heureCLÉA
COLLABORATIVE LITERATURE EXPLORATION & ANNOTATION

Thomas Bögel, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris, Jannik Strötgen

Eine digitale Heuristik zur Unterstützung geisteswissenschaftlicher MarkUps

Projektkonzeption

Projektbeschreibung

Gegenstand des Vorhabens heureCLÉA ist die Entwicklung einer "digitalen Heuristik". Darunter verstehen wir ein Funktionsmodul, das Geisteswissenschaftler bei der Deutung und Annotation von Texten unterstützt, indem es ihnen MarkUp-Aufgaben erleichtert oder abnimmt. Exemplarisch widmen wir uns dazu der semantischen Erschließung von Zeit-Referenzen in erzählenden Texten.
Für die Entwicklung des Moduls sind drei Arbeitsschritte vorgesehen:

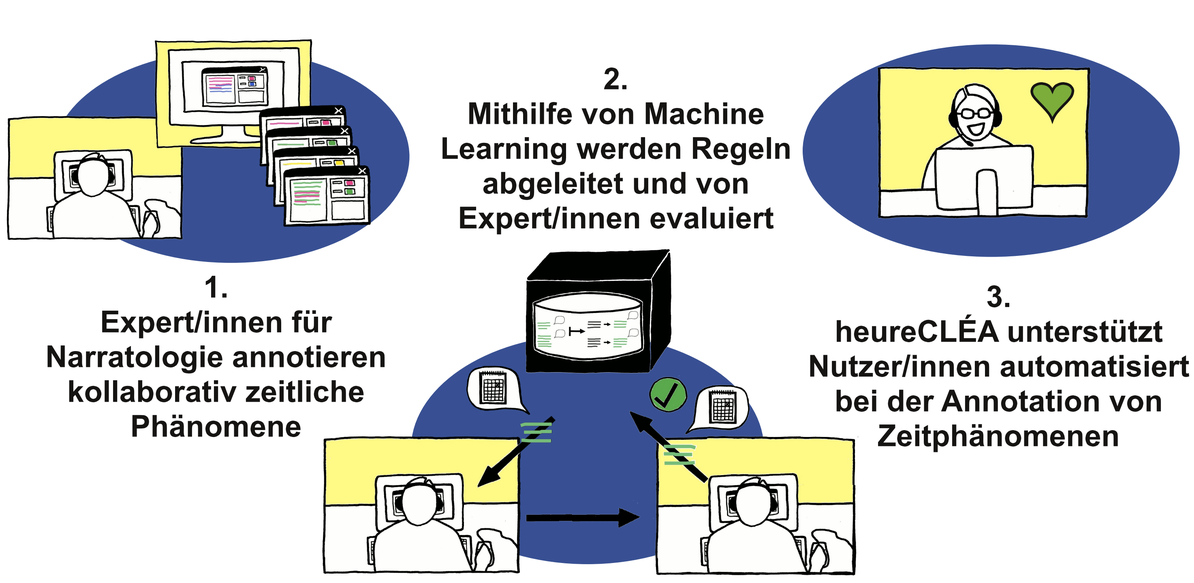
1. Automatisierung niedrigkomplexer MarkUp-Aufgaben,
2. Exploration und Analyse der automatisch generierten sowie komplexerer, manuell erstellter MarkUp-Varianten,
3. Computer-unterstützte Modellierung und Generierung von MarkUp-Varianten.

Entwickelt wird damit eine digitale Heuristik, die dem geisteswissenschaftlichen Nutzer automatische Deutungsangebote zur Verfügung stellt. Das Modul wird in die webbasierte Analyseplattform CATMA integriert, wo die MarkUp-Angebote vom Nutzer validiert werden können. So stehen auch nach Implementierung des Moduls weiterhin Korrekturmechanismen zur Verfügung.

Methode

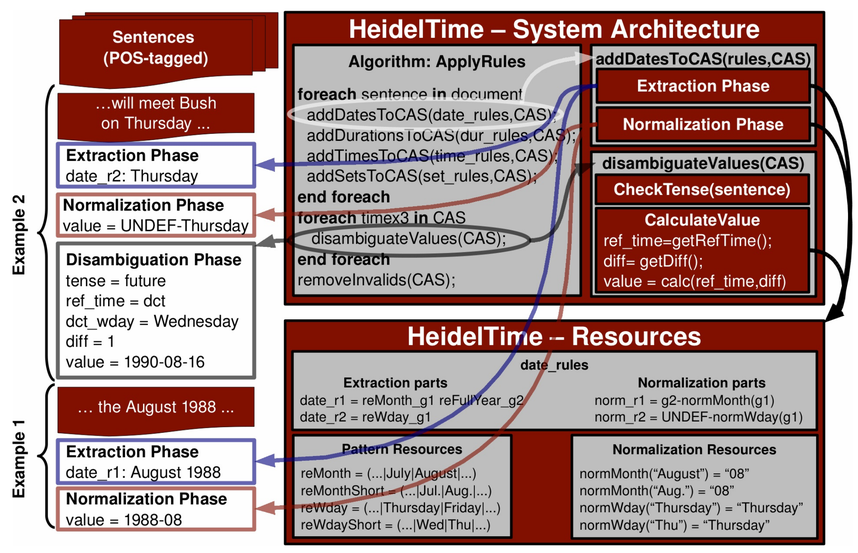
Die Taxonomie von Zeit-Phänomenen in Erzähltexten wird unter Rückgriff auf narratologische Kategorien erstellt [1]. Anschließend erfolgt das Tagging der Zeit-Repräsentationen und -Relationen in einem Korpus aus 21 literarischen Erzählungen durch:

