Improved Infrastructure for Historical Dutch

INT

July 1, 2021

Abstract

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: We propose tasks to

- 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

1 Introduction

We propose to 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-technical 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 gold 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.

Tools and annotated data will benefit the maintenance and improvement of the Nederlab infrastructure.

1.1 Impact

Targeted/Actual users Digital humanities, historical linguists, historians, any digital humanities scholar working with (historical) text.

Actual use (quantify!) It is difficult to quantify the use of historical annotation tools. One needs to distinguish the smaller group of corpus builders from the arger group of corpus users. Both benefit from an improved infrastructure, directly and indirectly. Obviously,

- There are many researchers using historical text
- The amount of corpus material that is processed by these tools is huge
- Easy accessibility to tools may increase the number of corpus builders

Social Impact All types of research involving historical language is still hindered by the complexities of historical Dutch. A good infrastructure will greatly benefit many researchers. (WP6 part:) By having the opportunity to involve digital humanities scholars in the development of user interfaces of tools from the linguistic community, the tools will become more accessible, also for the individual scholar with no particular technical background and with little or no technical support.

1.2 What is available

(this list is not complete; it shows that substantial work is available to build on, but harmonization and integration is necessary)

1.2.1 Tools for automatic linguistic annotation

- Adelheid[1], Developed by Hans van Halteren. Adelheid is already available in a CLAM-based web service for CLARIN-authenticated users at http://corpus1.mpi.nl/adelheid/main/
- Midas[2], developed by Mike Kestemont, now maintained by a.o. Enrique Manjavacas.
- INL labs, a supervised tagger trained on the Letters as Loot corpus[3]
- Frog middle Dutch (trained on CRM and Gysseling). Frog has recently been extended to deal with non-CGN tagsets [4]
- Nederlab spelling normalization, cf. a.o. [5]

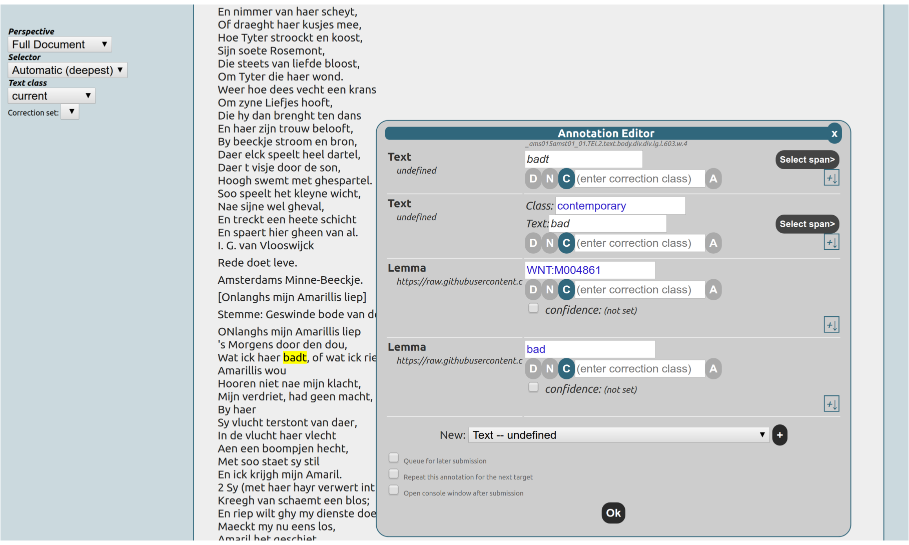
1.2.2 Tools for manual annotation / correction of automatic annotation

- FlaT[6]
- COBALT[7]
- Gustave[8]
- Adelheid also comes with a manual correction tool
- Generic annotation tools like Brat[9]

1.2.3 Training and evaluation data

- Corpus Oud-Nederlands
- Corpus Gysseling
- Corpus van Reenen-Mulder
- Extensions to CRM

[1] http://adelheid.ruhosting.nl/
[2] https://github.com/mikekestemont/Midas
[3] http://inl-labs.inl.nl/
[4] https://github.com/LanguageMachines/frog/issues/53
[5] https://ifarm.nl/clin2017st/paper/final.pdf
[6] https://flat.science.ru.nl/
[7] https://github.com/INL/COBALT
[8] https://languagedynamics.wp.hum.uu.nl/projects/
[9] http://brat.nlplab.org/



Perspective: Full Document

Selector: Automatic (deepest)

Text class: current

Correction set:

En nimmer van haer scheyt,
Of draeght haer kusjes mee,
Hoe Tyter stroockt en koost,
Sijn soete Rosemont,
Die steets van liefde bloost,
Om Tyter die haer wond.
Weer hoe dees vecht een krans
Om zyne Liefjes hooft,
Die hy dan brenght ten dans
En haer zijn trouw belooft,
By beeckje stroom en bron,
Daer elck speelt heel dartel,
Daer t visje door de son,
Hoogh swemt met ghespartel.
Soo speelt het kleyne wicht,
Nae sijne wel gheval,
En treckt een heete schicht
En spaert hier gheen van al.
I. G. van Vlooswijck

Rede doet leve.

