

Technischer Bericht Nr. 2009-01

GSCL-SYMPOSIUM
SPRACHTECHNOLOGIE UND eHUMANITIES

Wolfgang Hoepfner (Hrsg.)

14.02.2009

ISSN 1863-8554

IMPRESSUM:

Technische Berichte der Abteilung für Informatik und Angewandte
Kognitionswissenschaft, Universität Duisburg-Essen

ISSN 1863-8554

Herausgeber:

Abteilung für Informatik und Angewandte Kognitionswissenschaft
Fakultät für Ingenieurwissenschaften
Universität Duisburg-Essen
Campus Duisburg
47048 Duisburg

<http://duepublico.uni-duisburg-essen.de/informatik/berichte.xml>

Symposium
**»Sprachtechnologie
und eHumanities«**

26.02.2009 - 27.02.2009

UNIVERSITÄT DUISBURG-ESSEN, CAMPUS DUISBURG

G S C L

Gesellschaft für
Sprachtechnologie
& Computerlinguistik

PROGRAMMKOMITEE UND ORGANISATION

Wolfgang Hoepfner
Angela Klutsch
Marc Lechtenfeld
Nino Simunic

Universität Duisburg-Essen
Universität Duisburg-Essen
Universität Duisburg-Essen
Universität Duisburg-Essen

INHALTSVERZEICHNIS

D. BAUM, B. SAMLOWSKI, T. WINKLER, R. BARDELI, D. SCHNEIDER DiSCo – A Speaker and Speech Recognition Evaluation Corpus for Challenging Problems in the Broadcast Domain	1
G. BÜCHEL Datenbank- und XML-Technologien im Projekt NAPROCHE.....	10
M. BURGHARDT, C. WOLFF Werkzeuge zur Annotation diachroner Textkorpora	21
F. FRITZINGER, M. KISSELEW, U. HEID, A. MADSAK, H. SCHMID Werkzeuge zur Extraktion von signifikanten Wortpaaren als Web Service	32
K. IGNATOVA, C. TOPRAK, D. BERNHARD, I. GUREVYCH Annotation Question Types in Social Q&A Sites.....	44
F. JUNGERMANN Information Extraction with RapidMiner	50
P. KOLB, AMELIE KUTTER, CATHLEEN KANTNER, MANFRED STEDE Computer- und korpuslinguistische Verfahren für die Analyse massenmedialer politischer Kommunikation: Humanitäre und militärische Interventionen im Spiegel der Presse	62
A. MEHLER, R. GLEIM, U. WALTINGER, A. ERNST, D. ESCH, T. FEITH eHumanities Desktop – eine webbasierte Arbeitsumgebung für die geisteswissenschaftliche Fachinformatik	72
M. SCHEFFEL Relationserkennung auf deutschen Fließtexten.....	91
C. TOPRAK, C. MÜLLER, I. GUREVYCH Extracting Professional Preferences of Users from Natural Language Essays	103
C. VERTAN Multilinguality in an on-line platform for classical philology – beyond localisations of the user-interface.....	111
B. WAGNER, A. MEHLER, C. WOLFF, B. DOTZLER Bausteine eines Literary Memory Information System (LiMeS) am Beispiel der Kafka-Forschung	119

Vorwort

In den 90er Jahren des letzten Jahrhunderts war ich mehrere Jahre Dekan des Fachbereichs ‚Literatur- und Sprachwissenschaften‘ an der Gerhard-Mercator Universität Duisburg, die dann 2003 mit der Universität Essen fusioniert wurde und dabei ihren Namenspatron eingebüßt hat.

Damals gab es auf einmal größere Geldtöpfe, die für die informationstechnische Ausstattung der Fachbereiche genutzt werden sollten. In meinem damaligen Fachbereich habe ich natürlich die Professoren angesprochen, ob sie sich nicht für derlei Dinge interessieren könnten. Das Ergebnis war niederschmetternd: die meisten haben gar nicht reagiert, und auf persönliche Befragung traten dann Meinungen zutage, die in Computern ein Teufelswerkzeug sahen, das ihre wissenschaftliche Arbeit verderben würde und Maschinen zu ästhetisch urteilenden Instrumenten missbrauchen würde. Dies war insbesondere in der Literaturwissenschaft so; die Linguisten waren da schon zugänglicher, vermutlich weil sie mit formalen Methoden besser vertraut waren.

