CLÉA – *Literature Éxploration and Annotation Environment* for Google Books corpora

Application for a
Google Digital Humanities Research Award (Europe)

Prof. Dr. Jan Christoph Meister
University of Hamburg, Germany

Submission date: 30 October 2010

Abstract

We propose to supplement Google Books with a web based collaborative text exploration, markup and analysis environment, hereafter referred to as CLÉA (*Collaborative Literature Éxploration and Annotation*)[1]. The development of CLÉA builds on our open source desktop application CATMA which already supports high-level semantic annotation through TEI compliant, non-deterministic stand off markup. Development of the CLÉA system will be supported by a pilot study in which a Google Books subcorpus of approx.15-20 German pre-20th Century literary texts are tagged for specific narratological phenomena. The crowd sourced markup will then be analysed with a view to augmenting existing automated pre-processing in CLÉA via machine learning.

[1] In Ancient Greek mythology, Clea—like Pythia—was a priestess at the oracle of Delphi. Just to be different, use non-English characters and emphasise "exploration" as the key activity of literary scholars, we spell it with an accent.

Table of contents

Principal Investigator 3
Principal Investigator CV 3
Project related publications 2003 – 2010 4
Collaborators 6
Hamburg University 6
CATMA Development Team 6
External Collaborators 6

Proposal Body 7
0. Rationale 7
I. Research Objectives and Expected Results 7
I.1 Conceptual foundation: Hermeneutical approach to markup 8
I.2 Objectives and deliverables 8
I.2.1 Object-/Meta-Data and User Management Module 8
I.2.2 Textual markup and analysis module 8
I.2.3 System complex integration and web GUI 9
I.2.4 Pilot study analysis & extension of automated routines 9
II. Benefits to the Research Community 10
III. Data and services needed from Google 11
IV. Validity of Approach 11
IV.1 Availability of Data 11
IV.2 Computational Soundness 11
IV.2.1 Systems development expertise 11
IV.2.2 Machine learning expertise (ML) 11
IV.2.3 Computational narratology 12
V. Description of Funding Usage 12

VI. Project Implementation 12
VI.1. Preliminary Schedule 12
VI.2. CLÉA Component Diagram 13

VII. Bibliography 14

Principal Investigator

Jan Christoph Meister, Dr. phil. habil.
Professor for Modern German Literature
(Literary Theory, Text Analysis, Computational Philology)
University of Hamburg, Faculty of Humanities
Department SLM I – German Studies II
Von-Melle-Park 6
D- 20146 Hamburg

Office: +49 40 42838 2972
Cell: +49 172 40 865 41
eMail: jan-c-meister@uni-hamburg.de
Web: http://www.jcmeister.de

Principal Investigator CV

Born 1955 in Montreal / Canada

Academic Career

as of 2006: University of Hamburg

Professor of Modern German Literature (Theory of Literature, Methodology of Textual Analysis and Literary Computing) in the Department of Language, Literature and Media I, Faculty of Humanities, University of Hamburg.

2006: LMU, Munich

Professor of Modern German Literature and Humanities Computing ('Professor für Neuere deutsche Literaturwissenschaft und Computerphilologie ') at the Ludwig-Maximilians-University, Munich.

2001 - 2006: University of Hamburg

Since 4/2004 project leader for the project on 'Story Generator Algorithms' and for NarrNetz, an E-Learning project on Narrative Theory. 2002-2004 Chairperson of the Arbeitsstelle Computerphilologie (Working Group Literary Computing); co-author and coordinator for the E-Learning course C-Phil Online,

4/2001 - 3/2004 academic researcher in the DFG (German Research Foundation) funded Narratology Research Group (FGN) at Hamburg University, project on 'The Temporality Effect'.

2001 Habilitation at the University of Hamburg with a study on a narrative theory of action and Humanities Computing.

1986 - 1995: University of the Witwatersrand, Johannesburg

1986-88 Lecturer, 1988-90 Senior Lecturer, 1990-95 Associate Professor ad hominem (tenured) and Head of German Department / Deputy Chairperson Modern Languages at the University of the Witwatersrand, Johannesburg/South Africa.