Amsterdams Minne-Beeckje.

[Onlanghs mijn Amarillis liep]

Stemme: Geswinde bode van d

ONlanghs mijn Amarillis liep
's Morgens door den dou,
Wat ick haer badt, of wat ick rie
Amarillis wou
Hooren niet nae mijn klacht,
Mijn verdriet, had geen macht,
By haer
Sy vlucht terstont van daer,
In de vlucht haer vlecht
Aen een boompjen hecht,
Met soo staet sy stil
En ick krijgh mijn Amaril.
2 Sy (met haer hayr verwert int
Kreegh van schaemt een blos;
En riep wilt ghy my dienste doe
Maeckt my nu eens los,
Amaril het geschiet

Annotation Editor

_ams015amst01_01.TEI.2.text.body.div.div.lg.l.603.w.4

Text
undefined
badt
(enter correction class)

Text
undefined
Class: contemporary
Text: bad
(enter correction class)

Lemma
https://raw.githubusercontent.c
WNT:M004861
(enter correction class)
□ confidence: (not set)

Lemma
https://raw.githubusercontent.c
bad
(enter correction class)
□ confidence: (not set)

New: Text -- undefined

□ Queue for later submission
□ Repeat this annotation for the next target
□ Open console window after submission

Figure 1: The FoLiA Linguistic Annotation Tool.

  - KANTL southern extensions
  - 15th century extensions (Margit?)

- Letters as Loot
- Gentse Spelen
- Hooft letters[10]
- Nederlab evaluation sets
- Corpora compiled by Rik Vosters and colleagues at VU Brussel (early 19th century legal transcripts)

1.2.4 Lexica

- GiGaNT historical lexicon and web service

1.2.5 Tagsets and annotation guidelines

- Tagset Gysseling/CRM
- Extended subset of CGN tagset to incorporate features of Gysseling/CRM
- Main PoS according to INT lexica (Letters as Loot, Gentse Spelen)
- Extended CGN subset developed for the Hooft Letters
- Reduced CGN tagset in Nederlab Evaluation Samples
- GiGaNT tagset[11] and lemmatization principles[12]

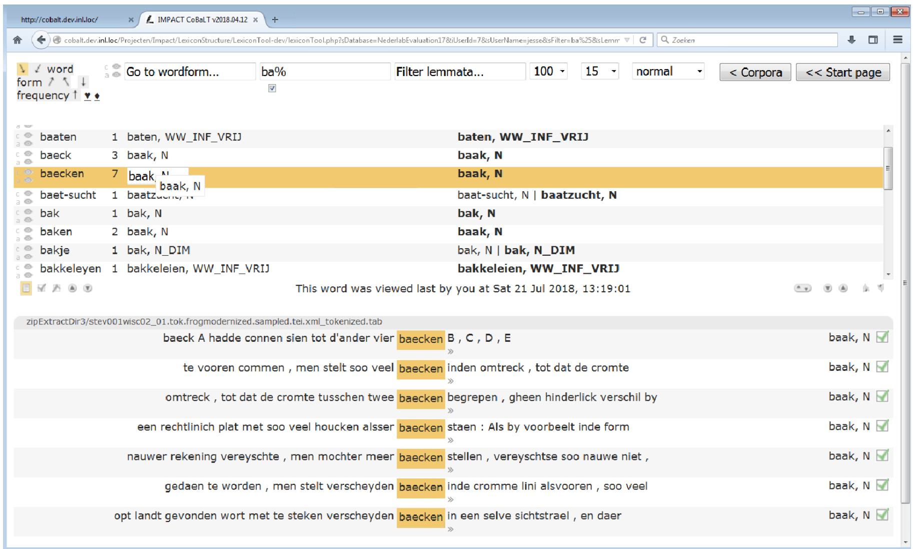
1.2.6 Workflow systems

- The PICLL-based Nederlab corpus processing workflow

[10] https://languagedynamics.wp.hum.uu.nl/projects/

[11] https://www.ivdnt.org/images/stories/onderzoek_en_onderwijs/publicaties/TaalbankWorkingpaper3.pdf

