Developing GiGaNT, a lexical infrastructure covering 16 centuries

Tilly Ruitenberg†, Jesse de Does and Katrien Depuydt
Instituut voor Nederlandse Lexicologie, Leiden, Nederland

GiGaNT is a new INL initiative which sets out to develop a computational lexicon (lexical database) covering 16 centuries of Dutch language. This means that all lexical data of the dictionaries, corpora and computational lexica of the Institute for Dutch Lexicology (INL) will be stored into a central database, functioning both as computational lexicon and central infrastructure for the maintenance of lexical data. Dictionaries, corpora and this computational lexicon are all part of the Dutch Language Bank (DLB).
The immediate incentive to develop GiGaNT was the need for a diachronic computational lexicon, to serve both as a link between texts and dictionaries in the DLB and as a solid infrastructure for other, similar lexical data at the INL. The GiGaNT lexicon will be used for text or corpus annotation, facilitating the retrieval and investigation of the annotated texts.
Integration of existing material into GiGaNT and its subsequent adaptation to enable it to function within computational applications will be a huge step towards another aim: the systematic screening of the complete Dutch word stock for 'gaps' in lexicographic description. This applies to both neologisms and hitherto undescribed historical words.
Users will benefit from the possibility to link from word forms in running text to lexicographical definitions in the INL dictionaries. Researchers, who now only have access to separate collections, will benefit as well: in the future they will have one single starting point for their searches and one single basis from which to develop new lexical material. GiGaNT will also give expert users better access to the lexical data maintained by the INL. The infrastructure will function as a database which will be accessible to API's and as a 'service' that enables researchers to compare their data with GiGaNT and eventually to contribute their own material to GiGaNT.

1. Short description

GiGaNT is a new INL initiative which sets out to develop a computational lexicon (lexical database) covering 16 centuries (6th – 21th century) of Dutch language. This means that all lexical data of the dictionaries, corpora and computational lexica of the Institute for Dutch Lexicology (INL) will be stored into a central database, functioning both as computational lexicon and central infrastructure for the maintenance of lexical data. Dictionaries, corpora and this computational lexicon are all part of the Dutch Language Bank (DLB).

The immediate incentive to develop GiGaNT was the need for a diachronic computational lexicon, to serve both as a link between texts and dictionaries in the DLB and as a solid infrastructure for other, similar lexical data at the INL. The GiGaNT lexicon will be used for text or corpus annotation, facilitating the retrieval and investigation of the annotated texts. Applying the GiGaNT tagset and lemmatization principles will guarantee easy 'storage' in the database as well as easy compatibility, both with each other and with the primary data of the GiGaNT database: the five large scholarly dictionaries of the INL.

Integration of existing material into GiGaNT and its subsequent adaptation to enable it to function within computational applications will be a huge step towards another aim: the systematic screening of the complete Dutch word stock for 'gaps' in lexicographic description: word forms lacking in the paradigms or word senses (which have) not yet (been) described in the dictionaries. This applies to both neologisms and hitherto undescribed historical words.

Users will benefit from the possibility to link from word forms in running text to lexicographical definitions in the INL dictionaries: ONW, VMNW, MNW, WNT[1] for historical texts and ANW[2] for modern material.

Researchers, who now only have access to separate collections, will benefit as well: in the future they will have one single starting point for their searches and one single basis from which to develop new lexical material. GiGaNT will also give expert users (e.g. computational linguists) better access to the lexical data maintained by the INL. The infrastructure will function as a database which will be accessible to API's and as a service that enables researchers to compare their data with GiGaNT and eventually to contribute their own material to GiGaNT.

GiGaNT starts out as a morphosyntactic, corpus-based lexicon. Eventually, it will also incorporate semantic and syntactic information.

2. Source material for the development of GiGaNT

The database will only contain written language, including both general vocabulary and named entities. GiGaNT does not primarily focus on dialectical variation, but since it covers 16 centuries of Dutch, and standardization only starts from the 17th century, dialect material cannot be excluded.

