlstetter and Klaus Schulz of the CIS IMPACT-team.

[2] De Denker 1, 1763 <www.dbnl.org>.

[3] EE2/3 and TR5 are cooperations between the Centrum für Informations- und Sprachverarbeitung at Ludwig-Maximilians-Universität München and the Institute for Dutch Lexicology. INL leads EE2/3, LMU leads TR5. Also involved in EE2/3 are the Österreichische Nationalbibliothek, the Deutsche Nationalbibliothek and the Staats- und Universitätsbibliothek Göttingen.

2. Using historical lexica and linguistic models to improve text recognition and accessibility

Two simple examples from the WNT[4] give an indication of the kind of historical language variation we are up against.

Lemma UITERLIJK ('exterior')

> uytterlijcste uyterlijkste d'uyterlijke uiterlyke uyterlijcke uiterlijke uyterlijck uiterlyken uiterlijkste uiterlicke wterlicke wterlijcke ulterlijk uiterlyk uiterlijk uyterlick wterlicken d'uyterlijcke uiterlijken uiterlijks wterlijck uytterlicke uitterlijke ujterlijke uytterlijk uyterlycke uyterlicken uijterlicke d'uiterlijcke wtterlijcke wterlyke wtterlijk (uiterlijke uuterlick uuterlic uyterlijke uyterlijcken uyterlicke d'uiterlyke wterlijke vuyterlijcke uuterlycke uuterlicke wterlijken uyterlijcksten uuyterlicke uuyterlick uuyterlycke uytterl uytterlijcke uytterlycke uytterlick vuytterlicke uiterlijker uyterlyck uterliek wterlijcken uiterlijkst uitterlijk uytterlijcken uyterlyk uiterlijk-net wterlick uutterlijck uuyterlicken uyttelijck uijterlijk uytterlijck uuterlijck uiterlick uitterlyk uuyterlic uuyterlyck uuyterlijck uiterlijck uytterlyck uterlyc wterlijk

Lemma WERELD ('world'):

> 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 werld vvereld weerelts werlde tswerels werreldts weereldt wereldje waereldje weurlt wald weëled

A few orthographical rules would obviously suffice to account for a large part of the variation encountered in the first example. This example also makes clear that for longer words, we can hardly hope to list all variants extensively in the lexicon with reasonable effort. Accounting for the variants in terms of orthographical rules is less obvious for the second example: many variants are largely unpredictable and can only be dealt with by listing them in the lexicon. This is why both linguistic modeling and extensive data development are essential to deal with historical language.

3. Requirements for the IMPACT lexica and linguistic tools

Our aim is to develop historical lexica combining scholarly precision with broad coverage for use in digitization (for both text recognition (TR5) and enhanced retrieval (EE2/3), and to deliver guidelines and a set of tools for the efficient production and deployment of such lexica.

This particular application imposes a few requirements.

First, the lexica need to be allow for specialization to periods or subject matter (for instance, *waereld* should not be included for OCR of texts after 1850). An unstructured, ever-growing set of word forms, without information about the kind of text (in terms of period and subject matter) in which we can expect the words to occur, is neither usable in text recognition nor in enrichment. Frequency information, essential in OCR, will also be added to the lexicon.

Second, the lexica should be 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*).

Lexica used for OCR and retrieval are necessarily incomplete due to the immense amount of possible orthographic variants found in the texts. Hence they need to be complemented by linguistic tools and models to deal with this problem[5].

Since the computational linguistic tools are developed within the context of a European project focusing on mass digitization of historical text, they should be language-

[4] *Woordenboek der Nederlandsche Taal* (cf. <http://gtb.inl.nl>).

[5] The tools and models deal mainly with variation of the 'predictable' type (cf. *uiterlijk* above).

independent (generic) whenever possible, and fit to quickly process large quantities of data.
The fact that linguistic modelling cannot account for all variants entails that the tools should part of a lexicon development workflow involving both automatic and manual processing[6].

4. Corpus-based lexicon structure

The core objects in the lexicon structure developed for IMPACT are word forms, lemmata and documents. All other objects define some kind of relation between these.
In order to enable the OCR's spellchecking mechanism to assess the plausibility of the occurrence of a word in a certain text, 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 words actually encountered in the language and record occurrences in actual texts, with frequency information (attestation);
- record in what kind of texts these words occur (document properties).