1981 - 1986: University of Hamburg

Academic Assistant in the Department of German; Doctoral Dissertation and Doctorate in Philology (1985)

1974 - 1981: Undergraduate Studies at Hamburg University

Research focus

Literary Computing, Textual Analysis, Narratology, 20th Century Austrian Literature, Fantastic Literature. For a list and descriptions of research projects related to this proposal see http://www.jcmeister.de/html/projects.html

DH related offices and activities

- Member of the Executive Committee of the Association for Literary and Linguistic Computing (ALLC) and the Alliance of Digital Humanities (ADHO) Multi-Lingual / Multi-Cultural Subcommittee.
- Chair of the Hamburg Digital Humanities (HDH) Initiative (www.hdh.uni-hamburg.de)
- Local organizer of the Digital Humanities International Conference DH 2012, University of Hamburg 16-22nd July 2012.

Project related publications 2003 – 2010

Books

Computing Action. A Narratological Approach. Translated by Alastair Matthews. Foreword by Marie-Laure Ryan. Berlin, New York (de Gruyter) 2003 (= Narratologia, Bd.2) Abstract, introduction and table of contents at http://www.jcmeister.de/html/computing-action1.html

Narratology beyond Literary Criticism. Mediality, Disciplinarity. Edited by Jan Christoph Meister in collaboration with Tom Kindt and Wilhelm Schernus. Berlin, New York (de Gruyter) 2005 (= Narratologia, Bd. 6)

Einführung in die Erzähltextanalyse. Silke Lahn, Jan Christoph Meister. Unter Mitarbeit von Matthias Aumüller, Benjamin Biebuyck, Anja Burghard, Jens Eder, Per Krogh Hansen, Peter Hühn und Felix Sprang. Stuttgart und Weimar (Metzler) 2008.

Chapters in books

"Computational approaches to narrative." In: Herman, David; Ryan, Marie-Laure (eds.), Routledge Encyclopedia of Narratology. London and New York 2005, 78-80

"Minimal Narrative"; In: Herman, David; Ryan, Marie-Laure (eds.), Routledge Encyclopedia of Narratology. London and New York 2005, 312

"Narrative Units." In: Herman, David; Ryan, Marie-Laure (eds.), Routledge Encyclopedia of Narratology. London and New York 2005, 382-384

"Le Metalepticon: une étude informatique de la métalepse." In: Pier, John; Schaeffer, Marie-Jean (ed.): Métalepses. Entorses au pacte de la représentation. Éditions de lécole des hautes études en sciences sociales. Paris 2005, 225-246 – English online version “The Metalepticon: a Computational Approach to Metalepsis” at http://www.jcmeister.de/downloads/texts/jcm-metalepticon.html

"‘Narrativité’, ‘événement’ et l’objectivation de la temporalité." In: Pier, John: Théorie du récit. L’apport de la recherche allemande. Villeneuve d’ Ascq (Presses Universitaires du Septentrion) 2007, 189-208

“Computational Narratology oder: Kann man das Erzählen berechenbar machen?” In: Müller, Corinna, Scheidgen, Irina (Eds.): Mediale Ordnungen. Erzählen, Archivieren, Beschreiben. Schriftenreihe der Gesellschaft für Medienwissenachften (GfM) 5. Marburg (Schüren Verlag) 2007, 19-39

(with Jörg Schönert): The DNS of Mediacy. In: Peter Hühn, Wolf Schmid, Jörg Schönert (Hg.): Point of View, Perspective, and Focalization. Berlin, New York (de Gruyter) 2009 (= Narratologia 17), 11-40. Online at http://www.jcmeister.de/downloads/texts/meister_schönert_2009.pdf

Meister, Jan Christoph: "Narratology". In: Hühn, Peter et al. (eds.): the living handbook of narratology. Hamburg: Hamburg University Press. Online at hup.sub.uni-hamburg.de/lhn/index.php ?title=Narratology&oldid=850

Journal articles

"Tagging Time in Prolog. The Temporality Effect Project"; in: Literary and Linguistic Computing; 2005 vol. 20: 107-124. Online at http://www.jcmeister.de/downloads/texts/jcm-tagging-time.html

Birte Lönneker, Jan Christoph Meister, Pablo Gervás, Federico Peinado and Michael Mateas (2005): "Story Generators: Models and Approaches for the Generation of Literary Artefacts.; in: ACH/ALLC 2005 Conference Abstracts. Proceedings of the 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, Victoria, BC, Canada, June 15-18, 2005. 126-133. Online at http://www.jcmeister.de/downloads/texts/loenneker-meister-etal-2005.pdf