The initial population will be from existing material: the four historical INL dictionaries. This means that the period 600 – 1976 will at least be covered by the existing dictionary content. These scholarly dictionaries include ample quotations with full bibliographical information. This 'dictionary quotation corpus' is, together with the lemmata, one of the primary sources for GiGaNT. A special toolset for the extraction of lexicon content from this corpus has been developed and has successfully been applied to the WNT.

Apart from the dictionaries, GiGaNT will be fed by corpora maintained by the INL and by large external text corpora such as old newspaper collections digitized by the KB[3] and the Dutch DBNL. Priority will be given to text corpora covering the period of the lexical material under development. For example, the INL is currently involved in the IMPACT project, which focuses, in the case of Dutch on two periods: the 18th and 19th centuries. This means that we are now concentrating on texts from these periods.

Material from existing morphosyntactic lexica, such as e-Lex[4], will, in due time, also be incorporated. In any case, external sources will always need to be converted to the GiGaNT database structure, and the annotation will need to be 'translated' to the GiGaNT value set.

[1] ONW: Dictionary of Old Dutch; VMNW: Dictionary of Early Middle Dutch; MNW: Dictionary of Middle Dutch; WNT: Dictionary of the Dutch Language. Online at http://gtb.inl.nl.

[2] ANW: Contemporary Dictionary of Dutch; a demo version is currently online at http://anw.inl.nl.

[3] National library of the Netherlands.

[4] http://www.inl.nl/index.php?option=com_content&task=view&id=356&Itemid=668.

2.1. Paradigmatic expansion and 'hypothetical' lexicon content

Most full-form lexica contain word forms which have never been found in corpora. Strictly limiting lexicon content to attested forms means that complete paradigms will only be found for frequent words and that the content will be incomplete from the point of view of applications which require full form word lists, like OCR or OCR postcorrection[5].

Apart from so-called attested word forms, GiGaNT will also incorporate expanded or generated word forms. This means that paradigms with gaps will be completed with hypothetical word forms, which have been generated on the basis of rules derived from the paradigms in question.

These 'hypothetical' word forms may belong to both historical and modern language and can be generated by our own tools or taken from existing lexica with expanded forms. Generated word forms will of course be 'flagged' as 'not attested'. Eventually they may be looked up and actually be found in new text corpora. Of course, for some applications, expansion is unnecessary and hitherto unattested word forms may as well be analyzed on the fly. For other tasks, for instance OCR-postcorrection, a simple list of expanded forms is preferred to a morphological solution. However, the larger the lexicon and the amount of attested word forms, the smaller the need for expanded word forms. Expansion is a good temporary solution.

3. Requirements for the lexicon

Our aim is to develop a diachronic lexicon combining scholarly precision with broad coverage both to be used in compuational linguistic applications and as a central database with information on the vocabulary of Dutch. This imposes a few requirements.

First, the lexica need to allow for specialization to periods or subject matter. It should for instance be possible to exclude historical variants like *waereld* for *wereld* when working with recent text material. An unstructured, ever-growing set of word forms, without attestation information about the kind of text (in terms of period and subject matter) in which we can expect the words to occur, is usesless for most purposes (both computational and lexicographic). This means that every single word form (entry) will be linked to its occurrence in the text, together with its 'attestation data', i.e. context and type, location, author and date of the text. Frequency information will also be added to the lexicon.

Second, the lexicon 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*).

Any diachronic lexicon is necessarily incomplete due to the immense amount of possible orthographic variants found in Dutch historical texts. Hence it is currently being complemented by linguistic tools and models to deal with this problem. 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.

[5] Application of lexicon content for these purposes is part of IMPACT.

Finally, the lexicon must be interoperable with other data. The lexicon structure is therefore not only implemented as a relational database. XML export and import modules to (an extension of) LMF are currently under development.

4. Data categories in GiGaNT

Apart from lemma, PoS (coarse-grained or fine-grained) and attestation information (including frequency information), each entry will be provided with a flat morphological analysis. In due time, semantic and syntactic information will be added as well.