It is impossible to extract all possible word forms from the limited amount of available reliably transcribed historical text. Hence, we need mechanisms to extend the lexicon and to enable us 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:

- 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.

In order to effectuate word searches without having to worry about inflection and variation of word forms, 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 unlabelled (i.e. not yet lemmatized or labelled with Part of Speech) or labelled (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-new-words prediction.
- Lexical Source.

Hence, to each labelled or unlabelled word form, we link *attestation* objects which are basically just verified occurrences of the words in documents. The attestations enable us to derive the relevant information about the domain of applicability of word forms from the properties of the documents they occur in. When a word form is taken from a lexicon or dictionary, or when it originates from automatic analysis expansion, we also keep track of its provenance. Apart from the link to the relevant word form and a location in a document, the attestation objects contain the following information:

- verification (yes/no): Whether the occurrence of a labelled word form is checked manually by an expert;
- frequency in a document or document collection.

Two distinct kinds of attestation may be relevant: we may just link a word form to a

[6] In order to deal with variation of the second type (cf. *wereld* above).

document, recording the frequency of occurrence ('attestation at text level'), or we may link to an individual occurrence of the word ('attestation at the token level')[7]. The latter kind of attestation is especially relevant to tagged corpora. In the lexicon building workflow, lemmata may first be assigned on the text level, and ambiguity is not completely resolved. At a later stage, ambiguity may be resolved by assigning lemmata on the token level.

5. Lexicon building

IMPACT will not only deliver a set of tools and lexicon content, but also guidelines and a documented workflow for lexicon development. We will focus on the description of the application of the tools involved in lexicon building, the input -and output management and the manual verification procedures, all of these concerning both regular workflow and workflow specific to the learning cycle leading to improvement of the lexicon. We will not describe the workflow in the form of a 'manual' for the different tools we develop. Instead, we will follow the cookbook metaphor, and describe the 'ingredients' (linguistic data initially available) and the 'utensils' needed for lexicon building, in order to be able to refer to them in the description of the 'recipes'.