Jan Christoph Meister: "Events are Us." In: Amsterdam International Electronic Journal for Cultural Narratology, (AJCN) No. 4 / Autumn 2007. http://cf.hum.uva.nl/narratology/a07_meister.htm [22.04.08]

Stefan Gradmann, Jan Christoph Meister: "Digital document and interpretation: re-thinking 'text' and scholarship in electronic settings". In: Poiesis & Praxis. International Journal of Ethics of Science and Technology Assessment, 2008. Online at http://www.springerlink.com/content/g370807768tx2027/fulltext.html [22.04.08]

In print

"Tales of contingency, contingencies of telling: Towards an algorithm of narrative subjectivity." In: Barry Mazur et.al. (eds.), Narrative and Mathematics. Princeton UP. Pre-publication (not for circulation): http://www.jcmeister.de/dh-prepub/algnarrsub.pdf

"Crowd sourcing true meaning. A collaborative markup approach to textual interpretation." In: Marilyn Deegan, Willard McCarty (eds.): Festschrift for Harold Short. OUP. Pre-publication (not for circulation): http://www.jcmeister.de/dh-prepub/crowdmean.pdf

Collaborators

Hamburg University

- Prof. Christopher Habel, University of Hamburg, Department of Informatics, Knowledge and Language Processing (collaborating NLP expert)
- Prof. Christoph Klauck, Hamburg University of Applied Sciences (collaborating ML expert)

CATMA Development Team

- Marco Petris, Dipl. Informatiker (University of Hamburg)
- Evelyn Gius, M.A. Computational Linguistics (University of Hamburg)
- Lena Schüch, MA German Studies (University of Hamburg)
- Malte Meister, BSc Computer Science (ICT, Cape Town)

External Collaborators

- Prof. Pablo Gervas, Departamento de Ingeniería del Software e Inteligencia Artificial (http://nil.fdi.ucm.es/index.php?q=node/92)
  Facultad de Informática, Universidad Complutense de Madrid
- Prof. Worthy Martin, Department of Computer Science, School of Engineering and Applied Science, University of Virginia
  (http://www.cs.virginia.edu/people/faculty/faculty.php?member=martin)
- Prof. Stéfan Sinclair, Communication Studies & Multimedia, McMaster University, Hamilton/Ontario, Canada
  (http://csmm.mcmaster.ca/faculty/profile_sinclair.html)

Proposal Body

CLÉA – a *Collaborative Literature Éxploration and Annotation* environment for Google Books corpora

0. Rationale

Humanities researchers in the field of literary studies *access* and *read* literary texts in digital format via the web in increasing numbers—but with few exceptions, they do not yet use web-based tools to *interpret* these texts. For most scholars, apart from search and find, the actual processing of a text still takes place outside the digital realm. The interest essentially motivating human encounters with literature does thus not seem to benefit from the new paradigm: hermeneutic, i.e. "meaning" oriented high-order interpretation that transcends a mere decoding of information.

The creation of web based text analytical services such as *Voyeur* (http://hermeneuti.ca) marks an important step in this regard. However, high-order hermeneutic interpretation requires more than automated pattern, string- or word-level analysis of the source object, namely the ability to add semantic markup and analyse both the object-data and the meta-data in combination. Also, true hermeneutic activity is not deterministic, but explorative: in the scholarly interpretation of literature we are not looking for *the right* answer, but for *new, plausible and relevant* answers. This has significant consequences for the design of a system aimed at supporting this activity.

The vision motivating our project addresses this shortfall: If we can pool the *source data*—in our case, literary texts—in a comprehensive open access repository, such as Google Books, and use digital technology to analyse it, then why not use the same collaborative approach and technological means to create, collect, aggregate and analyse their related *meta data* that will help us to *interpret* literary texts?

I. Research Objectives and Expected Results

We propose to supplement Google Books with a web based collaborative text exploration, markup and analysis environment, hereafter referred to as CLÉA (Collaborative Literature Éxploration and Annotation). Development of the system will be supported by a pilot study in which a Google Books subcorpus of approx.15-20 German pre-20th Century literary texts are marked up for specific narratological phenomena in a crowdsourcing approach.