Diese Zeiten haben sich geändert. Auch Literaturwissenschaftler haben gelernt, das Internet zu nutzen, und schreiben beispielsweise gerne E-Mails an Fachkollegen. Linguisten nutzen weltweit verfügbare Korpora aus und nehmen diese als Grundlage für ihre Forschungen.

In den letzten Jahren hat die Informationstechnologie die Geisteswissenschaften entdeckt, möglicherweise kann man auch sagen, die Geisteswissenschaften hätten die Informationstechnologie schätzen gelernt. Wie auch immer: es gibt zahlreiche Förderungsinitiativen in diesem Bereich. Und dies hängt auch mit den Digitalisierungsinitiativen der letzten Jahrzehnte zusammen. Je mehr digitalisierte Korpora existieren, desto mehr bisher unbekannt oder ungenutzte Quellen sind ortsunabhängig verfügbar. Die Frage stellt sich jetzt, wie diese Datenmassen einer Wissenschaftsgemeinschaft verfügbar gemacht werden können? Als ein Beispiel soll hier das kürzlich abgeschlossene

BMBF-Projekt Wikinger erwähnt werden¹, an dem die Fraunhofer Gesellschaft (St. Augustin), die Kommission für Zeitgeschichte (Bonn) und die Duisburger Computerlinguistik gemeinsam gearbeitet haben. In diesem Projekt wurde ein System für die semantische Annotation von Texten zum deutschen Katholizismus der letzten 200 Jahre entwickelt.

Das BMBF hat vor mehreren Jahren ein Sonderprogramm unter dem Namen eScience aufgelegt (<http://www.bmbf.de/de/298.php>), bei dem es um die informationstechnologische Unterstützung zahlreicher Wissenschaften ging; die Geisteswissenschaften waren dort ein Element, wenn auch noch kein sonderlich prominentes, aber das oben erwähnte Projekt Wikinger und auch TextGrid (<http://www.textgrid.de/>) gehören dazu. Im Rahmen der europäischen Projekte eContent^{plus}² und Clarin (<http://www.clarin.eu/>) sind diese Aktivitäten ausgeweitet worden, die deutsche Variante ist die D-SPIN Initiative (<http://www.sfs.uni-tuebingen.de/dspin/>). Im Jahr 2007 hat das BMBF außerdem ein Programm „Wechselwirkungen zwischen Natur- und Geisteswissenschaften“ aufgelegt mit den Pilotanwendungen ‚Archäologie und Altertumswissenschaften‘ sowie ‚Sprach- und Literaturwissenschaften‘ (<http://www.bmbf.de/foerderungen/7774.php>), in dem derzeit 16 Projekte gefördert werden. Die in Deutschland wichtigsten Wissenschaftsorganisationen haben eine Schwerpunktinitiative „Digitale Information“³ herausgegeben. Zurzeit sind außerdem Aktivitäten bei der DFG für ein Schwerpunktprogramm ‚Digital Humanities‘ begonnen worden.

Es tut sich also etwas bei der Annäherung zwischen den Geisteswissenschaften und der Informationstechnologie, und das ist für beide Seiten gut. Das Duisburger Symposium ‚Sprachtechnologie und eHumanities‘ bietet ein breites und interdisziplinäres Forum, auf dem die Verflechtungen zwischen den Geisteswissenschaften und der Informatik diskutiert werden sollen.

¹ Lars Bröcker, Stefan Paal, Andreas Burtscheidt, Bernhard Frings, Marc Rössler, Andreas Wagner, Wolfgang Hoepfner. "WIKINGER - Wiki Next Generation Enhanced Repositories". German e-Science Conference 2007. Baden-Baden, Germany, 2007 (<http://www.ges2007.de>).