[12] https://www.ivdnt.org/images/stories/onderzoek_en_onderwijs/publicaties/TaalbankWorkingpaper4.pdf



cobalt.dev.inl.loc/Projecten/Impact/LexiconStructure/LexiconTool-dev/lexiconTool.php?sDatabase=NederlabEvaluation17&iUserId=7&sUserName=jesse&sFilter=ba%25&sLemm

Go to wordform... | ba% ☑ | Filter lemmata... | 100 | 15 | normal

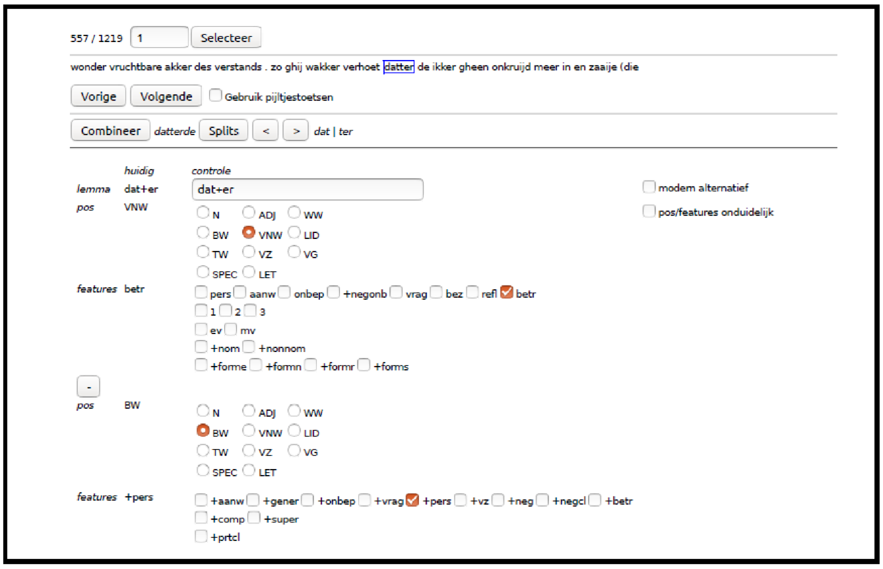
baaten	1	baten, WW_INF_VRIJ	**baten, WW_INF_VRIJ**
baeck	3	baak, N	**baak, N**
baecken	7	baak, N baak, N	**baak, N**
baet-sucht	1	baatzucht, N	baat-sucht, N | **baatzucht, N**
bak	1	bak, N	**bak, N**
baken	2	baak, N	**baak, N**
bakje	1	bak, N_DIM	bak, N | **bak, N_DIM**
bakkeleyen	1	bakkeleien, WW_INF_VRIJ	**bakkeleien, WW_INF_VRIJ**

This word was viewed last by you at Sat 21 Jul 2018, 13:19:01

zipExtractDir3/stev001wisc02_01.tok.frogmodernized.sampled.tei.xml_tokenized.tab

baeck A hadde connen sien tot d'ander vier	baecken	B , C , D , E »	baak, N ✓
te vooren commen , men stelt soo veel	baecken	inden omtreck , tot dat de cromte »	baak, N ✓
omtreck , tot dat de cromte tusschen twee	baecken	begrepen , gheen hinderlick verschil by »	baak, N ✓
een rechtlinich plat met soo veel houcken alsser	baecken	staen : Als by voorbeelt inde form »	baak, N ✓
nauwer rekening vereyschte , men mochter meer	baecken	stellen , vereyschtse soo nauwe niet , »	baak, N ✓
gedaen te worden , men stelt verscheyden	baecken	inde cromme lini alsvooren , soo veel »	baak, N ✓
opt landt gevonden wort met te steken verscheyden	baecken	in een selve sichtstrael , en daer »	baak, N ✓

Figure 2: The CoBaLT Lexicon Annotation Tool.