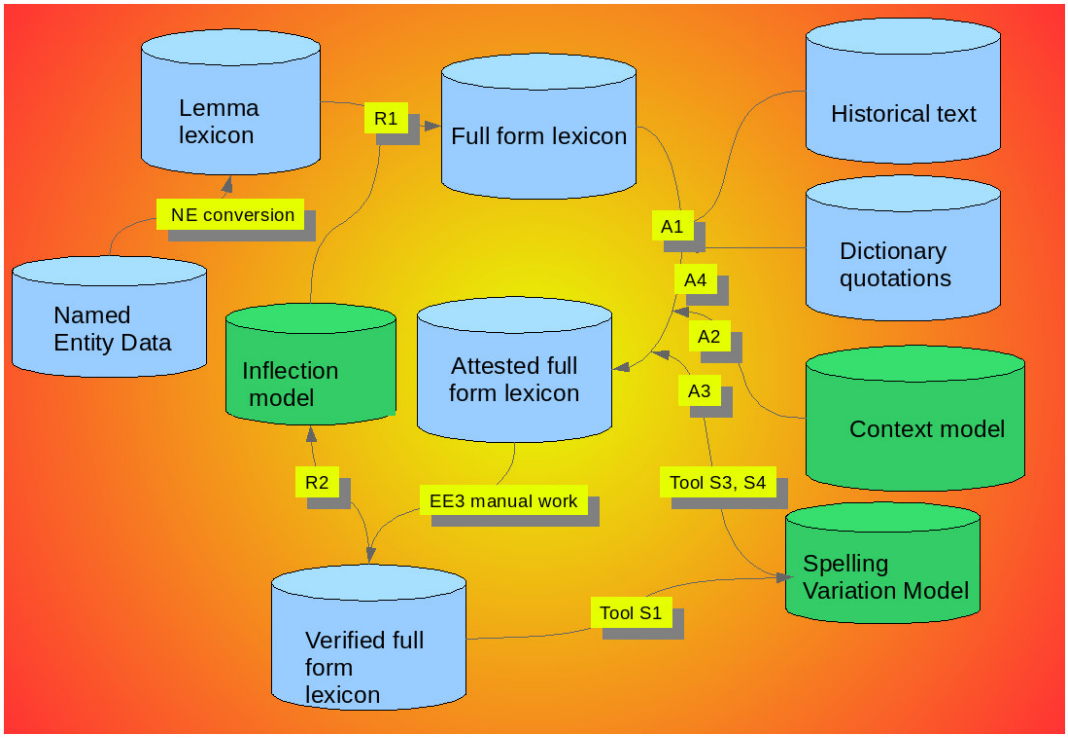


Fig. 1: Lexicon building workflow

6. Data sources to start with

There are different sources from which lexicon building may start: A lemma lexicon: (list of lemmata, for instance the entry list of a historical dictionary); a full form lexicon (list of lemmata with their paradigmatic word forms); historical text, untagged; historical text, lemmatized; historical text, with part-of-speech tags. In the cookbook, we will describe the types of resources that can be useful, give pointers as to where to obtain the material, and suggest parameters for the evaluation of possible sources regarding applicability and quality.

[7] A type is a word form, a token is a particular instance (occurrence) of the type in a text.

Other preconditions for historical lexicon building are:
- an existing modern full form lexicon or at least a lemma list;
- a decent tokenizer, type-frequency list generator;
- graphical User Interface for lexicon building.

7. Computational linguistic tools for lexicon building

This section lists an important part of the cooking utensils, the computational linguistic tools involved in lexicon building. With the exception of tool A5, all are planned IMPACT deliverables. The numbers of the tools correspond to figure 1.

7.1 Tools for reverse lemmatization

Reverse lemmatization is the task of expanding a headword list or a partial full form lexicon to a full form lexicon containing all inflected forms for all headwords.

Reverse lemmatization may start by first extending the headwords of historical dictionaries to the full inflectional paradigm, by using either hand-crafted rules or rules induced from example material. Actual *attestations* confirming the existence and the relevance of the constructed forms will be gathered subsequently by processing text material.

During the lexicon building process, reverse lemmatization will be applied to 'new' lemmata and partial paradigms arising from incorporation of data from corpora or other lexical information.

Tool R1: *Rule-based reverse lemmatization*
INPUT: a set of lemmata (base forms) with part of speech and information about inflectional classes *and* a set of rules specifying how to produce the inflected forms
OUTPUT: a database of inflected forms

Tool R2: *Inflection pattern inference from example material for reverse lemmatization*
INPUT: a partial full form lexicon + the input for previous tool
DESCRIPTION: *b*ased on the rule set, partial lexicon and inflection classes from the previous, the Tool collects statistical patterns relating formal properties of base forms to the inflection pattern (like 'singular ends on *y* → plural may end on *ies*', etc).
*OUTPUT: e*nhanced database of inflected forms and inflection class information.

7.2 Tools for attestation

The automatically generated inflectional paradigm of a dictionary entry does not cover all the language variation found in actual texts. Tools are needed for extracting linguistic information from corpus material in order to add new entries to the lexicon, and inflected forms for existing entries in the lexicon, including provenance (text type) and frequency information necessary for the lexicon deployment in OCR. A special case is the extraction of linguistic knowledge from dictionary quotations.

Tool A1: *attestation with basic matching*
Description: this will simply perform an elementary match of forms in the lexicon with text material, collecting document information and frequencies. Tokenization and case-mapping will be part of this Tool.
INPUT: a partially attested full form lexicon and historical text
OUTPUT: a partially attested full form lexicon (with more attestations)

Tool A2: *Context-aware matching*
Description: this will perform a match of forms in the lexicon with text material, collection

document information and frequencies, using part of speech information and sentence context to discard some of the inherent ambiguity.
INPUT: a partially attested full form lexicon
AUXILIARY DATA: a model assigning a probability distribution of Part of Speech to tokens, based on immediate context
OUTPUT: a partially attested full form lexicon (with more attestations)

Tool A3: *Spelling variation aware matching*
Description: this will perform a match of forms in the lexicon with text material, collection document information and frequencies, using a model of historical spelling variation
INPUT: a partially attested full form lexicon and historical text + a model of spelling variation in the form of a weighted pattern set
AUXILIARY TOOL: historical spelling matcher
OUTPUT: a partially attested full form lexicon (more attestations)

Tool A4: *Matching the headword in dictionary quotations*
Description: this is a special case of the attestation task, by which the form of the headword needs to be attested in a dictionary quotation. The purpose of this special case is to exploit the fact that finding the match is easier than in free text of arbitrary length, and to use the fact that dictionary quotations tend to record non-standard variants.
INPUT: a file consisting of pairs of: {lemma, quotation containing a word form of the lemma}, supplemented by lexicon, spelling variation model
OUTPUT: the input file with the word form corresponding to the lemma marked in each quotation, lexicon with more attested word forms.