² http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm

³ http://www.dfg.de/aktuelles_presse/das_neueste/download/pm_allianz_digitale_information_details_080612.pdf

Werkzeuge zur Annotation diachroner Textkorpora

Manuel Burghardt, Christian Wolff
Institut für Information und Medien, Sprache und Kultur
Universität Regensburg
93040 Regensburg
{manuel.burghardt, christian.wolff}@sprachlit.uni-regensburg.de

Keywords: Annotationswerkzeuge, Evaluation, diachrone Korpora, historische Sprachwissenschaft, Benutzerfreundlichkeit, Funktionalität, ISO 9126, ISO 25000

Abstract

We discuss the problem of annotating syntax in diachronic corpora. A study analysing functionality as well as usability characteristics of more than 50 annotation tools is presented. For this study, we have developed a quality model based on international standards ISO/IEC 9126-1:2001 (*Software engineering – Product quality – Part 1: Quality model*) and ISO/IEC 25000:2005 (*Software Engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE*).

Zusammenfassung

Wir diskutieren zunächst die Problematik der (syntaktischen) Annotation diachroner Korpora und stellen anschließend eine Evaluationsstudie vor, bei der mehr als 50 Annotationswerkzeuge und -frameworks vor dem Hintergrund eines funktionalen und software-ergonomischen Anforderungsprofils nach dem Qualitätsmodell von ISO/IEC 9126-1:2001 (*Software engineering – Product quality – Part 1: Quality model*) und ISO/IEC 25000:2005 (*Software Engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE*) evaluiert wurden.

1 Diachrone Korpora als Herausforderung der Texttechnologie

Obwohl die Korpuslinguistik mittlerweile zum Standardrepertoire sprachwissenschaftlicher Methodik gehört (Tognini-Bonelli 2001), stellen diachrone Korpusanalysen immer noch eine Herausforderung für die menschlichen Annotatoren dar: In der Regel liegt mit steigendem Alter der Sprachentwicklungsstufe nur wenig Datenmaterial vor und Ergebnisse (sowie Analyseverfahren) synchroner Sprachbetrachtung lassen sich nicht ohne Weiteres auf ältere Sprachstufen übertragen, da bei diachronen Ansätzen in

erster Linie Wandelprozesse auf den verschiedenen sprachlichen Beschreibungsebenen (z.B. Morphologie, Phonetik, Syntax oder Semantik) entlang einer Zeitachse untersucht werden. Die explizite Annotation bestimmter sprachlicher Parameter soll dabei Entwicklungstendenzen aufdecken helfen. Problematisch sind hierbei vor allem orthografische und syntaktische Ambiguitäten, welche eine eindeutige Annotation und spätere Auswertung der Daten erschweren. Doch gerade die explizite Annotation einer Mehrzahl von Lesarten kann dabei helfen, Sprachwandelprozesse zu beschreiben. Zudem sind diachrone Annotationen zwangsläufig durch ein hohes Maß an Diskontinuität gekennzeichnet (Kroymann et al. 2004), da Texte unterschiedlicher Sprachstufen oftmals nicht direkt miteinander vergleichbar sind. Der Grund hierfür liegt in der unterschiedlich starken Ausprägung bestimmter sprachlicher Parameter wie etwa Textgattung oder syntaktische Funktion. Das Aufzeigen solcher diskontinuierlichen Entwicklungen liefert neben der Annotation von Ambiguitäten wichtige Erkenntnisse über diachrone Sprachwandelprozesse. Um diese komplexen und kontextuell schwer zu bewertenden sprachlichen Phänomene korrekt annotieren zu können, ist bei der Erstellung diachroner Korpora immer noch ein hohes Maß an manueller Arbeit nötig. Aufgabe eines geeigneten Annotationswerkzeugs muss es somit sein, den Menschen bei seiner intelligenten und kreativen Arbeit so gut wie möglich zu unterstützen und ihn auf lange Sicht zu entlasten. Dabei sollte ein Annotationswerkzeug einerseits funktional sein, und dem Benutzer immer wiederkehrende Routineaufgaben durch Automatisierungsmechanismen abnehmen, andererseits muss ein entsprechendes Tool Anforderungen an Benutzbarkeit und Softwareergonomie erfüllen. Inwiefern die derzeit verfügbaren Werkzeuge den Anforderungen einer solchen „Annotationsergonomie“ gerecht werden, soll im Folgenden untersucht werden.