4.1. Lemmatization

The lemma in GiGaNT is a modern lemma, assigned according to well-established principles. One single lemma will link diachronic, regional and orthographic variants. We add the corresponding modern form rather than a modern translation of the word. Consistent application of this principle means that in the case of words which have disappeared from modern usage, we assign a 'modern' form synthesized by following regular etymological developments. For instance, the modern lemma for 'aenvaerdighen' is not 'aanvaarden', but 'aanvaardigen'.

Lemmatization is enormously helpful in circumventing spelling variation in older texts during retrieval. Apart from this, the modern lemma is also the main instrument in cross-linking historical corpus texts with modern language data, the main historical dictionaries of Dutch, and other corpora lemmatized in this way.

4.2. Approach to part of speech tagging

Traditionally, morphosyntactic tagsets contain the following types of information beyond the basic part of speech:

1. functional/syntactic information (transitive/intransitive verb usage, attributive/predicative/adverbial, etc).
2. traditional paradigmatic information (person, number, case, mood, tense)
3. strictly formal features (prefix/suffix information, vokalstufe, ..)

For the part of speech set in GiGaNT, we propose to combine paradigmatic information with formal information by dividing the part of speech information into two distinct parts: $T_{f(ormal)}$ and $T_{p(aradigmatic)}$. In the T_f part, the tag directly mirrors the form, for instance VRB(infl=-t). In the T_p part, more traditional paradigmatic labels are assigned, such as person, number, tense, mood.

word forms	*main part of speech*	*formal tag part (T_f)*	*paradigmatic tag part (T_p)*
Modern dutch			
zeggen	*VRB*	(infl=-en)	(1,pl,pres,ind)
zeggen	*VRB*	(infl=-en)	(2,pl,pres,ind)
zeggen	*VRB*	(infl=-en)	(3,pl,pres,ind)
zeggen	*VRB*	(infl=-en)	(-,-,-,inf)
Early middle dutch			
sech, seg(h), segg,	*VRB*	Infl=0	1e sg.ind.pres.
secge, seche, seg(h)e, segg(h)e	*VRB*	Infl=e	1e sg.ind.pres.
secg(h)en, segg(h)en, zegghen,	*VRB*	Infl=en	1e pl.ind.pres.
secghe, segg(h)e, zegghe (when followed by **wi**)	*VRB*	Infl=e	1e pl.ind.pres.

sech, seg(h),	*VRB*	Infl=0	imp.sg.
seg(h)e	*VRB*	Infl=e	imp.sg.
sagit, secget, sec(h)t, segg(h)et	*VRB*	Infl=t/et	imp.pl.
segg, segh, *uncertain*	*VRB*	Infl=0	1e sg.conj.pres.
sage, segge	*VRB*	Infl=e	1e sg.conj.pres.
segg,	*VRB*	Infl=0	3e sg.conj.pres.
sage, secghe, segghe	*VRB*	Infl=e	3e sg.conj.pres.

Example (from VMNW article *zeggen*)

Full paradigmatic tagging of the complete lexicon content is extremely time-consuming; manual tagging of verbal mood, tense, number and person has for instance proven to be unfeasible in several tagging projects[6] (the example illustrates the problems). However, completely discarding the traditional paradigm would be a considerable loss to researchers of inflectional morphology and the usage of inflectional categories. We intend to make up for this loss in two ways:

1. Similar to the description of inflection in the dictionaries of Early Middle and Old Dutch, we will list, for each lemma, all occurring different word forms with their full paradigmatic tagging, together with attestation data. We will use these data to keep track of the (many-to-many) *mapping* between formal and paradigmatic features, enabling at least the retrieval of potential occurrences of the categories one is looking for. That way we may not support extraction, with full precision and recall, of verbs in the subjunctive mood, but the researcher will at least be offered a set of possible candidates.
2. At least one attestation will be linked for each separate quadruple consisting of word form, lemma, T_f -tag part and full T_p tag.

4.2.1. Attestation

GiGaNT will be growing continuously. This will increase its effectiveness for certain tasks: the larger the lexicon, the more effective the tools which deploy the lexicon and which will fill the gaps. It is equally true, however, that eventually the growth of lexicon content irrelevant to certain text types or tasks might become a nuisance rather than an asset. To address this problem, sensible and goal-oriented selections of lexicon content based on, for example, frequency and period will be needed. Attestations and document-specific metadata (dating, localization, text type) are, of course, prerequisites for this solution.