These are the tools mentioned in the IMPACT description of work. Of course, there are other possibilities, like for instance the following:

Tool A5: *Matching by using text alignment*
This involves the use of parallel texts in old and newer language, for instance a respelled edition and an edition in original spelling of a certain work. For Dutch, one could try to use the 1637 and 1888 Statenvertaling versions of the bible. For German, there are 1554 and 1912 versions of the Luther bible. This can give useful example material. An alignment with GIZA++[8] of the latter yields for instance (historical variants underlined):

> Am/AM Anfang/anfang schuf/schuf Gott/Gott *Himmel/Himel und/vnd Erde/Erden*
> *Und/Vnd* die/die Erde/Erde war/war wüst/wüst *und/vnd* leer/leer
> *und/vnd* es/es war/war finster/finster *auf/auff* der/der *Tiefe/Tieffe und/Vnd* der/der Geist/Geist Gottes/Gottes *schwebte/schwebet auf/auff* dem/dem Wasser/Wasser

7.3 Tools for dealing with spelling variation

Tool S1: *pattern inference from example material in the form list of pairs (normalized word form, historical word form)*
INPUT: a set of pairs {normalized word, historical word}, as a text file
OUTPUT: a set of patterns with weights
EXAMPLE: this will infer typical patterns like *ae/aa*, *y/ij*, *ck/k*, *qu/kw* from a set of word pairs like (*aengenaem, aangenaam*), (*qualiteit,kwaliteit*), etc.

Tool S2: *pattern inference from material in the form (historical word list, normalized word*

[8] Och, ney 2003: 19-51.

list)
INPUT: a normalized word list and a historical word list
OUTPUT: a set of weighted patterns and a partial mapping between items on the two lists.
COMMENT: this tool can be used to match a modern lexicon against historical text when no example material in the form of the previous example is available.

Tool S3: application of patterns: matching against a word list
INPUT: a set of weighted patterns, a word list, a target word w
OUTPUT: a weighted set W = (w_i,p_i) of words from the word list, such that application of patterns maps each word w_i to w with probability p_i.
USE: this can be used either to find matches for a normalized search term in historical text, or to match a historical word form against a word list in normalized spelling

Tool S4: constrained variant generation
INPUT: a set of patterns, a set of normalized words N (for instance a modern lexicon), a set of historical words H (for instance the words occurring in a given document)
OUTPUT: a weighted partial multi-valued mapping (n_i, h_i, p_i) from N to H, such that application of patterns maps each word n_i to h_i with probability p_i.
USE: this is logically equivalent to a repeated application of the previous to each item I of a word list, but much more efficient.

8. The lexicon building workflow[9]

This section describes the two major recipes for lexicon building resulting in attested word forms, involving different data sources for lexicon development and the tools as described in the previous section. Our purpose is, in both cases, to build a diachronic word form lexicon that contains spelling variants and morphological variants of words that have appeared in documents over a certain period.
Some important properties of the resulting word form lexicon are:
- it contains the modern lemma corresponding to the historic word form;
- it provides attestations representing genuine usage of the words in historical texts;
- the attestations have bibliographical information, including date.

There are several ways to build such a lexicon starting from language data like a diachronic corpus, a modern full form lexicon, and a historical dictionary.

8.1 Lexicon building using a full form lexicon and historical text (without morphological analysis)

[9] See also figure 1.

Initialization:
Lexicon := some full form lexicon, f.i. CISLEX for German, e-LeX for Dutch
Patterns := some initial set of spelling patterns (aa/ae), perhaps the empty set.

While (not satisfied with coverage of lexicon)
{
Step 1. Process selected texts with lexicon and orthographical variant patterns.
Step 2. Split the words from the texts in 3 subsets (using CL deliverable Tool A1, Tool A3)
W_1 = exact match with lexicon
W_2 = match with lexicon, using patterns (= match in the 'hypothetical lexicon'
W_3 = not found at all
Step 3. Manual checking, using the corpus-based lexicon building GUI (0) in combination with the context view for token-level attestations.
- For *w* in W_1, possibly check ambiguous word forms for lemma assignment
- For *w* in W_2,
  - check the matched lemma (e.g. word form: bieck/ lemma: bakken)
  - check the matched 'normalized' word form (e.g. historical bieck, normalized biek
  now either:
  (i) The match ok (no action required)
  (ii) The match is not ok, but a match with existing lemma/normalized word form is possible (correct)
  (iii) No match is possible with existing data. (Move w to W_3)
- For *w* in W_3, there are three possibilities.
(i) *w* can be matched with lexicon, using hitherto an unknown pattern. In this case we manually add match with modern (normalized) word form, so the pattern inference tool S_1 can infer a new pattern
(ii) word is a new word form corresponding to an existing lemma
(iii) word belongs to a hitherto unknown lemma
In case (ii) or (iii): add lemma and/or normalized word form to the database
Step 4. Rerun pattern inference (CL Tool S1) (we now have new example data)[1]
Step 5. back to 1.
}