557 / 1219 | 1

wonder vruchtbare akker des verstands . zo ghij wakker verhoet datter de ikker gheen onkruijd meer in en zaaije (die

□ Gebruik pijltjestoetsen

datterde | dat | ter

	huidig	controle
lemma	dat+er	dat+er
pos	VNW	○ N ○ ADJ ○ WW ○ BW ● VNW ○ LID ○ TW ○ VZ ○ VG ○ SPEC ○ LET
features	betr	□ pers □ aanw □ onbep □ +negonb □ vrag □ bez □ refl ☑ betr □ 1 □ 2 □ 3 □ ev □ mv □ +nom □ +nonnom □ +forme □ +formn □ +formr □ +forms
pos	BW	○ N ○ ADJ ○ WW ● BW ○ VNW ○ LID ○ TW ○ VZ ○ VG ○ SPEC ○ LET
features	+pers	□ +aanw □ +gener □ +onbep □ +vrag ☑ +pers □ +vz □ +neg □ +negcl □ +betr □ +comp □ +super □ +prtcl

□ modern alternatief

□ pos/features onduidelijk

Figure 3: The Gustave Annotation Tool.

1.3 What must be created

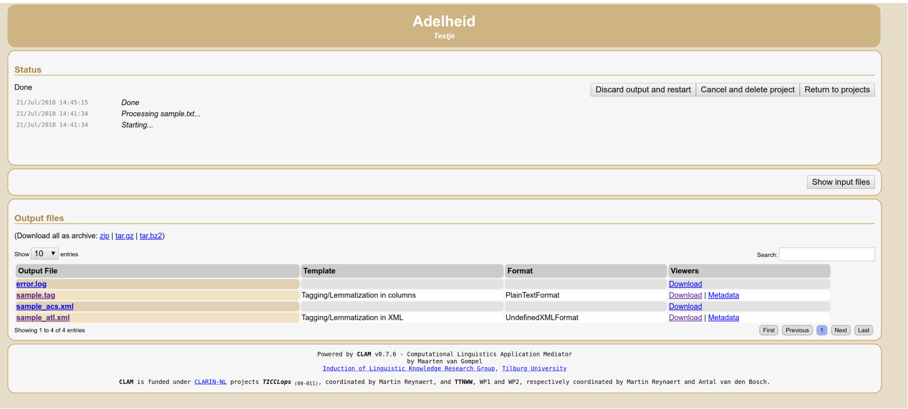
While the most obvious thing to do is to sit down, take a deep breath and start annotating as much training data as possible, we think that history has shown that fragmentation will continue without an infrastructure and "community building" effort. (hence the involvement of research partners).

We propose tasks to

1. Develop a common tag set and annotation guidelines, such that

   – It allows different levels of linguistic detail to enable projects with different aims to profit from each others efforts – the tag set and guidelines should cover both extremely coarse-grained tagging (main PoS and lemma) and detailed annotation up the the annotation of difficult features like case[13].
   – It includes a standardized approach to tokenization issues in corpus (XML) file formats. Neither TEI nor FoLiA currently offer a completely satisfactory solution. For both

[13] Currently, only the Old Dutch corpus has case. This is obviously a significant gap, as is the (almost, cf. InPolder, [14]) absence of syntactically annotated material



Adelheid

Testje

Status

Done

```
21/Jul/2018 14:45:15    Done
21/Jul/2018 14:41:34    Processing sample.txt...
21/Jul/2018 14:41:34    Starting...
```

Output files

(Download all as archive: zip | tar.gz | tar.bz2)

Show 10 entries Search:

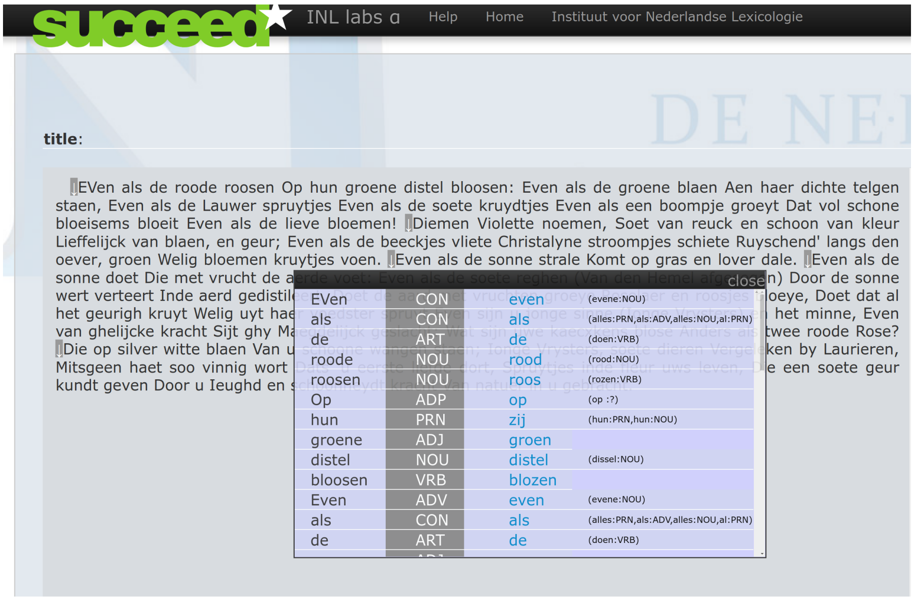
Output File	Template	Format	Viewers
error.log			Download
sample.tag	Tagging/Lemmatization in columns	PlainTextFormat	Download | Metadata
sample_acs.xml			Download
sample_atl.xml	Tagging/Lemmatization in XML	UndefinedXMLFormat	Download | Metadata