As CLÉA aims to support high-level hermeneutic text interpretation, it must be based on an approach to markup that transcends the limitations of low-level text description. We define this distinction as follows: Description cannot tolerate ambiguity, whereas an interpretation is an interpretation if and only if at least one alternative to it exists. Note that alternative interpretations are not subject to formal restrictions of binary logic: they can affirm, complement or contradict one another. In short, interpretations are of a probabilistic nature and highly context dependent.

To meet these conceptual requirements CLÉA will be implemented as a web based client-server architecture featuring

- a core server module version of our existing desktop application CATMA (http://www.catma.de) functionally complemented by
- a source- and meta-data document & user management server module that (prospectively) interfaces with Google Books and
- a web front-end emulating the traditional hermeneutic workflow pattern established in literary studies disciplines.

Rather than starting from scratch the implementation will make use of open source third party components where feasible, in particular pre-processing routines implemented in *Voyeur*.

I.1 Conceptual foundation: Hermeneutical approach to markup

Unlike structural markup, which in most cases can be expressed by a more or less fixed tag set, hermeneutical markup needs to be flexible and extensible. This requires a data structure which supports such flexibility, yet is standardized enough to enable tools interoperability. The markup should be stored in a stand-off manner to enable overlapping and contradictory markup, and it should be stored separated from the literary artefact to support the coexistence of markup for different purposes, or even the coexistence of conflicting markup for the same purpose. Only this data model will acknowledge the standard practice in literary studies, i.e. a constant revision of interpretation (including one's own) that does not necessarily amount to falsification.

In view of these requirements, our approach is based on a non-deterministic "one-to-many" data model which we aim to (a) implement in a fully-functional web-based system, (b) test in a multi-user pilot study with a corpus of German language literary texts and (c) analyse and potentially exploit in a machine-learning approach in order to enhance the system's automated pre-processing functions. This project brief translates into the following objectives and deliverables:

I.2 Objectives and deliverables

I.2.1 Object-/Meta-Data and User Management Module

Our **objective** is to develop a fully functional *object-/meta-data and user management server module.* As a **result**, we expect the implementation of a module which

- enables users to compile a virtual source document corpus (consisting of 1 to n individual documents taken from the Google Books repository) and make it available to the markup and analysis module;
- stores and manages all source & standoff markup documents generated by users so that all interpretations associated with a defined text or text corpus can be dynamically aggregated, re-used, edited and extended and shared among users.

I.2.2 Textual markup and analysis module

Our **objective** is to transform our existing desktop application CATMA[3] into a scalable server module. As a **result**, we expect the implementation of a second module which

- relates a corpus (= a defined set of source documents) to n stand-off markup documents;

[3] Key technical features of CATMA are: open source, TEI/XML compliant, fully JAVA based and platform independent. For further details see http://www.catma.de .

- gives the user full control over the creation and management of TEI compliant tag-sets;
- integrates an expandable set of automated pre-processing routines (indexing, distribution and collocation analysis etc.);
- enables users to generate non-deterministic, low to high-level markup that can be exploited in subsequent manual as well as automated interpretive procedures.[4]

This CLÉA module will also enable 'on-the-fly' crowdsourced correction of e.g. scanning errors in source documents applied in parallel to semantic markup. In order to guarantee integrity of the original source document CLÉA will not intervene with the source document, but will capture corrections as a specific category and layer of markup: text correction is conceptualized as creating a text variant, not as a retrospective alteration of the primary text witness.

I.2.3 System complex integration and web GUI

Our **objective** is to develop an intuitive web GUI (frontend) for the system complex. As a **result**, we expect the implementation of a system frontend providing

- a workflow oriented browser based GUI that integrates
- the object-/meta-data & user management module (including provision for a potential Google Books API) and
- the textual markup and analysis core module.

I.2.4 Pilot study analysis & extension of automated routines

In parallel to the project, we will conduct a multi-user pilot study in collaborative markup using a corpus of non-English (German) literary texts. The markup will focus on the phenomenon of narrative 'perspective' as defined in the narratological theory of Genette and particularly Schmid (2010)[5]. The study will be undertaken as part of two regular BA/MA-courses in German Studies at Hamburg University towards the end of the summer semester.[6] Results from this study will be utilized as follows:

Our **objective** is to apply a machine-learning approach to the body of markup documents created by the pilot group. As a **result**, we expect to

- detect regularities in the outcome as well as in the procedures of markup that may be modelled by way of algorithmic formalization;
- develop a feature set based on the detected regularities as a starting point for feature selection and/or feature engineering and the exploration of suited machine learning algorithms with the WEKA[7] data mining software;

[4] CLÉA will also implement a key feature request of CATMA users, namely the possibility to correct scanning errors in the source document in parallel to markup. In order to guarantee integrity of the original source document CLÈA will capture crowdsourced document correction as a specific category and layer of markup.

