# T036: Improved infrastructure for Older Dutch

**IDs in Google Sheets:** T036. T125, T126 (\*)

**Author:** INT

**What exists / What must be adapted / extended / created anew:**

For the detailed description cf. <https://github.com/INL/clariah-plus-planning/blob/master/Aviertjes/PDF/Improved_Infrastructure_for_Historical_Dutch.pdf>

We quote the abstract here:

High quality linguistic annotation of historical Dutch is still problematic.This is clearly exemplified by a project like Nederlab covering Dutch language from the 6th- 21st century. Even though there are by now several tools that handle the problem,

– Efforts are fragmented

– Training material is scarce, absent for some periods and diverse in adopted tagset and annotation guidelines

– There is no common strategy to deal with typical issues of historical Dutch

- nontrivial word segmentation, spelling variation, and inflectional features which are hard to classify

There is need for a more thorough approach, comprising the following tasks:

– Develop a common tag set and annotation guidelines

– Reliable metadata for historical corpus data

– Harmonize available training data and historical lexica

– Extend training data where the gaps are most painful

– Optimization and minor adaptation of existing taggers

– Integrate tools in a workflow for corpus processing

– Create a community and a shared infrastructure to enable researchers to easily upload, annotate and correct their data

Most tasks belong to WP3; The last task is WP6

**Targeted/Actual users:** Digital humanities, historical linguists, historians  
**Actual use (quantify!):**

**Social Impact** **(concrete examples):** All types of research involving historical language is still hindered by the complexities of historical Dutch. A good infrastructure will benefit many researchers

**Estimate in PMs (try to justify):** Cf. detailed planning

**Lead + PMs:** INT

**Participants + PMs: (**INT personnel)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Taak** | **Type personeel** | **Available** | **2019** | **2020** | **2021** | **2022** | **2023** |  | **Summed** | **Task total** |
| T036 Older Dutch | Linguist 1 | 4 | 1 | 2 | 1 |  |  |  | **4** | 14 |
| T036 Older Dutch | Linguist 2 | 4 |  | 1 | 1 | 1 | 1 |  | **4** |  |
| T036 Older Dutch | Software developer | 3 | 1 | 1 | 1 |  |  |  | **3** |  |
| T036 Older Dutch | Computational linguist | 3 | 1 | 1 | 1 |  |  |  | **3** |  |

**Deliverables:**

|  |  |  |  |
| --- | --- | --- | --- |
| ID | D or M | Type | Description |
| T036D0a | Deliverable | Document | Tagging guidelines  Diachronic linguistic research based on linguistically annotated texts would be better facilitated if we had a tagset applicable to all language stages of historical Dutch. The document contains a proposal for the Tagset for Diachronous corpus material of Dutch (TDN).  For the determination of this tagset we have based ourselves on the tagset of GiGaNT (Groot Geïntegreerd lexicon van de Nederlandse Taal), the corpus tagset CGN/D-Coi, the tagset Corpus Gysseling, the tagset Corpus Van Reenen-Mulder, the tagset of Corpus Oudnederlands, the Nederlab project, the linguistc annotations of letters by P. C. Hooft and the historical dictionaries: the Old Dutch Dictionary (ONW) 500-1200, the Early Middle Dutch Dictionary (VMNW) 1200-1300, the Middle Dutch Dictionary (MNW) 1250-1550 and the Dutch Language Dictionary (WNT) 1500-1976.  We tested the proposals by performing a complete mapping of the CGN/D-Coi tags, the tags in Corpus Gysseling and the tags in Corpus Oudnederlands to TDN. |
| T036D0b | Deliverable | Document | Corpus encoding  This document contains recommendations for linguistic annotation of historical texts in CLARIN-standardized formats.  Our first target is a proposal for linguistic annotation in TEI P5. Further objectives are proposals for FoLiA and NAF. |
| T036D1 | Deliverable | Document | Evaluation of linguistic annotation for historical Dutch  The document contains evaluation results on ground truth data for state-of-the art systems tackling linguistic annotation of historical Dutch |
| T036D2a | Deliverable | data | Training and evaluation data  Gold standard linguistic annotated corpus material for Dutch, 1200-~1970.  We aim for about 100000 tokens per century. |
| T036D2b | Deliverable | data | Proof of concept corpus  Linguistically annotated corpus, using the best available automatic tagging system for the material at hand.  The corpus composition will be determined by use cases, a.o. those emerging from WP6. |
| T036D3 | Deliverable | Software | Tagging services/applications  Web application (code named 'GALAHAD') for historical enrichment of historical Dutch, designed to enable the user to chose the most fitting option from available systems.  The system is designed in such a way that models/systems can easily be added by proving a docker container satisfying a specified interface. |
| T036D4 | Deliverable | Software | Web-based environment for manual correction of linguistic annotation |
| T036D5 | Deliverable | Website | Website with demonstrator tools and documentation |

**PMs etcetera**