Showing 1 to 4 of 4 entries

Powered by **CLAM** v0.7.6 - Computational Linguistics Application Mediator
by Maarten van Gompel
Induction of Linguistic Knowledge Research Group, Tilburg University

CLAM is funded under CLARIN-NL projects ***TICCLops*** (09-011), coordinated by Martin Reynaert, and **TTNWW**, WP1 and WP2, respectively coordinated by Martin Reynaert and Antal van den Bosch.

Figure 4: The Adelheid tagging service.



title:

EVen als de roode roosen Op hun groene distel bloosen: Even als de groene blaen Aen haer dichte telgen staen, Even als de Lauwer spruytjes Even als de soete kruydtjes Even als een boompje groeyt Dat vol schone bloeisems bloeit Even als de lieve bloemen! Diemen Violette noemen, Soet van reuck en schoon van kleur Lieffelijck van blaen, en geur; Even als de beeckjes vliete Christalyne stroompjes schiete Ruyschend' langs den oever, groen Welig bloemen kruytjes voen. Even als de sonne strale Komt op gras en lover dale. Even als de sonne doet Die met vrucht de aerde voet: Even als de soete reghen (Van den Hemel afgegeven) Door de sonne wert verteert Inde aerd gedistileert; Doet de aarde met vruchten groeye Rooslaer en roosjes bloeye, Doet dat al het geurigh kruyt Welig uyt haer voedster spruyt Even sijn u jonge sinne (Ionge Vrysters) en het minne, Even van ghelijcke kracht Sijt ghy Maegdelijck geslacht: Wat sijn uwe kaecxkens blose Anders als twee roode Rose? Die op silver witte blaen Van u schoone wangen staen; Ionge Vrysters, soete dieren Vergeleken by Laurieren, Mitsgeen haet soo vinnig wort Dats u eerste lierde dort, Spruytjes inde fleur uws leven, Die een soete geur kundt geven Door u Ieughd en schoonheydt kracht Van natuer in u gebracht.

EVen	CON	even	(evene:NOU)
als	CON	als	(alles:PRN,als:ADV,alles:NOU,al:PRN)
de	ART	de	(doen:VRB)
roode	NOU	rood	(rood:NOU)
roosen	NOU	roos	(rozen:VRB)
Op	ADP	op	(op :?)
hun	PRN	zij	(hun:PRN,hun:NOU)
groene	ADJ	groen	
distel	NOU	distel	(dissel:NOU)
bloosen	VRB	blozen	
Even	ADV	even	(evene:NOU)
als	CON	als	(alles:PRN,als:ADV,alles:NOU,al:PRN)
de	ART	de	(doen:VRB)

Figure 5: The INL labs demonstrator.

issues we may build on work done in the context of Nederlab to harmonize CGN-based and historical tagging, and to elaborate TEI- and FoLiA based approaches to encoding historical "splits" (*te rugh*) and "merges" (*saghse*). Cf. for instance [Cou] for a discussion of tag set issues.

- Mapping to other tagsets, such as universal dependencies
- Adequate metadata for historical corpus data. This involves solid principles enabling a distinction between dating of "text", text witness and publication date, as well as assigning separate metadata to parts of documents which may be .e.g. of different date (In Nederlab, both a proposed FoLiA and TEI approach have been developed)

2. Harmonize available training data and historical lexica. It should be clear that this involves more than a mapping of tags; extensive manual work may be required to achieve agreement
3. Extend training data where the gaps are most painful (as shown by an extensive evaluation). There are obvious gaps in most periods.