[5] See http://books.google.de/books?id=Do6e5MZuADcC&pg=PA199&lpg=PA199 [seen 26/10/2010].—In narratological terms perspective is defined as "the way the representation of the story is influenced by the position, personality and values of the narrator, the characters and, possibly, other, more hypothetical entities in the storyworld" (Niederhoff 2010: http://hup.sub.uni-hamburg.de/lhn/index.php/Perspective/Point_of_View [27.10.2010])

[6] Normally first and second week of July. While strongly desirable, prior completion of the development detailed under objectives 1-3 is not critical as the existing desktop version of CATMA may be used as a fall-back option without compromising on the conceptual approach.

[7] See http://www.cs.waikato.ac.nz/~ml/weka/index.html (seen 26/10/10)

- implement new routines as an integral part of the textual markup and analysis module's heuristic pre-processing functionality
- provide an integration between these trained algorithms and a (productive or experimental) Google API for testing and executing algorithms with large scale data and/or large scale computing power.

II. Benefits to the Research Community

To date, digital text markup generated by literary studies projects has generally been highly specific to the projects' individual research questions. Unlike linguistic markup (e.g., POS tagging), literary markup is hardly ever re-used or made accessible to the community. From the point of view of the literary scholar, this is CLÉA(sic!)rly a waste of resources: instead of building on existing markup, new projects have to start from scratch. In addition, from the point of view of the digital humanist, we must ask: if the markup expertise which scholars of literature apply in their individual projects is not preserved and analysed, how can we expect to make progress in the automation of these processes?

The intention of the CLÉA-project is to crowdsource semantic markup via a web based environment that comprises document&user management as well as digital text markup and analysis routines. Combining a collaborative approach to markup with this technological framework will benefit the DH scientific community in the following ways:

- The outcome of the project will *facilitate future sharing and collaborative creation of markup among researchers*. This will eventually complement Google Books as a repository of source texts with a repository of public domain markup documents.
- The system is conceptually based on an *approach fully acknowledging the non-deterministic, potentially ambiguous semiotic nature of the literary object*. Because of this, and because the GUI aims to emulate the traditional workflow in order to guarantee a low threshold for the non-expert user, the system will help to further promote the use of digital resources among humanities scholars.
- By monitoring and analysing how humans make interpretive decisions through a machine learning approach, the project will help to *expand the range of interpretive procedures that can be computationally modelled and automated*.
- The multi-user pilot study run in parallel to the second phase of the system development process will markup a subcorpus of German texts taken from Google Books. Working with non-English source texts will enable us to *identify language-specific markup requirements* (for example, the need to account for historical variants of language use, such as the lack of a reliable norm for punctuation in pre mid-19th century German literature, which can turn the identification of sentence borders from trivial automated parsing into a complex interpretive task).

The markup created in the pilot study will focus on identifying phenomena of 'narrative perspective'—simply put, the way and modes by which a narrative account of something is profiled in terms of epistemological, normative and aesthetic constraints. Unlike many of the more 'fuzzy' and hard-to-model high-order functions of narrative, the relevant constituent phenomena of narrative perspective have been extensively analysed and explicated by narratology (the 'science of narrative') and can thus be considered theoretically well-defined, strong candidates for formalization. One of the anticipated outcomes of the project is thus the development of a narratological pre-processing routine that can detect 'perspective markers' in automatic or semi-automatic mode.

III. Data and services needed from Google

The project will require the following from Google:

- A subcorpus of approx. 20 German literary texts originally published between 1790 and 1890. Data can be provided in Google's raw digital text format(s) and will be imported into CLÉA via a routine which we will build;
- a (productive or experimental) Google Books API to simulate future direct access and corpus aggregation from within the CLÉA environment;
- a (productive or experimental) Google Algorithms API to simulate future integration of the developed algorithms in a Google supercomputing environment;
- expertise in terms of machine learning approaches. In particular, the project would greatly benefit from a start-up and scoping plus a mid-project assessment workshop with Google experts in the fields of pattern recognition and machine learning.

