CLARIN-NL CALL 1 Proposal

1 Project Title, Acronym and Abstract

Title: A Distributed Lemmatizer for Historical Dutch

Acronym: Adelheid

Abstract: This project aims at providing a web-application with which an end user can have historical Dutch text tokenized, lemmatized and part-of-speech tagged, using the most appropriate resources (such as lexica) for the text in question. The need to consistently use appropriate resources leads to the intuitively obvious strategy of placing this service in the Clarin infrastructure. For each specific text, the user can then select the best resources from those available in Clarin, wherever they might reside, and where necessary supplemented by own lexica. During the project a demonstrator for the distributed automatic lemmatization will be created, with some 14th century charters as test texts as well as corresponding resources.

Target Start Date: January 1, 2010
Target End Date: June 30, 2010
Type: Demonstrator Project

2 Coordinator

Name: Dr. Hans van Halteren
Function: Language Technology Expert
Organization: Radboud University Nijmegen
Address: Centre for Language and Speech Technology, P.O. Box 9103, 6500 HD Nijmegen
E-mail: hvh@let.ru.nl
Tel: 024 361 2836
Fax: 024 361 2907
Role: Technology Provider & User

3 Composition of the Research Team

Name: Dr. Margit Rem
Function: Historical Linguist
Organization: Radboud University Nijmegen
Address: Dept. of Dutch, P.O. Box 9103, 6500 HD Nijmegen
E-mail: M.Rem@let.ru.nl
Tel: 024 361 2899
Fax: 024 361 2907
Role: Data Provider & User

Name: Daan Broeder
Function: Head MPI Clarin Centre
Organization: Max-Planck Institute for Psycholinguistics
Address: P.O. Box 310, 6500 AH Nijmegen
E-mail: daan.broeder@mpi.nl
Tel: 024 352 1103
Fax: 024 352 1213
Role: Clarin centre representative

Name: NN, to be hired
Function: Web Application Programmer

Role: Technology Provider

Name: IS, to be assigned by Clarin
Role: Infrastructure Specialist

4 Clarin centre

Given that this project is to demonstrate the ability of the Clarin infrastructure to provide functionality even while data and services reside in different locations, it is vital that the components of the targeted web service are placed in multiple Clarin centres. We will determine the best exact placement of each component/resource to be determined later during the project, under guidance of the infrastructure specialist.

However, for the initial insertion of the data and services into the Clarin infrastructure, the intended Clarin centre is MPI, Nijmegen. The MPI has expressed its willingness to also host services, but remarked that a final decision can only be taken pending currently ongoing discussion about persistence and availability of such services.

5 Requested Budget

59975 Euro

6 Description of the Proposed Project

6.1 Research Questions

We target the full range of research on historical texts in fields such as historical linguistics, literature science and history. Examples of research questions would be: How did the system of plural formation change in Dutch from between the Middle Ages and the 17th century? Do differences in orthography follow geographic patterns which align with administrative areas? What is the pattern of introduction of chivalric virtues in originally Dutch literature? Can we identify influences of copyists on literary works by way of wordings linked to a specific time or place? What was the relative power of cities and persons within the County of Holland? How did prices for goods develop through the years?

6.2 Research Data

Answers to the research questions sketched in 6.1 generally have to be based on orthographically transcribed manuscripts in which the transcribers have interpreted missing or unreadable characters as well as markers for abbreviation. Ten example texts will be provided for the demonstrator scenario, taken from the charters in the Corpus of texts written by chancery clerks of the Count of Holland, 1300-1340 (Corpus of the Count's scribes; Rem, 2003), which have been chosen and transcribed by Margit Rem. The texts are currently available in a simple XML format (see appendix A). Metadata such as provenance of the original text and subsequent transcription are available, but still need to be transformed to the CMDI (or IMDI) format.

6.3 Technology

The proposed system will be based on an existing tagger-lemmatizer system which was built by Hans van Halteren on the basis of the manually tagged and lemmatized Corpus van Reenen-Mulder of 14th century Dutch charters (van Reenen and Mulder, 1993; Rem, 2003; NB The example texts do not form part of this corpus). In ten-fold cross-validation on this corpus, the system reached an accuracy of 94-95%. The system currently runs on a Linux machine and consists of various shell scripts, Perl scripts and C programs. It also accesses two external tagging systems, TnT (Brants, 2000) and SVMTool

(Giménez and Márquez, 2004), both of which are freely available for scientific purposes. Documentation is as yet in a rather primal state.