2 Auswahlkriterien Werkzeuge zur Annotation diachroner Korpora

Annotation wurde bereits lange vor Anbruch des Computerzeitalters betrieben und ist seit jeher ein wichtiges Instrument um Wissen zu akkumulieren, es zu verwalten und anderen besser zugänglich zu machen. Seitdem haben sich im Bereich der Annotationspraxis zahlreiche computergestützte Techniken und Vorgehensweisen etabliert, die den menschlichen Annotator bei seiner anspruchsvollen Aufgabe unterstützen. Vor diesem Hintergrund hat sich ein breites Spektrum an Annotationswerkzeugen etabliert, das ebenso heterogen und vielschichtig ist wie die Zahl denkbarer Annotationsszenarien selbst (Ide & Brew 2000). In unserer Studie haben wir durch Systematisierung bestehender Tools sowie durch Berücksichtigung der Anforderungen, die sich aus einem diachronen Annotationszenario ergeben, geeignete Werkzeuge für diachrone Korpora identifiziert und evaluiert.

Ein zweistufiger Selektionsprozess dient dabei der Reduktion von mehr als 50 nachweisbaren Werkzeugen und Frameworks auf eine handhabbare Menge: Die wesentlichen Kriterien für den ersten Schritt der Systematisierung der über 50 Annotationswerkzeuge sind dabei „Annotationsmodalität“ und „Softwaretyp“ (Wolff 2004). Obwohl die Verfügbarkeit von digitalisierten Sprach- und Videodaten in den letzten Jahren stark zugenommen hat (Dybkjær et al. 2001), sind solche multimedialen Datensätze für fundierte diachrone Untersuchungen, die sich meist auf einer mehrere Jahrhunderte umfassenden Zeitachse bewegen, (noch) nicht von allzu großer Relevanz. Tools, die für diachrone Korpusanalysen eingesetzt werden sollen, müssen in jedem Fall die Annotationsmodalität *Text* unterstützen. Beim *Softwaretyp* sind „fertig implementierte“ und sofort einsetzbare Programme mit grafischer Oberfläche (i. d. R. *monolithische* Softwaresysteme bzw. „rich clients“) abstrakten Klassenbibliotheken oder komplexen Frameworks vorzuziehen, da der Annotator möglichst effizient bei seiner Aufgabe unterstützt werden soll, in den meisten Annotationsszenarien jedoch weder die Zeit noch das technische *know how* zur aufwendigen Konfiguration oder Erstimplementierung eines Werkzeugs vorhanden ist (Dipper et al. 2004). Tabelle 1 gibt eine Übersicht der in der Studie untersuchten Annotationswerkzeuge:

Toolname	Softwaretyp
ACE Annotation Toolkit	Annotationswerkzeug (basiert auf dem AGTK)
ACT	Annotationswerkzeug
AGTK	Framework
Alembic Workbench	Framework und Annotationswerkzeug
Annotate	Annotationswerkzeug
Anvil	Annotationswerkzeug
Arboreal	Annotationswerkzeug, XML-Browser
ATLAS	Framework
CAVA	Annotationswerkzeuge
Callisto	Annotationswerkzeug (basiert auf jATLAS, Nachfolger der Alembic Workbench)
CBAS	Annotationswerkzeug
CLAN	Annotationswerkzeug für Texte eines bestimmten Formats (CHILDES), Analysetool

Toolname	Softwaretyp
CLaRK	Annotationswerkzeug, Lexikonerstellung
CSLU Toolkit	Framework, Annotationswerkzeug, Analyse-tool, TTS, Sprachtrainer
DAT	Annotationswerkzeug (benutzt das DAMSL ₁₁ Schema)
Dexter	Annotationswerkzeug
DitAT	Annotationswerkzeug
ELAN	Annotationswerkzeug
EUDICO	Framework, Workbench (Integration in GATE geplant)
EXMARaLDA	Annotationswerkzeug, Korpusmanager, Analysetool
FLEX	Annotationswerkzeug für Feldforschung, Lexikonerstellung
GATE	Framework und Annotationswerkzeug