IV. Validity of Approach

IV.1 Availability of Data

We have verified that a sufficient amount of suitable object data (i.e. German literary texts) is available on Google Books. Further technical details relating to data availability and format were the subject of exploratory discussions with Google's Leslie Yeh Johnson in October 2010.

IV.2 Computational Soundness

IV.2.1 Systems development expertise

The core module of the CLÉA-system, which we propose to develop, builds on the existing stand-alone application CATMA, a robust platform independent markup and text analysis software developed with start-up funding provided by the University of Hamburg since mid-2008. The CATMA development team includes narratologists, computer linguists and two software developers with extensive experience in the development and maintenance of computer systems for commerce and manufacturing industries.

CATMA fully complies with the conceptual model of hermeneutic markup described under I.1. and has been extensively tested in BA- and MA German literature and DH-courses since winter semester 2008/2009. The software was presented to the scientific community at the international digital humanities conferences DH 2009 in Maryland and DH 2010 in London; it is also an integral component of the newly inaugurated "Digital Commons Project" jointly funded by the Canadian SSHRC and the German Humboldt Foundation in the context of which we cooperate with the Voyeur team (Prof. Stéfan Sinclair, McMaster University; Prof. Geoffrey Rockwell, University of Alberta).

Among other CATMA is currently being used outside Germany by researchers in Finland (Prof. Lisa-Lena Oppinen of Oulu University), the US (Prof. William Kretzschmar of University of Georgia, Alexander Dorsk of MBLWHOI Library, Liang Zhou of Thomson Reuters R&D) and Italy (Prof. Maurizio Lana of Università del Piemonte Orientale).

IV.2.2 Machine learning expertise (ML)

ML expertise in the core CATMA development team is not sufficient. To address this shortfall, CLÉA investigation into ML will be supported by a group of CS experts with whom we have collaborated in past projects. The following have contributed to the current proposal and confirmed their interest to collaborate:

- AI, NLP, ML: **Prof. Pablo Gervas**, Departamento de Ingeniería del Software e Inteligencia Artificial, from the Facultad de Informática, Universidad Complutense de Madrid. (Previous cooperation: Story Generator Algorithm-project ; see http://www.jcmeister.de/html/sga1.html).
- ML: **Prof. Worthy Martin**, Department of Computer Science, School of Engineering and Applied Science, University of Virginia. (Current cooperation in the Digital Commons Initiative project; see http://www.jcmeister.de/html/projects.html).
- ML, AI, Distributed Systems: **Prof. Christoph Klauck**, Hamburg University of Applied Sciences (http://users.informatik.haw-hamburg.de/~klauck/index.html)
- NLP: **Prof. Christopher Habel**, University of Hamburg, Department of Informatics, Knowledge and Language Processing. (Previous cooperation: Project on time modelling in narratives; see http://www.jcmeister.de/html/temporality1.html).

IV.2.3 Computational narratology

The project draws on experiences from a number of previous projects investigating potentials for cross-fertilization among narratology and computational modelling (for details see http://www.jcmeister.de/html/projects.html). The narrative phenomenon of ‘perspective’ is the subject of Meister&Schönert (2009)[11].

With a particular view to the narratological markup of narrative perspective, a tagset for this and other narratological categories will be created on the basis of the narratological theory mentioned in I.2.4 and tested on several (English and German) texts during an independent research project starting in January 2011. These tag-sets can be implemented in CATMA and CLÉA so that they may be used as a basis for text analysis. In cooperation with students from McMaster University (Canada), means for the graphic representation of the created 'perspective'-tags both in *Voyeur* and CATMA/CLÉA will be developed. This research is part of a PhD. dissertation project supervised by the principal investigator.

V. Description of Funding Usage

Funding of US $ 50.000 will be allocated as follows:

- 80% = US $ 40.000 for the employment of 1 system developer as a Research Assistant (annual personnel cost for 1 x 50% post = € 28.500)
- 20% = US $ 10.000 for external programming contracts

VI. Project Implementation

VI.1. Preliminary Schedule

Month 1-3:

- Transformation of CATMA backend components to CLÉA Server components
- Development of basic Frontend support

Month 4-6:

- Development/integration of the User Management component
- Development of the Versioning component