6.4 Description

Research on transcribed manuscripts from the early centuries of the use of the Dutch language, say up to the middle of the 19th century, is seriously hampered by a lack of standardized orthography. As an example, in the corpus Van Reenen-Mulder, the 1055 occurrences of the lemma *penning* (which would be important for the last of the research questions) show 56 different spellings, all of which a researcher would have to guess in order to find the corresponding text passages. The accessibility of the material is enhanced enormously by adding annotation layers showing normalized representations for word boundaries, lemmas and preferably also part-of-speech tags. A fully manual annotation is not only expensive, but also runs the risk of deviations from the agreed upon normalizations, e.g. using the frequent form *penninc* as lemma instead of *penning*. Any annotation should therefore in principle be done semiautomatically, with an automatic system providing the most likely options and a human specialist providing further disambiguation only when deemed necessary.

In this project, we address the automatic system providing the most likely options, in the form of a web-application calling upon web-services performing tokenization, tagging and lemmatization. An input text is first split into tokens by a tokenizing component, then all possible lemma-tag combinations for each token are derived in a lexical component, after which the likelihood of each lemma-tag combination is estimated by a contextual component. Each of these three components accesses specific data resources and can be expected to work best if the resources are tuned to the text's domain and time period. For the current demonstrator project, we will provide resources tuned for 14th century charters.

The tokenizing and contextual component will be atomic web-services. The lexical component, however, will be a web-service which allows for a selection of compatible lexical resources, in the form of lexica relating specific word forms to lemma-tag combinations that might underlie them, each possibly with a numerical indication of its relative likelihood. The main lexica for the demonstrator scenario, tuned for 14th century charters, have been created by Hans van Halteren on the basis of the Corpus van Reenen-Mulder, basing the likelihood indication on counts in the corpus and statistics on spelling variation. It will also be possible for the end user to provide an own lexicon, e.g. concerning lists of specific terms or proper names, but this will have to conform to the format of the current lexicons. For the demonstrator scenario, an example "own" lexicon tuned to the example texts will be provided. All mentioned lexica are currently available in a simple XML format (see appendix B). Metadata describing the format and provenance of the lexicons will be created in IMDI/CMDI MD formats.

The demonstrator is also meant as a catalyst for the scholarly community dealing with historical Dutch texts to provide not only texts in a standard format but also mutually compatible text analysis tools for Clarin. We will attempt to further this cause in a symposium at the completion of the project.

6.5 Plan

Type: Demonstrator project

The overall system will be a web-application. The user interface will offer the user choices a) for the input texts (example texts, other texts in the infrastructure, own texts), b) tokenizer service (example tokenizer, other tokenizers in the infrastructure), c) lexica (example lexica, other lexica in the infrastructure, own lexica), d) disambiguation service (example disambiguator, other disambiguators in the infrastructure) and e) level of desired disambiguation (1-best, n-best or thresholded). Selected resources are checked for compatibility (by way of their metadata), after which the system calls upon the various components in sequence, providing input files and resources as required, and sends the end result to the user's workspace. The components are placed in CLARIN centers as web-services according to CLARIN specifications (which are currently still work in progress). This will also allow their future use with other applications. In this project, the system will not contain metadata search functionality to identify potentially useful resources, but any Clarin resources other than the example data will have to be indicated by ID. The demonstrator scenario will describe step-by-step how one or more of the example texts can be annotated using the example lexica and the additional dedicated lexicon file.

The activities during the project are (in parentheses bracketed lists of deliverables and estimate of effort):

- Create ISOcat-compliant XML-based formats for data streams moving along interfaces between the system and the components. Adapting components to handle these formats. (FRM1-FRM5, SRV1-SRV3, MLS2) (NN 80 hours, HvH 40 hours, IS 20 hours)
- Making web-services out of the components and building central web-application. Testing. (APP, MLS3) (NN 640 hours, HvH 100 hours, MR 20 hours)
- Placing web-application, web-services, resources on Clarin servers. Testing. (MLS4) (NN 80 hours, HvH 20 hours, DB 80 hours, IS 60 hours)
- Building and creating documentation for demonstrator scenario. (DOC5) (MR 20 hours)
- Creating documentation for software. (DOC4, DOC6-DOC8) (NN 80 hours, HvH 80 hours, IS 20 hours)
- Creating documentation for texts and lexica. (DOC1, DOC2) (MR 20 hours, HvH 20 hours)
- Creating documentation for annotation. (DOC3) (MR 40 hours, HvH 20 hours)
- Creating metadata for resources and web-services (MD1-MD6) (MR 40 hours, HvH 60 hours, IS 20 hours)
- Reporting on the experiences vis-à-vis the Clarin infrastructure and standards. (DOC9) (MR 20 hours, HvH 40 hours, NN 80 hours, IS 40 hours)
- Organizing symposium to disseminate results. (MLS5) (MR 40 hours, HvH 20 hours)

.7 Deliverables and Milestones

Deliverables (with indications at the end of which month completion is projected, and person responsible):