Figure 2: corpus-based lexicon building

Example:

Text = '*Terwyl wy hier van woningen spreken, moet ik zeggen dat my in deze Stadt vremt voorquam het maexel van huizen, die geheel voltoit hier op de markt te koop gebragt worden.*'
Initial Lexicon = {*terwijl, wij, hier, woning: {woning, woningen), van, spreken, moeten, zeggen, dat, mij, in, deze stad, ik, vreemd, het, huis, huizen, die, voorkwam, geheel, voltooid, hier, op, de, markt, te, koop, gebracht, worden*}
Initial Patterns = { *y/ij, qu/kw, ae/aa, g/ch, ch/g*}
After step 2:
W_1 = { *hier, van, woningen, …* }
W_2 = { *terwyl, wy, my, voorquam, gebragt* }
W_3 = { *stadt, vremt, maexel, voltoit* }
In step 3:
Add to lexicon: new lemma maaksel,
Add for pattern inference: examples (maaksel, maexel); examples (stadt, stad), (vremt, vreemd), (voltoit, voltooid)
After step 4 (Rerun pattern inference): new patterns { x/ks, dt$/d$, oi/ooi, t$/d$}
Please note that this example is not entirely realistic: pattern inference only works for a large example set

Figure 3: example for the workflow in figure 2

The purpose is to retrieve historical variants from historical corpus data. A significant part of all manual work involved in lexicon building is covered in this recipe. This means that this part of the workflow has to be extremely efficient. Figure 2 describes the acquisition process, in which not only the lexicon content grows, but also the model of orthographical variation

adapts to new examples.

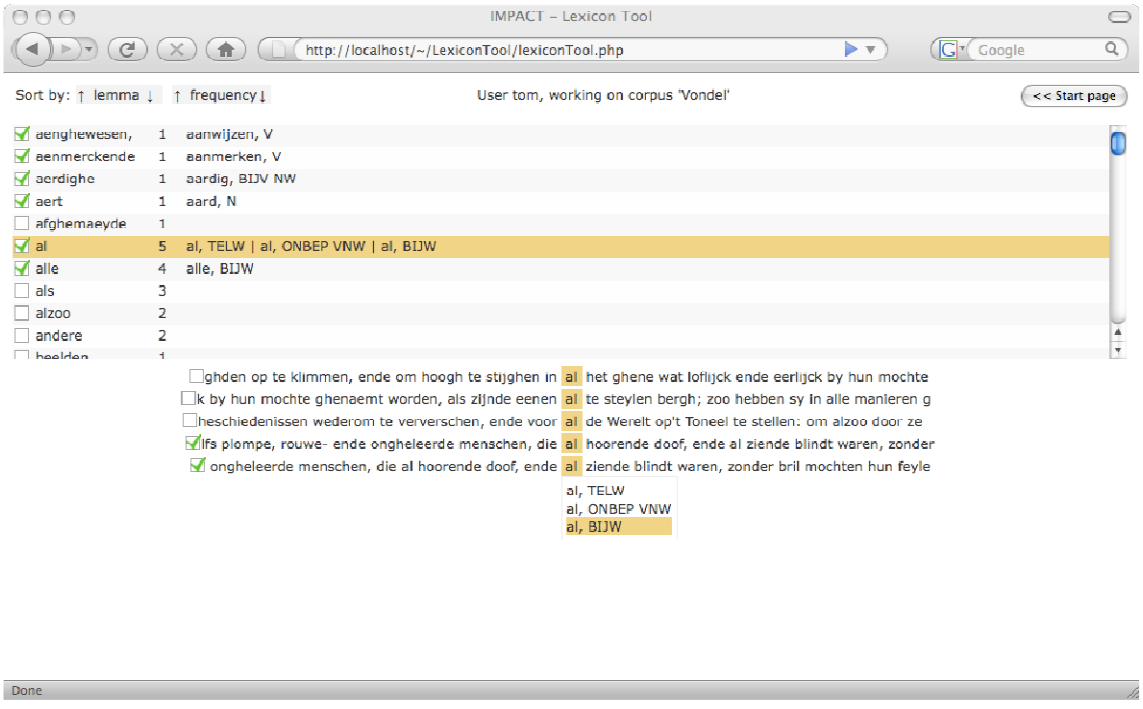
8.2 Lexicon building from historical dictionary quotations

