Searching and analyzing large annotated text collections in Nederlab

Hennie Brugman
Meertens Institute, Amsterdam, Netherlands
hennie.brugman@meertens.knaw.nl

Nederlab[1] is a five year project that started at the beginning of 2013 with the ambition to collect the full production of published Dutch texts from about 800 until present. Its main purpose is to present this text material in a homogeneous way through a virtual research environment aimed at primarily historians, literary scholars and linguists. Most important metadata dimensions to make this collection accessible are author information, dates, locations and genres. An important aspect of Nederlab is interlinking of objects within and between sub collections, thus providing a Nederlab thesaurus of authors and titles.

A rough estimate indicates that the full body of relevant texts comes down to a collection of the order of magnitude of 10-100 billion words. This implies that all of our back end technology has to scale very well.

Much of the added value that the Nederlab collection ingest pipeline creates, next to careful editorial curation of the metadata, stems from text annotations that are mostly automatically created. Wherever the quality of the base texts allows it we add the following layers of annotation during a preprocessing phase: an additional text layer that contains normalized and spell-corrected text (using TiCCL) (Reynaert 2014) lemma, part of speech (including several types of sub features), entities (the last three created using frog[2]), and a word by word automatic translation to modern wordforms, to help apply language tools that are built for modern Dutch to older texts (Tjong Kim Sang 2015). We intend to add a layer with entity linking information: for entities in the texts it contains references to external knowledge bases such as for example Wikipedia or person or location thesauri. Another layer will contain references to lemmata in the online historical lexicon of INL (the Institute for Dutch Lexicology). Finally, annotation layers that originate from processing of the original texts are: text segmentations such as paragraphs and sentences, chapters and chapter titles, editorial matter (vs main text), language used (at paragraph level). Some of the Nederlab use cases require hierarchical annotation structures, such as morphology or syntactic trees. A number of use case-specific custom annotation layers may be added as well. This means that it will not be exceptional to have text segments with 15 or more associated annotation layers.

[1] www.nederlab.nl

[2] http://languagemachines.github.io/frog/

Nederlab's scientific use cases require first of all that the search back end is able to retrieve complex sequential patterns over multiple layers, minimally as are covered by the Corpus Query Language (CQL) (Christ 1994). Additionally required are, directly available as part of the research environment or needed to feed data into special analytic tools:

- statistics overviews over sub collections or over arbitrary sets of query results: absolute and relative number of hits, number of words in all documents that contain hits, minimum, maximum and mean number of hits per document and statistical spread, etc. It is also important that multi-word hits (e.g. multi-word named entities like 'Den Haag') are counted correctly.

- keyword-in-context representations of search results: show hits with an adjustable left and right context and with all associated annotation layers.

- grouping of results over metadata dimensions, hit value or hit context: list all unique occurring hit values, accompanied by the number of times each value occurs.

- distributions of results over metadata dimensions and time intervals. We also want to be able to retrieve such distributions over multiple dimensions simultaneously and visualize them. Examples of this are distributions over genres and gender of the author, or over gender and year of publication.

- frequency lists over wordforms or annotation values, both for individual documents and random subcollections of Nederlab. These frequency lists can be used as input for more high level analytical tools, for example a tool to compare different subcollections of Nederlab to each other, or a collocation tool.

- multi-dimensional data arrays to export (numeric) results to often used analytical packages such as R

- import of all annotation types and structures that are supported by the FoLiA XML format (van Gompel 2013) and export of results in FoLiA format

- annotated text search should be seamlessly integrated with search and refinement over metadata, both types of searches can be applied in any order

From the above it is clear that the combination of scale, complexity of annotation structures and scientific requirements with respect to search and analysis form a serious technical challenge for the Nederlab back end infrastructure. We have to make sure that the system is scalable (for example by supporting parallel search indexes) and is easily manageable (by allowing easy and modular updating of the indexes). Substantial work has been done on this type of infrastructures in other projects. There are a number of systems available, each with its own strengths, weaknesses and track records, e.g. Blacklab[3], Corpus Workbench[4], Korap[5] and MTAS (Multi Tier Annotation Search, Meertens Institute).

[3] http://inl.github.io/BlackLab/

In the Nederlab project we have experience using both BlackLab and MTAS. Our current implementation (based on MTAS) already meets most of the requirements discussed, for a collection of 1.5 billion annotated words.

References

Christ, O. (1994) 'A Modular and Flexible Architecture for an Integrated Corpus Query System', In *Proceedings of COMPLEX '94: 3rd Conference on Computational Lexicography and Text Research*, Budapest.

Gompel, M. van, and Reynaert, M. (2013) 'FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study', *Computational Linguistics in the Netherlands Journal*, 3.

Martin Reynaert. (2014) 'Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up', In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. ELRA.

Erik Tjong Kim Sang. (2015) 'Converting seventeenth century Dutch to modern Dutch', *workshop Morfosyntactisch verrijken van historische teksten*, Utrecht, 16 november 2015.

[4] http://www.ims.uni-stuttgart.de/forschung/projekte/CorpusWorkbench.html

[5] http://korap.ids-mannheim.de/