Granularity versus Dispersion in the Dutch Diachronical Database of Lexical Frequencies TICCLAT

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

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

Abstract

The Nederlab project collected the digitized diachronical corpora of Dutch and made them available to all researchers in a single, explorable and exploitable portal within the CLARIN infrastructure. We are now building a database of lexical items and their frequencies collected according to the best known year of text production or publication on the basis of the 18.5 billion word tokens in the corpus. We here briefly discuss the corpus contents, major database design decisions we have taken, the tools we use and the approaches we take.

1 Introduction

We[1] have worked in 'spelling correction', very broadly put, for going on for two decades. The great paradox we see in non-words in texts, whether having been created by mistyping or misrecognition by some text digitization system or any other mishap, is that when they have been resolved to the real-word that 'should be there', they have in the large majority of cases been solved, once and for all. This was in fact the vision behind the work on spelling correction already performed by IBM researchers in the 1980s (Pollock and Zamora, 1984), who advocated 'absolute correction', i.e. if a known error is encountered: replace it by its correct form. In Reynaert (2005) we gave an example of the non-word 'onjections' that might variously have to be resolved to 'injections', perhaps given the context 'these painful *onjections', versus to 'objections' given the context 'her vehement *onjections'. At least for longer words, measured in numbers of characters, such ambiguities are in fact rather rare.

The above gives the main rationale behind the current project TICCLAT, which stands for 'Text-Induced Corpus Correction and Lexical Assessment Tool'. It is meant to help assess the validity of word forms encountered in Dutch diachronical text. Its databases, or selected subsets thereof, will assist OCR post-correction, but a great many more uses may easily be envisaged. We want to use the vocabulary present in what is today the largest finely preprocessed corpus of Dutch, actually a compilation of many corpora, collected in the prior project Nederlab (Brugman et al., 2016), to try and solve most of these non-words, once and for all. To resolve them and link them to their most likely real-word versions, then to have this database available online[2] to all comers, freely usable for whatever research purposes the community may find uses for. Apart from this, we hope to greatly enhance the historical lexica available to us with as many as possible of the historical spelling variants ever produced as reproduced in these subcorpora. Thirdly, we want to account for the morphological variants of words, starting with contemporary Dutch, gradually going back over time to at least the 13th. century.

In fact, the way we are proceeding is to start with the best contemporary resources we have, 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 we have in the Nederlab corpus. The current estimate is that only around 6% of the 18.5 billion word tokens of text in Nederlab is born-digital.

[1] Here: the first author.

[2] The TICCLAT database is to be hosted by Clarin Center Meertens Institute (`https://centres.clarin.eu/centre/23`)

2 From corpora to year-stamped frequency lists

Corpora Overview: the Nederlab corpora Table 1 gives an overview of the main Nederlab subcorpora ingested in TICCLAT.

Period	Corpus Title	Type	Size
13th. century	"Walewein ende Keye"	Book	S
13th. to 21st. cent.	DBNL: Digital Library of Dutch Literature	Books	L
14th. century	Corpus of 14th. Century Dutch by Van Reenen & Mulder	Acts	M
17th. century	Scholarly Correspondences (Geleerdenbrieven and Epistolarium)	Letters	M
1620-1640	Minutes of the States of Holland (by N. Stellingwerff and S. Schot)	Reports	M
1618 to 1700	KB or Dutch National Library Newspaper Collection (crowd-sourced by Nicoline van der Sijs, rekeyed)	Newspapers	L
1701 to 1940	KB or Dutch National Library Newspaper Collection (OCR)	Newspapers	H
1621-1700	Acta of the Particular Synodes of Southern Holland	Reports	M
1693-1701	Pieter van Dam's "A description of the Dutch East India Company (VOC)"	Book	M
17th. and 18th. cent.	"Prize Papers" aka: "Sailing Letters" (edited by N. van der Sijs)	Letters	M
1702-1720	Correspondences of Anthonie Heinsius	Letters	S
1780-1800	EDBO: Early Dutch Books Online	Books	L
1811-1831	Diaries of Willem de Clercq	Diaries	S
1814 - 2014	Dutch Acts of Parliament (Political Mashup version)	Reports	H
1874-1918	Diaries and Notes of Willem Hendrik de Beaufort	Diaries	S
19th. century	Vincent van Gogh – The Letters	Letters	S
1891-1947	Diaries of P.J.M. Aalberse	Diaries	S
1985-2005	STEVIN Written Dutch Reference Corpus SoNaR-500	30 genres	L

Table 1: Overview of the main subcorpora available in Nederlab.
Size estimates: S = Small, M = Medium, L = Large, H = Huge.