[11] see http://www.jcmeister.de/downloads/texts/meister_schönert_2009.pdf

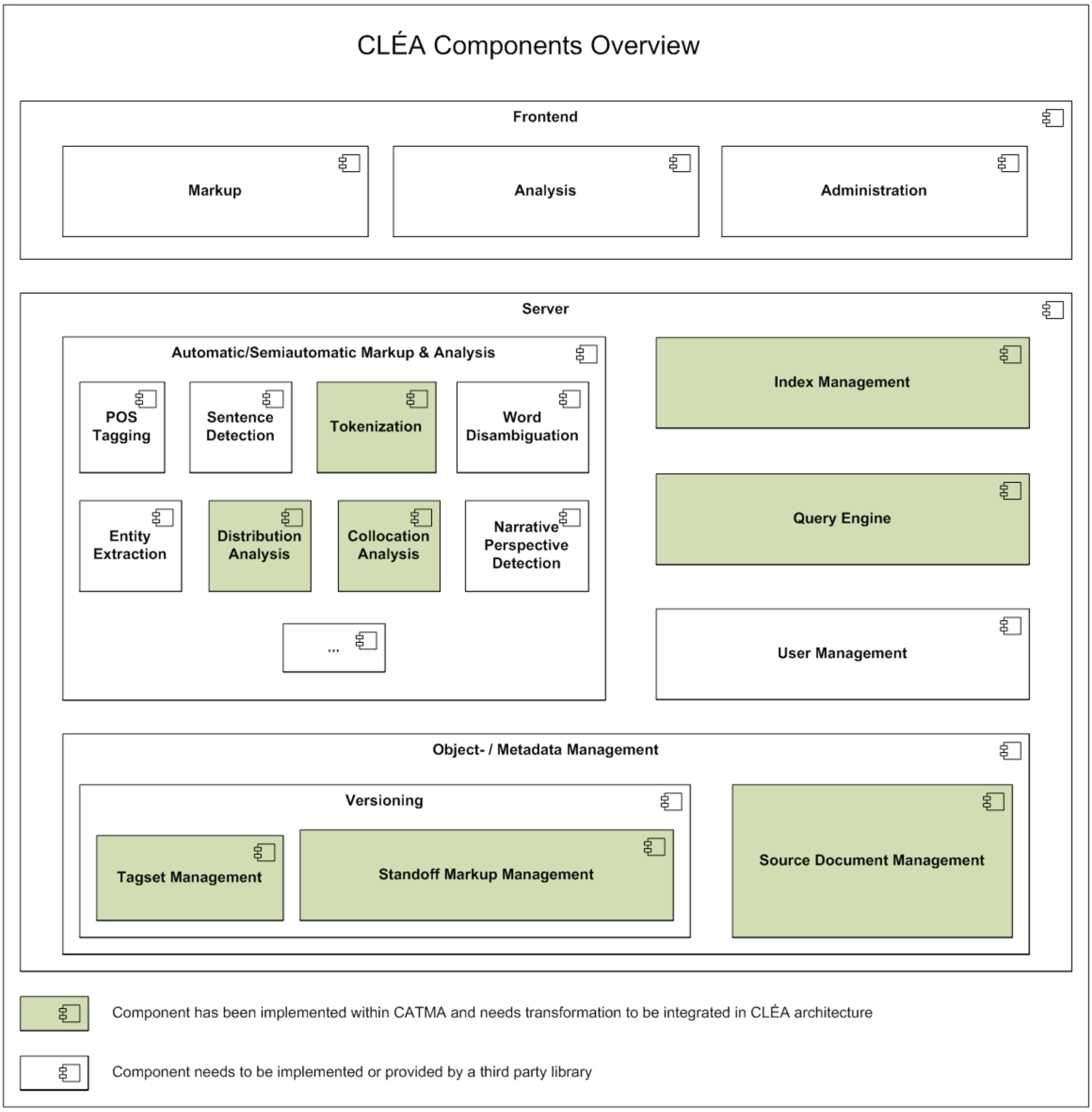
Month 7-9:

- Finalizing Frontend support
- Development/integration of missing subcomponents of the Automatic/Semiautomatic Markup & Analysis component
- Pilot analysis of routines for automated detection of narrative ‘perspective’

Month 10-12:

- Prototype implementation of automated detection of narrative ‘perspective’