Historical or diachronic scholarly dictionaries tend to include numerous quotations from different periods illustrating the usage of words in historic texts. The main idea is to use these dictionary quotations and the associated bibliographical information as attestations of word forms. These quotations exemplify the usage of the head word of a dictionary item; the lemma. Usually the word form in the quotation which corresponds to the lemma is not explictly marked in the digital versions of the dictionary. We developed a method to match the lemma to the corresponding word form in each quotation. This method consists of two separate processes. First, we apply automatic preprocessing to select the most probable candidate word form in the quotations. The results are stored in a database. Secondly, the results are manually verified and corrected using a specially designed tool (cf. section 0).

9. Graphical User Interface Tools for manual work in lexicon building

Obviously, we cannot do without GUI tools for manual work in lexicon building. This section presents a prototype of the tool for corpus-based lexicon building and the finished GUI for attestation from historical dictionary quotations.

9.1 User interface for corpus-based lexicon building



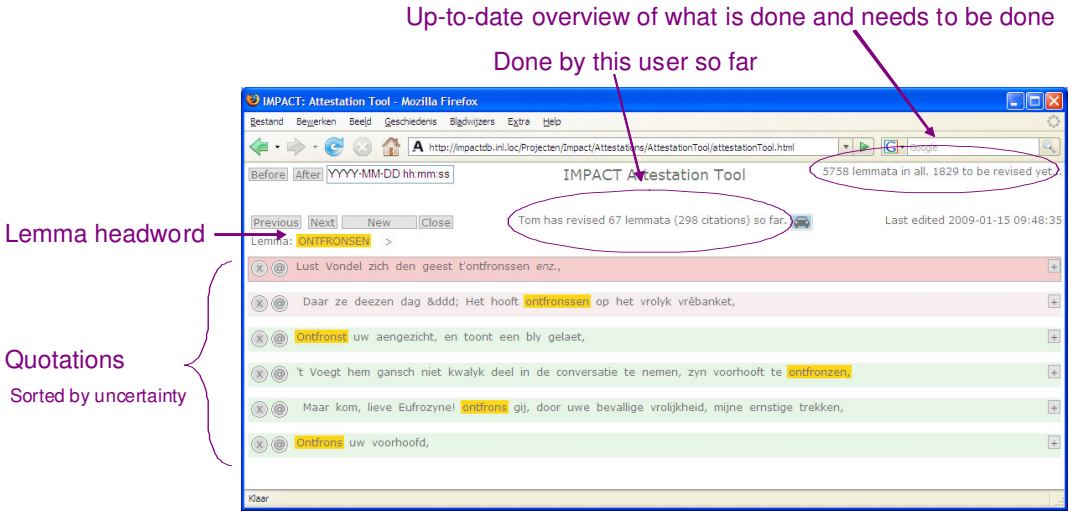
The manual work here consists of the checking and correction of the automatical lemma and part of speech assignment for (part of) the vocabulary found in the document or corpus under investigation. The vocabulary may be restricted to certain selections to speed up bulk processing, for instance to unknown words, words that can be matched to existing lexicon content using reliable patterns of spelling variation, etc.

The GUI presents a split view consisting of the type-frequency list from the text(s) in

the top part of the screen and a KWiC[10] view of the occurrences of the current word form in the text(s) in the bottom part of the window. The top part allows for the editing of lemma assignment and part of speech, resulting in text-level attestations of the current word form, with possible ambiguity.

In the KWiC view of the contexts in which the current word form occurs, token-level attestations may be added by choosing one of the possible interpretations for each corpus token.

9.2 Attestation in dictionary quotations



The manual work here consists of checking the automatically marked occurrences of the dictionary headword in the quotations belonging to the lemma. The tool lists quotations per lemma. Quotations are ordered by uncertainty. The most uncertain ones (containing words least similar to the headword) appear at the top and are marked red(dish). Literal matches are at the bottom, marked green.

