1. Project Title & Acronym and Abstract

Title: Text-Induced Corpus Clean-up online processing system
Acronym: TICCLops
Abstract: We propose a demonstration project which will allow CLARIN users to submit their corpora for fully automatic spelling correction and normalization by TICCLops, the online processing version of our core component TICCL. This system should be widely applicable in all manner of curation projects and lexicographical work.
Target Start Date: 15 November 2009
Target End Date: 15 May 2010
Type: Demonstrator Project

2. Coordinator

Name: dr. Martin Reynaert
Function: Researcher
Organization: Tilburg University, Tilburg centre for Creative Computing (UvT – TiCC)
Address: Kamer D 342, Postbus 90153, 5000 LE Tilburg
E-mail: Reynaert@uvt.nl
Tel: +31 (0)13 466 3116
Fax: +31 (0)13 466 2892
Role(s): Technology Provider

3. Composition of the Project Team

Name: Drs. Astrid Verheusen
Function: Head Digitisation Department
Organization: Koninklijke Bibliotheek (KB)
Address: Postbus 90407, 2509 LK Den Haag
E-mail: astrid.verheusen@kb.nl
Tel: +31 (0)70 3140911
Fax: +31 (0)70 3140450
Role(s): User, Data Provider

Name: Drs. Remco van Veenendaal
Function: Hoofd TST-Centrale
Organization: Instituut voor Nederlandse Lexicologie (INL)
Address: Witte Singel / Doelencomplex, Matthias de Vrieshof 2-3, 2311 BZ Leiden
E-mail: Remco.vanveenendaal@inl.nl
Tel: 071 - 5272495 / +32 32654601
Fax: 071 - 5272115
Role(s): CLARIN Center and Data Provider

4. CLARIN centre

The Institute for Dutch Lexicology (INL, Instituut voor Nederlandse Lexicologie) .

5. Requested budget: 58,501 euro

5. Description of the Proposed Project

The proposed project aims to fill an important need in the context of CLARIN-NL and CLARIN. The Call is directed at curation and demonstrator projects. The present proposal encompasses a demonstrator project for a generic tool which should prove widely applicable to virtually any curation project.

The technology which is to be brought into the CLARIN infrastructure is at the top of the state of the art in post-correction of large digitized text collections, whether these were obtained by Optical Character Recognition technology or as a result of rekeying. Given the current state of the art in OCR, any result of a digitisation project, however large or small, is qualitatively seriously encumbered. Rekeying is often off-shored and in part due to the language barriers involved still demonstrably error-prone. Finally, very few texts contain no typographical errors, even born digital well-edited text. Further, digital text collections are heir to historical spelling changes, transmission errors, encoding issues, etc.

Text-Induced Corpus Clean-up (TICCL) was developed first as a prototype at the request of the Koninklijke Bibliotheek - The Hague (KB) and reworked into a production tool according to KB specifications (currently at production version 2.0) mainly during the second half of 2008. It is a fully functional environment for processing possibly very large corpora in order to largely remove the undesirable lexical variation in them. It has provisions for various input and output formats, is flexible and robust and has very high recall and acceptable precision. As a spelling variation detection system it is to our knowledge unique in making principled use of the input text as possible source for target output canonical forms. As such it is far less domain-sensitive than other approaches: the domain is largely covered by the input text collection.

We here primarily propose to make this system available online so that texts or corpora can be submitted to it for automatic spelling correction. Given some necessary extensions, TICCL should be able to handle most scenarios and file formats involved in large or small corpora that are to be made available online, for whatever purposes.

1.1 Research Question(s)

Given the technology proposed here, all research questions raisable under the CLARIN framework should become better addressable. All the text collections involved and processed with the technology offered here should be qualitatively improved and therefore affording more gainful access and exploitation.

