TICCLAT: a Dutch diachronic database of linked word variants

Martin Reynaert
DHLab & Meertens Institute
KNAW Humanities Cluster, Amsterdam
& Tilburg University
The Netherlands
reynaert@uvt.nl

Janneke van der Zwaan / Patrick Bos
Netherlands eScience Center
Amsterdam
The Netherlands
j.vanderzwaan@esciencecenter.nl
p.bos@esciencecenter.nl

Introducing TICCLAT Digital Humanities (DH) research makes extensive and ever-increasing use of digital text corpora. Due to shortcomings in the digitization process, unknown quantities of word tokens cannot be found in the form in which they were originally written. In addition, words display different appearances over time as language evolves and spelling changes, whether or not standardized. Very many word forms produced at some time or another have not ever been attested. Due to the free compounding nature of Dutch, anyone can take just about any two or three existing words, join them and write them as a single new word. These compound neologisms have traditionally not really ever made it into lexicons or dictionaries. All this greatly hampers the reliability of the results of DH research. So, in hopes of remedying this situation, we build a system for lexical assessment based on the 18.5 billion word tokens in the diachronic Nederlab corpus (Brugman et al., 2016), which contains sources dating back as far as the 13th century. Of the text available in this largest corpus of diachronic Dutch, only about 6.5% percent was born-digital. We therefore set off Nederlab's vocabulary against a compendium of the finest Dutch lexical resources available to us. Our system is called TICCLAT, in full: 'Text-Induced Corpus Correction and Lexical Assessment Tool'.

The general idea in this work in progress has ever been that we start with the best contemporary resources we have and generalize as much as we can from the related word form information present in order to link related words so that the resultant word clusters can effortlessly be retrieved. Then to augment these with the evidence we encounter in the (re-)born-digital corpora we have and lastly to proceed to the much larger, but far and far noisier OCR-digitized corpora. In this abstract we focus on the first steps, that of fruitfully linking morphologically related contemporary word forms. These linked word types are then further linked to their known older or erratic word variants in the other available humanly validated lexicons and word lists on the basis of the contemporary lemmata present. Largely undiscussed in this short paper is how we will then proceed to further link the thousands of as yet unknown real-world diachronic word forms and the millions of non-word variants present in Nederlab.

The TICCLAT database The TICCLAT database builds further on the historical lexical database structure[1] developed by our project partner INT (Institute for the Dutch Language) in the European project IMPACT[2], which we extended especially to add links between words, the core concept of our system. We ingested the Nederlab corpora[3], converted into

[1] The database design can be found at `https://github.com/TICCLAT/docs`. A link there leads to the INT documentation.

[2] `https://www.digitisation.eu/`

[3] `https://www.nederlab.nl/onderzoeksportaal/`

word-form frequency lists subdivided by year for each subcorpus. We doubt not that DH researchers can greatly benefit from gaining insight into the dispersion (Baayen, 1996) of word forms over sources, and hence over time, place or languages and their varieties.

Supervised Morphology Induction We started off with Combilex[4], a freely available INT Dutch language resource that for a sizeable part of the Dutch contemporary vocabulary provides the morphological related word forms for the lemmata. Given e.g. a noun, for which the lemma is the singular form, it may list the plural form and the singular and plural diminutive forms. For each of the six word pair combinations derivable from these four word forms we calculate the anagram value differences (Reynaert, 2011) and sum these into a single morphological numerical signature value. This signature is in fact a summary of the character differences observed between the word forms that constitute the particular word's morphological paradigm. The major point here is that all nouns that have the exact same character confusions between their four morphologically related forms, will directly obtain the exact same signature value. Nouns for which one or more of the theoretically possible word forms is not present in the list, for whatever reason, will obtain a signature which represents a subclass of the full paradigmatic signature. In so far as it is possible that say an adjective with three listed word forms sums up to the same signature as a noun with just two, we also calculate the sum for just the pairs formable from the set where one member is the lemma and use this second sum as a control on the first. On the basis of the combination of these two signatures we finally build a five segment labeling code, the segments simply being prefixed with the characters Z to V, followed by an (alpha)numerical code. We then label the paradigms accordingly, for all the word classes, starting off with the numerically largest signature as 'Paradigm 1' in our code 'Z0001', which is further labeled as 'Subclass 1' or 'Y00001' and onwards with all the subclasses observed, descending according to their ever smaller numerical signatures, with a incremented sequential subclass number. The labels hereby obtained are further elaborated with an X-segment which gives the rank of the signature over all signatures observed. The W-segment assigns a defining cluster number on the basis of each individual lemma, these being descendingly sorted on the basis of their frequency in the Dutch contemporary written Dutch corpus SoNaR-500 (Oostdijk et al., 2013). The V-segment code classifies and numbers the word forms in the cluster.

We supplemented the basic Combilex-based TICCLAT database with the following humanly validated lexicons: the very recently released new Modern Dutch Gigant-Molex[5], a prerelease of what is to be its diachronic counterpart we currently refer to as the INT Historical Dutch Lexicon, eLex[6], Open Taal Lexicon[7] and a list of 12,558 Dutch typo's collected for (Reynaert, 2005).

Conclusion We have built the links between word-forms that form the core asset of our system. In addition, we build visual interfaces to explore the TICCLAT database and hence unlock its potential for doing research into word variants of all imaginable kinds. This includes research into lexical evolution over time, into spelling variants through time and over regional languages, and in general into word dispersion over the corpora in any dimension present in their metadata. Another envisioned application is query expansion. We will assess the performance of tasks related to spelling correction and modernization.

[4] CombiLex (Version 1.0.1) (2014) INT Data set: `http://hdl.handle.net/10032/tm-a2-k2`

[5] Gigant-Molex (Version 1.0) INT data set: `http://hdl.handle.net/10032/tm-a2-p9`

[6] e-Lex (Version 1.1.1) INT data set: `http://hdl.handle.net/10032/tm-a2-h2`

[7] https://www.opentaal.org

References

Harald Baayen. 1996. The effects of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, 22:455–480, 12.

Hennie Brugman, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Antal van den Bosch. 2016. Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In Nicoletta Calzolari et al., editor, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*, pages 1277–1281, Portoroz, Slovenia. ELRA.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, chapter 13, pages 219–247. Springer Verlag.

Martin Reynaert. 2005. *Text-Induced Spelling Correction*. Ph.D. thesis, Tilburg University.

Martin Reynaert. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2):173–187.