# Topic: Improved infrastructure for Older Dutch

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

**Author:** INT

**What exists:** Individual tools and resources for processing of “Older” Dutch:

Taggers etc:

* Adelheid (Hans van Halteren)
* Frog
* Midas (Mike Kestemont)
* INL labs
* Nederlab enrichment tools

Annotation tools

* FLAT
* COBALT

Resources

* GiGaNT lexicon
* Training data (Corpus Gysseling, CRM, Brieven als Buit, small datasets developed in Nederlab and IMPACT, …)

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

Work on an infrastructure with the necessary tools, lexica and training data to deal with historical language material. The infrastructure should be beneficial to both tool developers and non-expert users/researchers, who would like to annotate their own material.

In DH projects, a lot of research is based on historical corpus material. This research could benefit from good quality linguistic annotation. Linguistic annotation of historical language data however still remains a challenge. Work has been done in several projects, national and international (e.g. IMPACT, Brieven als Buit, and recently Nederlab), but results are fragmented and far from providing a complete solution. There are several approaches to linguistic annotation of historical material and tools which are valuable enough to develop further. What all approaches need is golden standard material to train, develop and evaluate on. Not only is there not enough training material. Currently available training material for various tasks (NER tagging, PoS tagging) is fragmented and should be harmonized for optimal benefit. We would like to develop an infrastructure which provides this necessary material and an environment in which tool developers could make their tool available to potentially function in a tagging service.

For the non-technical researchers, the infrastructure should offer a user friendly environment which offers a means to upload their data and choose a tool for automatic annotation. The environment should also offer potential users options to assess the quality of the offered tools. There should also be a tool (service) by means of which efficient correction of linguistically annotated corpus material can be done.

INT already makes a historical lexicon available (GiGaNT) in the Lexicon Service.

We would also like to build further on work done for:

INL-LABS: (demonstrator):http://inl-labs.taalbanknederlands.inl.nl/succeed/tagger/ui (linguistic annotation (lemma; PoS; NER);

OPEN CONVERT http://openconvert.clarin.inl.nl/openconvert/tagger/ui#file (text conversion).

COBALT: https://github.com/INL/COBALT (an application in which a corpus of texts can be loaded so as to be able to annotate its tokens (lemmatize and more). The annotation work in CoBaLT gives two products: an annotated corpus, and a lexicon consisting of word forms and their corresponding lemmata and such assigned to them).

For already existing Golden Standards for historical language material, e.g. see: <http://www.ivdnt.org/taalmaterialen>.

(\*)There is a relation with the tasks mentioned by VU (T20, 21 and 22), but here the focus would be on linguistic annotation of historical text material.

**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 greatly benefit many researchers

**Estimate in PMs (try to justify):** 12PM

Integration and harmonization of training material: 8PM

Retraining and deployment of tools: 4PM

Interfaces for manual annotation: 6PM (Cobalt)

**Lead + PMs:** INT

**Participants + PMs:** INT/RU