- automatische Extraktion und Normalisierung von Zeitausdrücken durch das System HeidelTime [2] (sowie automatische Annotation von Hilfskategorien durch Morphisto und TreeTagger);
- kollaboratives manuelles Tagging gemäß der narratologischen Taxonomie mit der Webplattform CATMA [3] nach folgendem Schema:



Mithilfe von Machine Learning-Verfahren werden Regularitäten und Automatisierung einer Teilmenge der in CATMA manuell vorgenommenen MarkUp-Operationen ermittelt, wobei insbesondere Supervised Learning-Methoden basierend auf manuell erstellten und verifizierten Zeit-MarkUps verwendet werden. Anschließend werden Verfahren der Dependenzanalyse auf Textebene und der Korrelation von Zeitausdrücken in verschiedenen Satz- und Zeitkontexten für die Ableitung von Regeln und Umsetzung in HeidelTime eingesetzt.

Das Programmmodul heureCLÉA wird dann den Nutzern in der kollaborativen Arbeitsumgebung CATMA als "digitale Heuristik" zur partiell automatisierten, partiell interaktiven Generierung von Zeit-MarkUp bereitgestellt. Die Hauptkomponente des Moduls wird HeidelTime sein, das, basierend auf einer kontinuierlich erweiterten Regelbasis, zur automatischen Spezifikation von Zeit-MarkUps in Texten verwendet wird.



Systemarchitektur und Verarbeitungsprozesse in HeidelTime

Referenzen

[1] Narratologische Grundlagen sind u.a.:
- Genette, Gérard: Die Erzählung. München: Fink 1998.
- Lahn, Silke/Meister, Jan Christoph: Einführung in die Erzähltextanalyse. Stuttgart: Metzler 2013.
- Meister, Jan Christoph/Schernus, Wilhelm (Hrsg.): Time. From Concept to Narrative Construct. A Reader. Berlin, New York: de Gruyter 2011.

[2] HeidelTime: dbs.ifi.uni-heidelberg.de/heideltime
[3] CATMA: www.catma.de

Erste Ergebnisse in Auszügen

Automatisierung

Bis dato konnten bereits erste Ergebnisse bezüglich der Automatisierung grundlegender temporaler Phänomene (Tempus und Zeitausdrücke) erzielt werden. Eine erste Version des heuristischen Funktionsmoduls, das automatische Vorschläge zur Annotation der genannten Kategorien enthält, kann dem Nutzer deshalb frühzeitig zur Verfügung gestellt werden. Aktuell erfolgt die manuelle Annotation der komplexeren narratologischen Zeitphänomene (Ordnung, Frequenz und Dauer), die die Datenbasis für die weitere Automatisierung liefert. Beispielhaft sei hier eine Übersicht der Werte für die Automatisierung von Tempus vor Erstimplementierung der Funktionalität dargestellt:

Tempus	korrekt getaggte Verben
Präsens	93,10
Präteritum	95,73
Perfekt	96,43
Plusquamperfekt	84,71
Futur	90,00

Intradisziplinäre Erkenntnisse durch interdisziplinäres Arbeiten

Im Rahmen kollaborativer Projekte bleibt es nicht aus, dass sich Diskrepanzen zwischen theoretischen und praktischen Grundlagen der beteiligten Disziplinen auftun. Es ist eine besondere Herausforderung, derartige Probleme auf eine Weise zu lösen, die nicht nur die Zusammenarbeit der Disziplinen ermöglicht, sondern darüber hinaus auch intradisziplinären Gewinn mit sich bringt.

In heureCLÉA schienen zunächst die literaturwissenschaftliche Praxis und die informationstheoretischen Anforderungen an auszuwertendes Datenmaterial zu konfligieren: Während sich die literaturwissenschaftliche Praxis durch exemplarisches Arbeiten und durch Deutungspluralismus auszeichnet, fordert die Informatik eine Datengrundlage, die weder *sparse* (d.h. zu dünn) noch *noisy* (d.h. nicht eindeutig interpretierbar) ist.

Exemplarisch sei an dieser Stelle auf die besonders fruchtbaren Ergebnisse verwiesen, die die hierdurch angestoßenen Problemlösungsprozesse für die Narratologie liefern konnten: Durch das Bestreben, eine aus informationstheoretischer möglichst gut interpretierbare Datengrundlage zu liefern, wurden diskrepante manuelle Annotationsergebnisse vermehrt diskutiert. Diese Praxis offenbarte die Abhängigkeit einiger temporaler Phänomene von bestimmten theoretischen Grundannahmen. Diese Erkenntnis ermöglicht einen neuen Zugang zu Annotation und Automatisierung: Beides kann nur unter Berücksichtigung der relevanten theoretischen Grundannahmen erfolgen, was zu einer Parametrisierung der automatischen Annotationsvorschläge führt.

Projektlaufzeit: 02/2013 - 01/2016

Kontakt: info@heureclea.de

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Durchführende Forschungseinrichtungen:

Universität Hamburg
Fakultät für Geisteswissenschaften
Interdisziplinäres Centrum für Narratologie (ICN)
Prof. Dr. Jan Christoph Meister

Ruprecht-Karls-Universität Heidelberg
Fakultät für Mathematik und Informatik
Lehrstuhl für Datenbanksysteme
Prof. Dr. Michael Gertz