- Data formats (data/documents)
  - FRM1: Input to tokenizing component (month 1, HvH)
  - FRM2: Output of tokenizing component, input to lexical component (month 1, HvH)
  - FRM3: Output of lexical component, input to contextual component (month 1, HvH)
  - FRM4: Output of contextual component (month 1, HvH)
  - FRM5: Final output of annotated text (month 1, HvH)
- Metadata (data)
  - MD1: Metadata for input texts (month 2, MR)
  - MD2: Metadata for lexica (month 2, HvH)
  - MD3: Metadata for tokenizing component (month 2, HvH)
  - MD4: Metadata for lexical component (month 2, HvH)
  - MD5: Metadata for contextual component (month 2, HvH)
  - MD6: Metadata for annotated text (month 2, HvH)
- System Components / Web-Services (software)
  - SRV1: Tokenizing component (month 3, NN)

  - SRV2: Lexical component (month 3, NN)
  - SRV3: Contextual component (month 3, NN)
- System / Central Web-Application (software)
  - APP1: Web-Application (month 6, NN)
- Documentation (documents)
  - DOC1: Documentation for input texts (month 3, MR)
  - DOC2: Documentation for lexica (month 3, HvH)
  - DOC3: Documentation for annotated texts (month 3, HvH)
  - DOC4: Documentation for web-application (month 6, NN)
  - DOC5: Documentation for demonstrator scenario (month 6, MR)
  - DOC6: Documentation for tokenizing component (month 4, HvH/NN)
  - DOC7: Documentation for lexical component (month 4, HvH/NN)
  - DOC8: Documentation for contextual component (month 4, HvH/NN)
  - DOC9: Report on experiences with Clarin infrastructure and standards (month 6, HvH/NN)

Milestones (MLS):

- MLS1: Texts and lexica have received metadata and have been inserted in Clarin infrastructure. (month 3, HvH)
- MLS2: Components have been upgraded as to interfacing. (month 3, HvH)
- MLS3: Web-application is running on dedicated machine. (month 5, NN)
- MLS4: System is available through Clarin infrastructure and demonstrator scenario is working. (month 6, NN)
- MLS5: Symposium is held to disseminate the results of the project, demonstrate the system and discuss further contributions to Clarin related to historical Dutch texts. (month 7, MR)

8 IPR and Ethical Issues: Risks

No IPR issues can be expected to occur.

- We have created all data resources ourselves.
- We have created the main software components ourselves.
- Any externally obtained software components are freely available for academic use.

No ethical issues can be expected to occur.

9 Expertise of the applicant(s)

Hans van Halteren has an experience of decades in building taggers and tagger generators. He was invited by Kluwer to edit one of the best-known overview works in this field (van Halteren, 1999) and was involved in the EAGLES standardization initiative and the construction of the tagset and tagger for the Corpus of Spoken Dutch. Using his own and other state-of-the-art tagger generators he built the system from which this project will depart.

Margit Rem has many years of experience with the Corpus van Reenen-Mulder, both as contributor and in later research. She wrote her PhD thesis on the basis of this corpus and the Corpus of the Count's scribes from which the example texts for this project will be taken. She also compiled a corpus of 14th century charters from Brussels, which is to be integrated into the Corpus van Reenen-Mulder. Finally, she cooperated closely with Hans van Halteren in creating the tagger-lemmatizer for 14th century charters.

10 Project budget details

Personnel costs are listed in the table below:

Participant	Organization	Effort (PM)	Salary Costs/PM (Euro)	Salary Costs (Euro)	Travel & subsistence (Euro)	Total (Euro)
Hans van Halteren	CLST/RUN	2.5	5400	13500	300	**14100**
Margit Rem	RUN	1.25	5400	6750	200	**7050**
NN	CLST/RUN	6	5700	34200	1200	**35700**
Daan Broeder	MPI	0.5	5400	2700	125	**2825**
Total		**10.25**		**57150**	**1825**	**58975**

In addition, the symposium at the completion of the project will cost 1000 Euro.

11 Literature

Thorsten Brants. 2000. TnT - A Statistical Part-Of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Jesús Giménez and Lluís Márquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.*

Hans van Halteren. 1999. *Syntactic Wordclass Tagging.* Dordrecht, Kluwer Academic Publishers.

Hans van Halteren. 2009. A tagger-lemmatizer for 14th century Dutch charters. Presentation at: *CLIN2008, Groningen University, Januari 22, 2009.*

Piet van Reenen en M. Mulder. 1993. Een gegevensbank van 14de-eeuwse Middelnederlandse dialecten op computer. In: *Lexikos 3 (Afrilexreeks/series 3 1993), Stellenbosch..*

Margit Rem. 2003. *De taal van de klerken uit de Hollandse grafelijke kanselarij (1300-1340). Naar een lokaliseringsprocedure voor het veertiende-eeuws Middelnederland.* Amsterdam, Münster.