1.2 Research Data

Within this project we will work on huge text collections which have been or are being digitized by the KB. We will work on OCR-ed Acts of Parliament (http://www.statengeneraaldigitaal.nl) covering 1920 to 1995, i.e. in historical and contemporary spelling. Further on digitized newspapers: we also have a gold standard for 'Het Volk-1918' (http://www.kb.nl/hrd/digi/ddd/index-en.html). The technology provider is thoroughly acquainted with these text collections and provides statistics on the lexical variation within them in Reynaert (2008). We now specifically wish to work on the new collections of recently digitized copyright-free books and magazines.

IPR, availability: All KB-collections involve copyright free works. All are or will be made freely available online.

Standards in use at the Koninklijke Bibliotheek:
For the opening up of collections the KB works to a fixed infrastructure. This infrastructure makes use of open standards and proven working methods. We here list only those most relevant for the current proposal, the full list is at: http://www.kb.nl/hrd/digitalisering/standaarden-en.html)

- For the descriptive metadata Dublin Core is used. KB-specific elements are created if the specific metadata do not occur in the Dublin Core. The descriptive metadata are stored in XML format.
- The page layout (the different zones, such as images, columns and headings) is stored with the help of a segmenting standard ALTO which is an XML format.
- All files will be made accessible with persistent URLs.
- For search queries made from a web application, the SRU-protocol, is used. With this the search queries can be included in a URL in a standard manner.

Gold standards to be used for evaluation purposes within this project:
Gold standards are created by querying the text collection's web site for an unidentifiable text string encountered within the collection. Thanks to the underlying Alto xml-file the corresponding area on the image that was used to derive the digital text by means of OCR is then highlighted, allowing for resolving the otherwise unidentifiable text string.

1.3 Technology

TICCL is now in production version 2.0 according to the specifications of our user and data partner, the Koninklijke Bibliotheek. It is currently undergoing acceptance testing at the KB. TICCL production version 2 has been written entirely in Perl, version 5.8.8. It requires no extra external modules. The code base currently consists of about 4,500 lines of code. TICCL can make use of various types of lexicons, among which the open source spelling corrector Ispell's lexicons, which are freely available for a wide range of languages. It also makes productive use of the Ispell affix files, which gives it morphological knowledge for the particular language.
Documentation: TICCL is fully documented in both English and Dutch. Besides the user manual, there is a functional and a technical description. The prototype underlying the

current production version is described in Reynaert (2008), exciting new developments towards parallel look-up of all the pairs of word tokens in a corpus displaying a particular character confusion are described in Reynaert (2009).
IPR: TICCL is open source software. The core correction algorithm developed in the framework of Reynaert (2005) has been distributed under the GPL license since 2005. In the context of our current work for the Stevin project SoNaR, we need to seek approval of the system's Open Source status from the Nederlandse Taalunie (NTU). Whatever the outcome of this, the INL and Dutch HLT Agency being part of the NTU, the IPR on TICCL will be in the hands of a partner of the current project's consortium.

1.4 Description

TICCLops should find widespread application. Any digital text project should profit from its availability. It addresses and answers huge needs in the current expensive and often frustratingly low-quality transition from paper-bound to digital society. It has been designed bottom-up to be flexible, easily adaptable to other languages and language varieties, whether contemporary or historical. In that the vocabulary of the corpus to be corrected plays an integral role in the correction process by supplementing the vocabulary of the validated lexicon(s), the system is not vulnerable to domain variation.

In the Stevin project SoNaR a 500 MW balanced reference corpus of contemporary written Dutch is being built. All texts are converted to D-Coi/SoNaR xml format, which contains an IMDI-header for the metadata. This format is destined to become a de facto standard for Dutch corpora and is a very likely candidate to become a CLARIN-NL accepted standard. As a byproduct of the proposed project, TICCLops will be able to handle the D-Coi/SoNaR xml-format.

1.5 Plan

Type: Demonstrator Project

Demonstrator project: Project duration is 6 months. The codes T/D refer to TICCLops deliverables, the codes T/M to TICCLops milestones.