VI.2. CLÉA Component Diagram


CLÉA Components Overview
Frontend
Markup
Analysis
Administration
Server
Automatic/Semiautomatic Markup & Analysis
POS Tagging
Sentence Detection
Tokenization
Word Disambiguation
Entity Extraction
Distribution Analysis
Collocation Analysis
Narrative Perspective Detection
...
Index Management
Query Engine
User Management
Object- / Metadata Management
Versioning
Tagset Management
Standoff Markup Management
Source Document Management
Component has been implemented within CATMA and needs transformation to be integrated in CLÉA architecture
Component needs to be implemented or provided by a third party library

VII. Bibliography

Ambati, V, Vogel, S & Carbonell, J 2010, 'Active Learning and Crowd-Sourcing for Machine Translation'. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10),* eds Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias, European Language Resources Association (ELRA), Valletta, Malta.

Argamon, S, Koppel, M, Fine, J & Shimoni, AR 2003, 'Gender, genre, and writing style in formal written texts', *TEXT*, vol. 23, pp. 321-346.

Bajwa, IS 2010, 'Context Based Meaning Extraction by Means of Markov Logic', *International Journal of Computer Theory and Engineering (IJCTE)*, vol. 2, no. 1, pp. 35–38. Available from: http://www.ijcte.org/papers/113-G609.pdf.

Bethard, S, Martin, JH & Klingenstein, S 2007, 'Timelines from Text: Identification of Syntactic Temporal Relations'. *ICSC '07: Proceedings of the International Conference on Semantic Computing,* IEEE Computer Society, Washington, DC, USA, pp. 11-18.

Boot, P 2009, 'Towards a TEI-based encoding scheme for the annotation of parallel texts', *Literary & [and] linguistic computing*, vol. 24, no. 3, pp. 347–361.

Buzzetti, D 2002, 'Digital Representation and the Text Model' in *New Literary History,* pp. 61–88.

C. Biemann, U. Quasthoff, G. Heyer & F. Holz 2008, 'ASV Toolbox - A Modular Collection of Language Exploration Tools'. *Proceedings of the 6th Language Resources and Evaluation Conference (LREC) 2008*. Available from: http://www.lrec-conf.org/proceedings/lrec2008/pdf/447_paper.pdf [28 October 2010].

Donmez, P, Carbonell, JG & Schneider, J 2010, 'A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy' in *Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA,* SIAM, pp. 826–837.

Elson, DK & McKeown, K 2010, 'Automatic Attribution of Quoted Speech in Literary Narrative' in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010,* eds Maria Fox & David Poole, AAAI Press.

Greevy, EP & Smeaton, AF 2004, 'Text Categorisation of Racist Texts Using a Support Vector Machine' in *Le poids des mots JADT. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles,* Louvain, pp. 533–544.

Holthausen, K, Onnasch, E & Ziche, P 2007, 'Dynamisierte Textcorpora. Anwendungen neuronaler Netze für editorische und texterschließende Fragestellungen', *editio. Internationales Jahrbuch für Editionswissenschaft*, vol. 21, pp. 169–188.

Juola, P & Bernola, A 2009, *Automatic Conjecture Generation in the Digital Humanities,* MITH, DH 2009 College Park, MD.

Kit, C, Pan, H & Webster, JJ 2002, 'Example-Based Machine Translation: A New Paradigm'. *Translation and Information Technology,* ed S.W. Chan, Chinese University of HK Press, pp. 57-78.

Lou Burnard 2001, 'On the hermeneutic implications of text encoding' in *New media and the humanities: research and applications,* Humanities Computing Unit, Oxford, pp. 31–38.

Pietz, W 2010, *Towards Hermeneutic Markup: An architectural outline,* King's College, DH 2010, London. Available from: http://piez.org/wendell/papers/dh2010/index.html [28 October 2010].

Schilder, F, Habel, C & Schilder, F 2001, 'From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages', *IN PROCEEDINGS OF THE ACL-2001 WORKSHOP ON TEMPORAL AND SPATIAL INFORMATION PROCESSING, ACL-2001. TOULOSE,* pp. 65–72.

Taboada, M, Brooke, J & Stede, M 2009, 'Genre-based paragraph classification for sentiment analysis'. *SIGDIAL '09: Proceedings of the SIGDIAL 2009 Conference,* Association for Computational Linguistics, Morristown, NJ, USA, pp. 62-70.

Vincent, L 2007, 'Google Book Search: Document understanding on a massive scale'. *In Proceedings, IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR'07).*