Toolname	Softwaretyp
Interact	Annotationswerkzeug
ITE	Annotationswerkzeug
LT XML	Framework, Greptool
MATE	Framework
MediaStreams	Ikonisches Annotationswerkzeug
MMAX	Annotationswerkzeug
MMAX 2	Annotationswerkzeug (Nachfolger von MMAX)
Multext Tools	Annotationswerkzeug
MultiTool	Annotationswerkzeug, Analysetool
NITE (NXT)	Framework (Nachfolger von MATE)
Observer	Annotationswerkzeug
oXygen	XML Annotationswerkzeug
Palinka	Annotationswerkzeug (Nachfolger von Clinka)
Praat	Annotationswerkzeug, Analysetool, TTS
RST Tool	Annotationswerkzeug

Toolname	Softwaretyp
SignStream	Annotationswerkzeug, Analysetool
SmartKom	Framework (benutzen Anvil zur Annotation)
Snack	Framework
SyncWriter	Annotationswerkzeug
Synpathy	Annotationswerkzeug
Systemic Coder	Annotationswerkzeug, Analysetool
TASX	Framework und Annotationswerkzeug
Toolbox	Annotationswerkzeug für Feldforschung, Lexikonerstellung (Nachfolger von Shoebox)
Transcriber	Annotationswerkzeug
Transformer	Annotationswerkzeug
vPrism	Annotationswerkzeug
WaveSurfer	Annotationswerkzeug
UAM CorpusTool	Annotationswerkzeug (Nachfolger von Systemic Coder), Analysetool
Wordfreak	Annotationswerkzeug

Tabelle 1: Übersicht der in der Evaluationsstudie berücksichtigten Annotationswerkzeuge

Aus Darstellungsgründen sind in der Tabelle keine Links zu den Werkzeugen aufgeführt; diese sind aber unter <http://www.disynde.de> verfügbar, wo auch weiteres Dokumentationsmaterial zu den Tools vorhanden ist. Erfüllen die Annotationswerkzeuge im zweiten Schritt der Vorauswahl neben der geforderten Annotationsmodalität und einem entsprechenden Softwaretyp auch noch die Anforderungen *Verfügbarkeit und Aktualität der Applikation*, *Flexibilität der Annotationsschemata* sowie *Wiederverwendbarkeit des Annotationsformats*, so werden sie hinsichtlich ihrer Funktionalität und Benutzbarkeit ausführlich evaluiert. Nicht-funktionale Kriterien wie *Aktualität* und *Verfügbarkeit* einer Anwendung erscheinen als Auswahlkriterium gerechtfertigt, da Korpusaufbereitung typischerweise ein länger wählender Prozess ist, bei dem die Verfügbarkeit und Unterstützung der Werkzeugumgebung eine große Rolle spielt. Bei der Evaluation gilt es, zahlreiche weitere Anforderungen wie etwa *Zeichensatz*, *Mehrebenenannotation*, *Flexibilität der Ein- und Ausgabe*, *Adaptierbarkeit der Software*, *Automatisierbarkeit der Software* sowie die *Koordination verteilter Arbeitsabläufe* (Workflow) durch ein adäquates Qualitätsmodell zu operationalisieren (Burghardt 2008). Insgesamt qualifizieren sich vier Werkzeuge für eine ausführliche Softwareevaluation: Das auf dem ATLAS-

Framework basierende Tool *Callisto*, das Framework *GATE*, das auch ein Annotationswerkzeug beinhaltet, sowie die von kleineren Teams entwickelten Programme *MMAX2* und das *UAMCorpusTool*, das Ansätze aus der systemisch-funktionalen Linguistik (Halliday & Martin 1981) aufgreift.