   - 14th century: no literary material
   - 15th-19th century: very little for "standard" written text

4. We do not propose to build completely new taggers/lemmatizers. Instead we will concentrate on optimization and adapting existing annotation tools, both by retraining and possible tool maintenance where needed and feasible. Also, it should be determined by an extensive

Century	rate (tokens/hour)	tokens	hours	months
1350-1550 (Middle Dutch)	150	100,000	666	5
1550-1650 (Early New Dutch)	200	100,000	500	4
1650-1800 (New Dutch 1)	400	100,000	250	2
1800-1954 (New Dutch 2)	500	100,000	200	1.5
total		400,000	1616.7	12.5

Table 1: Estimated workload for training data production according to the periodization adopted in Nederlab. Please note that sharing of training material between centuries is possible, and that with the existing corpora, we will have around 3 million annotated tokens

evaluation which combination of tool and training data is best suitable for a certain type of material

5. Integrate tools in a workflow for corpus processing robust enough to handle large volumes of data and sensitive enough to select the right tools for the right data
6. Test the workflow on a considerable amount of data
7. Create a community and a shared infrastructure to enable researchers to easily upload, annotate and correct their data [Task for WP6].

2 Work plan

2.1 Technical plan

2.1.1 Evaluation and testing infrastructure

This is the first task to be tackled. Involves both web service testing and UI testing.

2.1.2 Deep learning domain adaptation methods

Due to recent advances in deep learning and more specifically domain adaptation[HE19], it is not unlikely that off-the-shelf deep learning methods will outperform dedicated tools. Furthermore, the existing tools for historical Dutch do not use domain adaptation, which makes them inefficient in terms of training data effort.

For a straightforward baseline application of this type of approach, we need

- A generic trained BERT or other context-sensitive embedding models[15]. Candidate corpora are

  - SoNaR (500M tokens) model built by LIACS, Susan Verberne)
  - CHN-large (1.2G tokens)
  - DBNL (historical, about 1G tokens)

  Since training such models is resource-intensive, we will need access to a cluster.
- Scripts for finetuning and task adaptation e.g. https://github.com/xhan77/AdaptaBERT. Note that fine-tuning is feasible on one GPU.

2.2 Workload for manual tagging

Usually, 300-500 tokens per hour, depending on expertise of the tagger and difficulty of the material, is to expected for historical material. It should be taken into account that for difficult material, the rate may slow down to slightly above 100 tokens per hour. A preliminary estimate is in table 2.2

[15] We should not concentrate on one particular type of embedding, XLNet (https://arxiv.org/abs/1906.08237) already appears to outperform BERT on a set of tasks. Training effort is so huge that it does not seem a practical option for now, but that might change

2.3 Involved Partners

- RU: Piccl, Frog, Corpus encoding, Adelheid, user interface (FLaT, Piccl)
- RU (Margit Rem): Tagging principles
- INT: Tagging guidelines, tagset, INL labs, tokenization guidelines, corpus encoding, user interface, metadata
- Leiden University (Sjef Barbiers) Tagging principles
- Utrecht (Marjo van Koppen) Tagging principles, annotation tool (Gustave)
- Meertens (Hennie Brugman) relation to Nederlab
- Karina van Dalen-Oskam (Huygens) superuser and user interface concerns

External collaboration will probably involve Mike Kestemont and Enrique Manjavacas.

2.4 Deliverables

ID	Description	Who	PM	WP
D0a	Tagging and lemmatization guidelines	INT, Leiden, RU, Utrecht	INT 3, Leiden, RU, Utrecht each 1	3
D0b	XML corpus encoding for historical corpora	INT (TEI), RU (FoLiA)	2	3
D1	Evaluation of automatic annotation tools per period and type of material	INT, RU	4	3
D2a	Training data development	INT	12	3
D2b	Automatically tagged proof of concept diachronic corpus, consisting of a substantial reliably metadated, non-OCR subset of Nederlab	INT	2	3
D3	Workflow, web service and web application(s) for historical tagging incorporating most essential tools and appropriate choice mechanism per period; Training of tools	RU, INT	4	3
D4	Web-based environment(s) for manual corpus annotation	INT, Huygens, RU	10	6
D5	Web site with demonstrator tools and extensive documentation	INT, RU, All	2	6

References

[Cou] Evie Coussé. Een digitaal compilatiecorpus historisch nederlands. *Lexikos*, 20:123–142.

[HE19] Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. *arXiv preprint arXiv:1904.02817*, 2019.