By using the arrow keys or the mouse, users can select or deselect words or move a selection. The 'X' button can be used to mark quotations requiring special attention (e.g. because they were extracted in the wrong way). The '?' button can be clicked to mark quotations that are 'unfortunate' (e.g. the headword doesn't appear in the quote as such but only in a compound).

Auto attestation: When a very frequent variant has been missed in automatic matching, auto attestation can come in handy. A user can select a word and, by hitting the auto attestation button, all occurrences of this word form can be highlighted.

Keyboard shortcuts: To enhance the usability the interface can be used with the mouse, with the keyboard or both.

Features and system requirements: The Attestation Tool is based on a LAMP[11]

[10] KWiC: Keyword in context.

[11] The acronym LAMP refers to a solution stack of software, usually free and open source software, used to run dynamic Web sites or servers. The original expansion is as follows:

- Linux, referring to the operating system;
- Apache, the Web server;
- MySQL, the database management system (or database server);
- PHP or others, i.e., Perl, Python, the programming languages.

architecture. Users only need a web browser. The interface consists of just one page: attestationTool.php. It is a so-called rich Internet application which means that it uses AJAX to communicate with the database server and display the results.

The tool has been built for speed.[12] When the automatic matching has worked out reasonably well users can very easily scan through the results, correct some mishaps and hit the spacebar to get the next lemma.

10. Related work

We are aware of other approaches to the building of lexical resources. Much work has been done on rule-base generation of full-form lexica, cf. f.i. Evans and Gazdar 1996. For statistical work on full-form lexicon generation or inference from corpus data cf. for instance (de Loupy & Gonçalves 2008; Sagot 2008). Graphical user interfaces to speed up the manual work are presented in Fontenelle 2008 and Ferreira et al. 2008.

Any approach to lemmatization of unknown words is relevant to our work (Cucerzan and Yarowski 2000, van den Bosch & Daelemans 1999). There are, however, three ways in which our approach stands out as different from existing ones:

- integration of approaches to spelling variation;
- corpus-based approach: always store information about the attestations of word forms;
- combination of automatic acquisition with a workflow for manual work.

Bibliography

Bosch, Antal van den, Walter Daelemans (1999), 'Memory-based Morphological Analysis', in: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*.

Chrupala, Grzegorz, Georgiana Dinu, Josef Van Genabith (2008), *Learning Morphology with Morfette*, LREC 2008.

Cucerzan, Silvio, David Yarowsky (2000), 'Language independent minimally supervised induction of lexical probabilities', in: *Proceedings of ACL-2000.*

Evans, Roger, Gerald Gazdar (1996), 'DATR: A language for lexical knowledge representation', in: *Computational Linguistics* 22.2, 167-216.

Ferreira, José Pedro, Sílvia Barbosa, Maarten Janssen (2008), *Mordebe Admin — A Lexical Management System*, Euralex 2008.

Fontenelle, Thierry, Nick Cipollone, Mike Daniels, Ian Johnson (2008), 'Lexicon Creator: A Tool for Building Lexicons for Proofing Tools and Search Technologies', in: *Proceedings of Euralex 2008.*

Gotscharek, Annette, Ulrich Reffle, Christoph Ringlstetter, Klaus Schulz (2009), *On Lexical Resources for Analyzing Historical Documents. Submitted to AND workshop on noisy data*, Barcelona, ICDAR 2009.

Hauser, A, M. Heller, E. Leiss, K.U. Schulz, C. Wanzeck (2007), 'Information Access to Historical Documents from the Early New High German Period', in: *IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*. Hyderabad, India - January 8, 2007, 147-154

Loupy, Claude de; Sandra Gonçalves (2008), 'Aide à la construction de lexiques morphosyntaxiques', in : *Proceedings of Euralex 2008*.

Och, Franz Josef , Hermann Ney. 'A Systematic Comparison of Various Statistical Alignment Models', in: *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

Sagot, Benoît (2005), 'Automatic Acquisition of a Slovak Lexicon from a Raw Corpus', in: Václav, Matousek, Pavel Mautner, Tomás Pavelka (eds), *Text, Speech and Dialogue, 8th International Conference*, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005, Proceedings.

---

http://en.wikipedia.org/wiki/LAMP_(software bundle)

[12] Currently, the speed of manual correction is: 1725 quotes/hour 231 lemmata/hour.