3 Aufbau eines standardisierten Qualitätsmodells

Die Qualitätsnormen zur Sicherung der Produktqualität von Software der ISO (*International Organization for Standardization*) und der IEC (*International Electrotechnical Commission*) (DIN 66272, 1994) sowie das Framework zur Evaluation von VNS-Software (*Software zur Verarbeitung Natürlicher Sprache*) der EAGLES-Evaluationsarbeitsgruppen (EAGLES 1999a, 1999b) dienen als Grundlage für ein weitestgehend standardisiertes und wieder verwendbares Evaluationsdesign. Dabei werden alle im vorangegangenen Kapitel allgemein formulierten Anforderungen in die beiden primären Qualitätskriterien der ISO 9126, *Funktionalität* und *Benutzbarkeit*, sowie deren Unterkriterien *Angemessenheit*, *Interoperabilität*, *Erlernbarkeit*, *Bedienbarkeit* und *Konformität* gegliedert und solange weiter aufgeteilt, bis messbare Attribute übrig bleiben. Das Ergebnis ist ein hierarchisches Qualitätsmodell, auf dessen höchster Ebene das Annotationswerkzeug als Ganzes steht und auf dessen unterster Ebene sich ein Katalog aus messbaren Attributen befindet.

Softwareprodukt (operationalisiert durch zwei Qualitätskriterien mit insgesamt fünf Unterkriterien und 30 Attributen)				
Funktionalität (operationalisiert durch zwei Unterkriterien mit insgesamt 13 Attributen)		Benutzbarkeit (operationalisiert durch drei Unterkriterien mit insgesamt 17 Attributen)		
Angemessenheit	Interoperabilität	Erlernbarkeit	Bedienbarkeit	Konformität
9 Attribute	4 Attribute	5 Attribute	8 Attribute	4 Attribute

Tabelle 2: Hierarchisches Qualitätsmodell nach ISO 9126

Die Funktionalität von Annotationswerkzeugen beschreibt, in welchem Maße Funktionen zur Erfüllung einer bestimmten Aufgabe durch einen Benutzer vorhanden sind. Damit beschreibt Funktionalität das *Verhältnis zwischen Werkzeug und Aufgabe*, während Benutzbarkeit das *Verhältnis zwischen Werkzeug und Benutzer* thematisiert. Anders als beim Kriterium der Funktionalität steht bei der Benutzbarkeit der Anwender mit seinen individuellen Bedürfnissen an Interaktionsverhalten und Visualisierung der Software im Vordergrund. Ein hoher Grad an Benutzbarkeit impliziert immer auch einen möglichst geringen Aufwand zum Erlernen und zur Bedienung der Software. Den Zusammenhang von Aufgabe, Werkzeug und Benutzer veranschaulicht Abbildung 1:

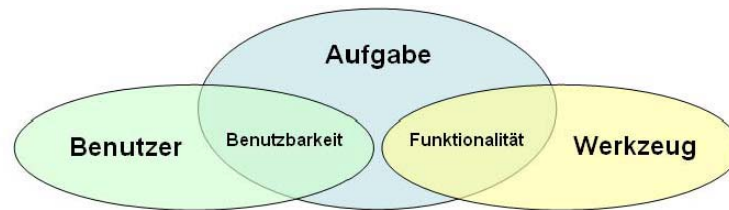


Abbildung 1: Benutzbarkeit und Funktionalität im Verhältnis zur Aufgabe

Funktionalität (I): Angemessenheit: Die Angemessenheit von Annotationswerkzeugen beschreibt deren funktionelle Eignung hinsichtlich einer bestimmten Aufgabe. Potenzielle Aufgaben, die ein angemessenes Werkzeug erfüllen muss, sind die korrekte Darstellung und Verarbeitung der Primärdaten, also der Rohtexte, und der Sekundärdaten, also der Annotationen, im Unicode-format. Außerdem sollte ein Werkzeug in der Lage sein, über einen Tokenizer grundlegende Annotationseinheiten wie Wörter und Sätze automatisch zu erkennen. Die teilweise oder vollständige Automatisierung von immer wiederkehrenden Routineaufgaben und Annotationsphänomenen durch die Verwendung von Lexika oder regulären Ausdrücken sowie die Möglichkeit der Annotation von Relationen und Referenzen zwischen einzelnen Annotationseinheiten gehören ebenso zum Funktionsumfang eines angemessenen Annotationswerkzeugs für diachrone Sprachdaten wie die Editierbarkeit der verwendeten Annotationsschemata. Konsistenzprüfungen und Validierung der Annotation gegen das zugrunde gelegte Schema garantieren eine einheitliche Annotation.