An important feature of the lexicon is therefore the inclusion of *attestation objects* which link tag and lemma assignments to an occurrence in a text. Attestation objects store the link to the relevant word form and a location in a document.

Two distinct levels of attestation are relevant: the first is the linking of a word form to a document, (‘attestation at text or corpus level’), the second is the linking to an individual occurrence of the word (‘attestation at the token level’)[7]. The latter type of attestation can be used to store corpus tagging. In the lexicon building and corpus annotation workflow, lemma and part of speech may first be assigned on the text level, and ambiguity is not completely resolved. At a later stage, ambiguity may be resolved by assigning annotation on the token level. Text-level attestations are linked to the occurrences of a word in a text, without specifying the location in the document.

[6] I.c. Corpus Gysseling and Corpus van Reenen-Mulder (13th and 14th-century Dutch, resp.).
[7] A type is a word form, a token is a particular instance (occurrence) of the type in a text.

4.2.2. Morphological Analysis

In due time, each lemma will be analyzed in terms of its immediate constituent parts. The parts will be linked to the corresponding lemma entries. A 'deep' analysis can be performed by storing, recursively, the analyses of the immediate constituents (*mandenmakersschaaf* is analyzed as a nominal compound of *mandenmaker* + *schaaf*, a deeper analysis can be stored if *mandenmaker* is analyzed in its turn as *mand* +*maken* + *'er'* etc.). We do not exclude resorting to a 'flat' analysis of a compound in cases where there is no clearly preferred hierarchy (*lucht+afvoer+kanaal*). There is often no need to choose between different 'bracketings' of a compound.

5. Feasibility and workflow

Adding all information categories to a lexicon of this envisaged size is an enormous task. Moreover, not all information categories are always necessary for all applications of the lexicon. We therefore designed the database structure and the lexicon development workflow in such a way that data in various states of processing and in various levels of detailedness can be incorporated into the lexicon. For example:

1. word forms with their lemma's and main Part of Speech (without any manual correction)
2. word forms only linked to their place in the text
3. word forms which have been verified for lemma and main Part of Speech
4. word forms with their attestation data and checked lemma and detailed Part of Speech

Despite the different status of the data, the lexicon will be accessible for searches and exploitation at any time during its construction. This setup is necessary not only because of the amount of data, but also because the database is perpetually evolving.

6. Lexicon development: tools

Lexicon development for GiGaNT comes with the following main tools: a tool for attestation in dictionary quotations, a tool for corpus-based lexicon development, tools for dealing with spelling variation, and a lemmatizer for historical Dutch. Some of the tools for lexicon building have been developed in accordance with the requirements of the current projects: IMPACT[8] and the processing of dictionary data to be entered in GiGaNT.

The tool set will be documented and the procedure for lexicon building will be described in a 'lexicon cookbook'. A detailed description of the lexicon building process in IMPACT can be found in De Does and Depuydt, 2009.

The tool set will be extended with at least a Part of Speech tagger for historical Dutch.

[8] Improving Access to Text, www.impact-project.eu. The project aims to significantly improve access to historical text material and to take away the barriers that stand in the way of / impede the mass digitization of the European cultural heritage.

7. Example: attestation data for 11 centuries in the life of a verb

For this example, we will use the verb WASSEN (*to wash)*. It corresponds to *waskan* (ONW) and *wasschen* (VMNW, MNW and WNT). We will consider a small part of the verbal paradigm in the lexicon: the preterite indicative singular forms.