- T/D1/core component: TICCL's input and output facilities need to be extended to handling compressed archives. TICCL production version 2 will further be extended to not only be able to handle raw text or element-based xml but also attribute-based xml such as the Alto xml format. This will be done in Month 1 by M. Reynaert. Throughout the project a scientific programmer will be employed to revise the whole code base with an eye on efficiency in terms of speed and memory use, besides his core task of turning TICCL into the TICCL online processing system.
- T/D1/application: TICCLops: TICCL will be given a CGI-interface. The user interface will provide upload facilities for the user's corpus (allowing for a range of compressed archive formats) and menu-based or API-directed parameter selection options. These options will allow the user to specify which of the available lexicons to use (availability will range from free lexicons to lexicons

available only after arrangement e.g. with the HLT-agency (TST-Centrale). The options will further enable the user to set specific system parameters, e.g. whether or not the system should perform word bigram based variant look-up. The CGI-environment and user interface will be added in Months 1-2 by M. Reynaert and a TiCC scientific programmer. In Month 3 TiCCLops will be locally put online at TiCC for thorough testing. It will be evaluated against the off-line production version 2 on the basis of the gold standards built in prior research by M. Reynaert and by the KB in the production version 2 acceptance tests.

- T/M2: Available on a server at CLARIN recognized center: Month 4 will be used by M. Reynaert and INL to install TiCCL and TICCLops at the INL CLARIN center. KB and TiCC will test the online processing system and verify the Windows server correction results in comparison with the Linux server results obtained locally at TiCC. We foresee no insurmountable difficulties, discrepancies in correction results or incompatibility with the CLARIN infrastructure architecture here in that prior tests and comparisons between Linux and Windows server deployments have proven near identical in the past, both at TiCC (Linux and Windows) and the KB (Windows).
- T/D3: Demonstration scenario: For the demonstration scenario we will choose a digitized copyright-free book and build a gold standard for it. The gold standard itself is a valuable reusable resource which allows for measuring the effectiveness of any extension or modification to the system or for comparison with competitive alternative systems. All aspects of using the system will be demonstrated to prospective users on the basis of this gold standard. These aspects include pre-packaging the corpus, in this case the book, as a compressed archive, choosing and setting the system parameters and receiving and interpreting the results. The actual evaluation results returned will then teach the prospective user what to expect from the system and to choose which output formats best suit his own needs. In Month 5 this scenario will be written by M. Reynaert in cooperation with Astrid Verheusen. M. Reynaert gradually builds the gold standard in the preceding months.
- T/D2: TICCLops documentation: Month 6 will be devoted by M. Reynaert and the scientific programmer to revising the existing TICCL production version 2 documentation in light of the extensions effected and to writing the new TICCLops specific user, API and developer documentation as well as document (T/D4/document): TICCLops requirements and desiderata. A mapping between TICCLops specific categories and ISOcat categories and extending ISOcat with required new categories will be undertaken within this same documentation period (T/M5 and T/D5/document and data).

6. Deliverables and Milestones

We distinguish between TICCLops milestones: T/M and TICCLops deliverables: T/D.

- Metadata of the resources and text collections dealt with in the project are already available online at the KB web sites. PIDs are available.

- T/M1: Demonstrator metadata made available on a recognized CLARIN server : 15 January 2010: M. Reynaert and R. van Veenendaal.
- T/D1: TICCLops: core component underlying the demonstrator 15 February 2010 : M. Reynaert. TICCLops application fully functional at INL CLARIN centre: 15 May 2010 : M. Reynaert and R. van Veenendaal.
- T/M2: TICCLops demonstrator available on a recognized CLARIN centre: 15 March 2010: M. Reynaert and Remco van Veenendaal.
- T/D2/documents: TICCLops Documentation : 15 May 2010 : M. Reynaert.
- T/D3/document and data: Demonstration Scenario with gold standard: 15 March 2010 : M. Reynaert.
- T/D4/document: TICCLops requirements and desiderata for the CLARIN infrastructure: 15 May 2010: M. Reynaert.
- T/M5: ISOcat extended with new entries: 15 May 2010 : M. Reynaert.
- T/D5: Mapping table defining a mapping between the resource-specific linguistic categories and ISOcat data categories: 15 May 2010 : M. Reynaert.