Method We want to account as best possible for as many of the lexical items as automatically as possible in terms of real-word versus non-word, contemporary real-word versus diachronical variant and name versus non-name. We first account for morphologically related word forms. These should give us a first handle on true word forms likely to be expected in a language, regardless of the time frame the particular language is inspected at. TICCL will next be used in line with new developments achieved since Reynaert (2011) for identifying historical variants and to account for non-word variants.

Granularity and Dispersion The main issue here is granularity. Our means to zoom in on the data is obviously to obtain frequency lists containing the lexical items and their frequencies. But, given e.g. the EDBO, should we get the overall corpus frequency ('how often does this word form occur in EDBO?'), the document frequency ('in how many EDBO books does this word form occur?'), its frequency for each book of the corpus, or even as the Hathi Trust provides for books in its collections, its frequency per page of each book in the corpus? What we currently opt for is yet another take: per subcorpus, as far as the metadata allows, we regard the documents originating in the same year (or range of years in case the exact year of text production or publication is not known) as belonging to each year's subsubcorpus and collect this subsubcorpus frequency for all its word types.

Having the corpus frequencies for the range of all the years of all subcorpora, we are then enabled to get a clear and humanly interpretable overview of the occurrence of a particular word form across time as well as across the corpora that make up the full Nederlab corpus. In this manner, the first link we have established between word forms in the TICCLAT database is that of their dispersion (Baayen, 1996) over time and over a range of diverse corpora. The term dispersion, "i.e. the degree to which occurrences of a word are distributed throughout a corpus evenly or unevenly/clumpily" is further qualified by Stefan Th. Gries in a new book chapter[3] as "one of the most crucial but at the same time underused basic statistical measures in corpus linguistics". Note that we have added the notion of subcorpora of the 'corpus', further subdivided into year-stamped further subsubcorpora, and that we wish to study and compare or contrast the distribution of 'words' over these. The main challenges we face are the result of the highly uneven quality of the text collections we work with, resulting in inordinate numbers of (non-)word types.

[3] Preprint at: `https://www.researchgate.net/publication/332120488_Analyzing_dispersion`

Impact of available metadata In this work we are at the mercy of the quality of the data and metadata available for each text and we encounter difficulties concerning this in e.g. the DBNL, which should by rights be considered one of the most valuable digital text collections extant for Dutch. The DBNL was largely built before (Moretti and Piazza, 2005) introduced the notion of Distant Reading. No usable distinction was made between e.g. contemporary commentaries and quoted historical text fragments. Also, the metadata available to us for at least this subcorpus is unsatisfactory to our purposes. Mediaeval works republished in recent years that have no mention of a text production year anywhere near the lifetime of the original author will definitely impact the reliability of our time-stamped word frequencies. Metadata is likely to remain problematic. Whoever compiles a corpus cares to record to the best of her abilities the best available metadata, given the goals of her project. Given the likely very divergent goals of later projects, given the availability of the particular corpus to the larger research community with different research interests, the available corpus metadata is unfortunately all too likely to be found wanting. We carry on regardless.

Means: the word frequency tools used We strive to get the most usable overview over the diachronical lexicon of Dutch as present in a wide and diverse range of word lists and digital text collections. These corpora are available to us in FoLiA XML and we have the tools to highly efficiently and in the best parallelized fashion process these (van Gompel et al., 2017). There are millions of files, however. So what we did was not bring the data to the tools, but rather bring the tools closer to the data. This is not yet what currently the Dutch CLARIAH-plus project works towards, i.e. infrastructure to bring the tools to the data, which can then remain in its proper repository and does not need to be copied and distributed, possibly leading to multiple copies differently (pre-)processed, transformed, linguistically enriched, etc. What we have now is a means of virtually reordering the otherwise stationary file collections for further processing, e.g. on the basis of metadata such as 'date of publication' or 'location' or any other criterion. We have extended two of the C++ TICCL modules[4] in order to extract the frequency lists from the Nederlab subcorpora in the best way possible.