Our starting point is a dictionary quotation corpus consisting of 2 ONW quotations, 16 VMNW quotations, 36 MNW quotations and more than 600 WNT quotations. Most of these dictionary quotations have been dated. In order to enable the date filtering mechanism for the extraction of period-specific lexica for various purposes, we rely heavily on dated attestations: to be able to ear-mark a word with any confidence for inclusion in a period-specific lexicon for the period [a – b], we should have an earliest attestation dated <= a and a latest attestion dated >= b. This is why we focus on finding the earliest and latest attestations.

word form	*person*	*earliest attestation*	*latest attestation*
uuosc	1	ONW, 901-100	ONW, 901-100 (=earliest) *in* ***uuosc*** *under unsculdigin hendi mina.*
wiesch	3	VMNW, 1276-1300	WNT, 1911
woysch	3	MNW, 1470	MNW, 1470 (== earliest)
wosche (l. woschse) (clitic combination, wosch + se)	3	MNW, 1451-1500	MNW, 1451-1500 (== earliest)
waschte	1,3	WNT, 1641 (1 sg)	WNT, 1889 (3 sg)
waste	3	WNT, 1693	WNT, 1693 (==earliest!)

These data illustrate that even a large dictionary quotation corpus will not suffice. The paradigm is not complete. There is, for instance, no recent attestation for the standard modern spelling '*waste'.* This is why we also need to explore corpus material, including Web data, which of course immediately yields recent attestations such as: '*Een achttien maanden oud apewijfje vond de oplossing : zij waste aardappelen in een nabijgelegen rivier schoon.'* [9]

8. Current situation and future work

The structure of the database has been designed and developed as part of the INL's participation in IMPACT. The Part of Speech tagset and the principles for lemma assignment have already been defined and described.

Lexicon content has been extracted and manually checked, by means of the application of the attestation tool, from the quotation corpus of the largest historical dictionary, the WNT. The lexicon currently contains 560,000 lemma-word form-part-of-speech entries from 211,000 lemmata linked to 1,3 million dictionary quotations. Currently, work is being done on elaborating the lexicon with corresponding corpus material. Morphological analysis, concerning both the systematics of the description as well as the tools for automatic morphological decomposition is also being tackled.

The four historical INL dictionaries are interlinked by two common elements: the modern lemma and the main part of speech assigned to each lemma in the dictionaries. One major challenge will be to link the lemmata on the basis of a more strict (etymological) identity

[9] http://aquariusonline.wordpress.com/2009/10/12/bizarre-feiten/ (October 2009).

criterion. We will also work on the extension of the lexicon content towards both newer and older material. Further development of modern lexical data will be undertaken in cooperation with the General Dictionary of Dutch (ANW); lexicon content for Middle Ducth will be extracted from the MNW quotation corpus (continuing work by John van der Voort van der Kleij).

References

Depuydt, Katrien; de Does, Jesse (2008). 'United in Diversity: Dutch Historical Dictionaries Online'. In Bernal, E.; DeCesaris, J. (eds.). *Proceedings of the XIII EURALEX International Congress*. 1237-1241.

Depuydt, K.; de Does, J. (2009). 'Computational Tools and Lexica to Improve Access to Text'. In Beijk, E.; Colman, L. et al. (2009). *Fons Verborum. Feestbundel voor prof. dr. A.M.F.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het INL*. Leiden/Amsterdam: Gopher Publishers. 187-199.

Francopoulo, G.; Bel, N.; George, M.; Calzolari, N.; Monachini, M.; Pet, M.; Soria, C (2007). 'Lexical Markup Framework: ISO standard for semantic information'. Tübingen: GLDV (Gesellschaft für linguistische Datenverarbeitung). (www.lexicalmarkupframework.org).

Francopoulo, G.; George, M.; Calzolari, N.; Monachini, M.; Bel, N.; Pet, M.; Soria, C. (2006). 'Lexical Markup Framework (LMF)'. In *Proceedings LREC 2006*. 233-236.

Kruyt, J.G. (2004). 'The Integrated Language Database of 8th - 21st-Century Dutch', in: M.T. Lino; M. F. Xavier; F. Ferreira; R. Costa; R. Silva (eds.). *Proceedings of the 4th International Conference on Language Resources and Evaluation*. 1751-1754.

van der Voort van der Kleij, J. (2005). 'Reverse lemmatizing of the Dictionary of Middle Dutch (1885-1929) using pattern matching'. In *Papers in Computational Lexicography, COMPLEX* 1. 233-242.