7. IPR and Ethical Issues: Risks

We see no ethical issues. Given the composition of the project consortium, we see no IPR issues: all data provided by the KB are copyright free. TICCL is open source. If this status were not to be ratified by the Nederlandse Taalunie, its IPR would be in the hands of the Dutch HLT-Agency which is part of consortium partner INL.

8. Expertise of the applicant(s)

TiCC: Martin Reynaert has long affinity with corpora and lexical variation in general and is the developer of anagram key hashing which forms the solid basis on which TICCL is built. The various approaches to spelling correction developed over the past years by the developer have variously shown great performance, wide and even multilingual applicability and are generally not language-specific. He worked in the Stevin project D-Coi and is now coordinator of corpus building in the Stevin project SoNaR. He is a member of staff at the Induction of Linguistic Knowledge research team (ILK), now part of TiCC at Tilburg University. ILK has a long tradition of and ample expertise in putting online demonstrators for its various language technology products. These are online at: http://ilk.uvt.nl.
KB: The National Library of the Netherlands fosters the national infrastructure for scientific information and plays an important role in the permanent access to digital information at an international level. It has developed a vast experience in the digitisation of images and text since the mid-1990s and its e-Depot, the world's first digital archiving system for academic publications, now contains more than 7.5m articles. The KB also hosts the offices of The European Library (TEL), and of the European Digital Library (EDL). The work on CLARIN-NL will be carried out within the Research and Development (R&D) Section of the KB which has taken part in many EC-funded projects over the years, e.g. TEL and the development of tools and services for digital preservation in the FP6 project PLANETS.

INL: The Institute for Dutch Lexicology (INL, Instituut voor Nederlandse Lexicologie) has a well-established track record in lexicology and lexicography, participated/s in (inter)national projects like DAM-LR, IMPACT, CLARIN and CLARIN-NL and is actively making digital Dutch language resources available through its Flemish-Dutch HLT Agency (TST-Centrale, Centrale voor Taal- en Spraaktechnologie), an initiative of and financed by the Dutch Language Union (NTU, Nederlandse Taalunie). The INL has a strong ambition to become a (type B) Center in the CLARIN infrastructure.

9. Project budget details

Participant	**Organization**	**Effort (PM)**	**Salary Costs/PM (Euro)**	**Salary Costs (Euro)**	**Travel & subsistence (Euro)**	**Total (Euro)**
M. Reynaert	TiCC-UvT	2.1	5,318	11,167	525	**11,692**
Scientific programmer	TiCC-UvT	6	5,695	34,173	1,500	**35,673**
A.Verheusen	KB	1	5,318	5,318	250	**5,568**
R. van Veenendaal	INL	1	5,318	5,318	250	**5,568**
Total		**10.1**		**38,000**	**2.525**	**58,501**

10. Literature

Reynaert (2005):
Martin Reynaert. Text-Induced Spelling Correction.
PhD thesis, Tilburg University, 2005. ISBN. 90-9020100-9.
Reynaert (2008):
Martin Reynaert. Non-interactive OCR post-correction for giga-scale digitization projects.
In Proceedings of CICLing 2008. Lecture Notes in Computer Science Vol. 4919/2008, pages 617--630, Berlin / Heidelberg, 2008. Springer. http://ilk.uvt.nl
Reynaert (2009):
Martin Reynaert. Parallel identification of the spelling variants in corpora.
In Proceedings of the Third Workshop on Analytics For Noisy Unstructured Text Data (Barcelona, Spain, July 23 - 24, 2009). AND '09. ACM, New York, NY, 77-84. DOI= http://doi.acm.org/10.1145/1568296.1568310