FoLiA-stats derives frequency lists from corpora. FoLiA-stats – there is also a TICCL-stats which works on plain text – in its original working mode can recursively traverse directory trees, locating the files to be jointly transformed into word ngram frequency lists. To do what we want to, this implies we would have had to copy all the files with text originating in the same year to a single directory for one year, which on its own and separately from the other years, would have to be processed by FoLiA-stats. We brought the tool closer to the data by extending it so that it can work on the basis of a list containing the full directory paths and file names of the thousands, for some subcorpora: millions, of texts to be processed. A label in the second column instructs the program to create a directory with the same name and to collect the vocabulary contained in all the files bearing the same label in a single frequency file to be output to that directory. So, regardless of the actual whereabouts of the files regarding directories or even storage partitions, a single frequency list containing the ngrams observed in the texts that originated in a single year can be created, in parallel, for a range of years.

The second tool, **TICCL-unk**, is next enlisted to 'clean' the word types from the corpus frequency list compiled by FoLiA-stats. Especially in OCRed corpora, which are untokenized, character strings occur that are, heuristically, deemed unsalvageable, i.e. OCR garbage. These are disregarded. Character strings having 'word' initial or final punctuation are written to another file and shorn of this punctuation added to the main list where, if already present, their corpus frequency is added to that of the clean version. Clean word strings are naturally written to the 'clean' file and their frequencies tallied.

3 The TICCLAT database

Structure of the TICCLAT database and tools The structure of the TICCLAT database is squarely adopted from the historical lexical database structure our project partner INT[5] developed in the European project IMPACT[6]. We have extended it with a number of fields required for our purposes, including

[4] Available from: `https://github.com/LanguageMachines`

[5] Institute for the Dutch Language: `https://ivdnt.org/the-dutch-language-institute`

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

being able to specify links between wordforms. An overview of the database schema and link to the IMPACT document can be found in the documentation[7]. The **TICCLAT software**[8] is used for database management, and ingesting and querying the data.

On linking based on word forms' relatedness It is one thing to fill a huge database with hundreds of year-stamped frequency lists of subcorpora derived from both born-digital and OCR-digitized texts. It is quite another to organize these many millions of word forms in a sensible way to start seeing through the trees and make them useful, whatever the intended use. In TICCLAT, we link the many related variants of what might constitute a single 'word'. Each word type is assigned a unique code which links and identifies through its prefix the overall cluster of its related words, by an infix specific to each of the following three subcategories: the morphologically related word forms, next the word types related diachronically or that are possibly divergent but accepted word variants and, finally, the incredible diversity of related erratic word forms misrecognized by the digitization processes. Numerical suffixes identify the word clusters and each unique word form in the cluster. The supervised morphology induction system we have developed to derive these codes in itself warrants a full paper to discuss the related work and provide a full evaluation.

4 Conclusions

What we have sketched is in fact a huge undertaking. Were it not that we actually have control over the granularity of both setting the timeline of the subsubcorpora we ingested in the database and that of the subcorpora we choose to TICCL in order to extract the lexical variants from and best-first rank these by, we might not achieve anything noteworthy in the limited time frame of the TICCLAT project. We are setting up the infrastructure for demonstrating on a sufficiently large scale that the ultimate goal, given the necessary resources, is achievable.

Acknowledgements

The authors acknowledge being funded in the call 'Accelerating Scientific Discovery in the Arts and Humanities' (ADAH) issued jointly by CLARIAH and the Netherlands eScience Center.

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.

F. Moretti and A. Piazza. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Joseph J. Pollock and Antonio Zamora. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM*, 27(4):358–368.

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.

Maarten van Gompel, Ko van der Sloot, Martin Reynaert, and Antal van den Bosch. 2017. FoLiA in practice: The infrastructure of a linguistic annotation format. In J. Odijk and A. van Hessen, editors, *CLARIN-NL in the Low Countries*, chapter 6, pages 71–81. Ubiquity (Open Access).

[7] https://github.com/TICCLAT/docs/blob/master/database_design.md

[8] https://github.com/TICCLAT/ticclat