Appendix A: Example input text in XML format

```
<charter>
 <text>2040.gecol</textid>
 <body>
   <w>Allen</w> <w>den</w> <w>ghenen</w> <w>die</w> <w>desen</w> <w>brief</w>
  <w>zien</w> <w>zullen</w> <w>jof</w> <w>horen</w> <w>lesen</w> <p>..</p>
  <w>Maken</w> <w>ic</w> <w>Cont</w> <w>gheraerd</w> <w>alewiins</w> <w>sone</w>
  <w>vte</w> <w>leyden</w> <p>.</p> <w>dat</w> <w>eersame</w> <w>lude</w>
  <w>ende</w> <w>religiose</w> <p>.</p> <w>here</w> <w>jan</w> <w>van</w>
  <w>hoenhorst</w> <w>landcomelduer</w> <w>van</w> <w>der</w> <w>balyen</w> <p>.</p>
  <w>ende</w> <w>dat</w> <w>ghemene</w> <w>conuent</w> <w>van</w> <w>den</w>
  <w>duytschen</w> <w>huyse</w> <w>tote</w> <w>vtrecht</w> <w>bi</w>
  <w>ghemenen</w> <w>consente</w> <w>ende</w> <w>bi</w> <w>goid</w> <w>denken</w>
  <w>heren</w> <w>dierix</w> <w>ende</w> <w>heren</w> <w>willems</w> <w>der</w>
```

```
   <w>prochiepapen</w> <w>te</w> <w>leyden</w> <p>.</p> <w>om</w> <w>bede</w>
   <w>willen</w> <w>mijns</w> <w>heren</w> <w>des</w> <w>grauen</w> <w>van</w>
   <w>heynegouwen</w> <w>ende</w> <w>van</w> <w>holland</w> <w>mi</w> <w>orlof</w>
   <w>ghegheuen</w> <w>hebben</w> <w>mine</w> <w>capelrie</w> <w>te</w> <w>done</w>
   <w>verdienen</w> <w>in</w> <w>hare</w> <w>kercken</w> <w>te</w> <w>sente</w>
   <w>pieters</w> <w>te</w> <w>leyden</w> <w>ende</w> <w>dat</w> <w>ic</w>
   <w>ende</w> <w>mine</w> <w>erfname</w> <w>ewelike</w> <w>die</w> <w>ghifte</w>
   <w>daer</w> <w>of</w> <w>behouden</w> <w>zullen</w> <w>ende</w> <w>gheuen</w>
   <w>sonder</w> <w>arghenlist</w> <w>behouden</w> <w>der</w> <w>moderkerc</w>
   <w>in</w> <w>allen</w> <w>anderen</w> <w>poynten</w> <w>haer</w> <w>recht</w>
   <p>.</p> <w>Jn</w> <w>orconde</w> <w>desen</w> <w>brieue</w> <w>beseghelt</w>
   <w>met</w> <w>minen</w> <w>zeghele</w> <p>.</p> <w>ghegheuen</w> <w>tote</w>
   <w>leyden</w> <w>vp</w> <w>sente</w> <w>michiels</w> <w>auond</w> <p>.</p>
   <w>Jnt</w> <w>iaer</w> <w>ons</w> <w>heren</w> <p>.</p> <w>M</w> <p>.</p>
   <w>CCC</w> <w>viue</w> <w>ende</w> <w>twintich</w> <p>.</p>
 </body>
</charter>
```

Appendix B: Example section of lexicon in XML format

```
<lexicon>
 ...
 <entry>
  <form>orconde</form>
  <annots>
   <annot><lemma>oorkonde</lemma><tag>N(sing,forme)</tag><rlik>0.970</rlik></annot>
   <annot><lemma>oorkonden</lemma><tag>V(fin,pres,lex,forme)</tag><rlik>0.024</rlik></annot>
   <annot><lemma>oorkonde</lemma><tag>N(plu,forme)</tag><rlik>0.006</rlik></annot>
  <annots>
 </entry>
 <entry>
  <form>orconden</form>
  <annots>
   <annot><lemma>oorkonden</lemma><tag>V(fin,pres,lex,formn)</tag><rlik>0.760</rlik></annot>
   <annot><lemma>oorkonde</lemma><tag>N(sing,formn)</tag><rlik>0.196</rlik></annot>
   <annot><lemma>oorkonde</lemma><tag>N(plu,formn)</tag><rlik>0.039</rlik></annot>
   <annot><lemma>oorkonden</lemma><tag>V(fin,past,lex,formn)</tag><rlik>0.005</rlik></annot>
  <annots>
 </entry>
 ...
</lexicon>
```