Funktionalität (II): Interoperabilität: Mit der Interoperabilität von Annotationswerkzeugen wird die Kompatibilität der Software zu anderen bestehenden Annotations- und Korpusprogrammen beschrieben. Dabei spielen sowohl umfangreiche Import- und Exportfunktionen als auch ein wohlgeformtes und standardisiertes XML *stand-off* Format (Dipper 2005) eine entscheidende Rolle. Idealerweise setzt das Annotationsformat auf existierenden Metastandards wie etwa dem *Syntactic Annotation Framework* (SynAF) auf (Trippel, Declerck & Heid 2005, Heilemann 2008).

Benutzbarkeit (I): Erlernbarkeit: Das Unterkriterium Erlernbarkeit beschreibt wie aufwendig es für einen Benutzer ist, sich Kenntnisse über die Funktionsweise und Bedienbarkeit der Annotationssoftware anzueignen. Dabei tragen Programmdokumentation und Tutorials zur *Lernförderlichkeit* des Werkzeugs bei, integrierte Hilfesysteme und Features wie interaktive *Tooltips* steigern die *Selbstbeschreibungsfähigkeit* des Programms. Wenn darüber hinaus noch zusätzliche Problemlösungsstrategien wie eine Mailing-Liste oder FAQs (*frequently asked questions*) angeboten werden, unterstützt das Annotationswerkzeug die *Erlernbarkeit* auf ideale Weise.

Benutzbarkeit (II): Bedienbarkeit: Ein gut bedienbares Annotationswerkzeug sollte ohne übermäßigen kognitiven Aufwand schnell und effizient gehandhabt werden können. *Bedienbarkeit* erstreckt sich von der ersten Installation und Konfiguration der Software, über die Steuerung des Programmablaufs bis zur Modifizierung der Annotationsschemata. Die Bedienbarkeit eines Tools wird durch eine angemessene Visualisierung und ein benutzerfreundliches Interaktionsdesign gesteigert. Im Idealfall können sowohl die Visualisierung als auch das Interaktionsverhalten individuell an den jeweiligen Benutzer angepasst werden.

Benutzbarkeit (III): Konformität: Konformität beschreibt, inwieweit gängige und bekannte Visualisierungs- und Interaktionskonzepte aus dem Bereich der Softwareergonomie auch im zu evaluierenden Annotationswerkzeug umgesetzt werden. Dies betrifft vor allem die Aufteilung und Gestaltung der Arbeitsfläche sowie konventionalisierte Metaphern und Funktionen zur Bearbeitung von Textdaten, wie etwa *Copy/Paste* oder *Undo/Redo*.

4 Metriken: Attribute, Werteskalen und Bewertungsregeln

Für insgesamt 30 Einzelattribute des standardisierten Qualitätsmodells werden Metriken eingesetzt, die es erlauben, dem jeweiligen Qualitätsmerkmal konkrete Werte zuzuordnen. Auf diese Weise können unterschiedliche Werkzeuge anhand ihrer individuellen Werteausprägungen verglichen und bewertet werden. Eine Metrik besteht dabei aus einem Maß und einer Messmethode, wobei die Attribute und alle potenziell zu erwartenden Werte das Maß darstellen. Die Messmethode dient dazu, einen konkreten Wert auf der Werteskala für ein Attribut zu bestimmen. Über Auswahlregeln kann definiert werden, ob mehrere Werte einer Werteskala für ein Attribut kombiniert werden dürfen. Bei Durchführung der Evaluation kann dann für jedes definierte Attribut ein konkreter Wert auf der vorgegebenen Werteskala ermittelt werden. Noch vor der Evaluationsdurchführung muss in so genannten Einstufungsniveaus festgelegt werden, welche Werte im akzeptablen, und welche Werte im nicht mehr akzeptablen Bereich liegen. Durch Aufrechnen der einzelnen Attribute lässt sich ermitteln, zu welchem Grad bestimmte Kriterien oder Subkriterien erfüllt sind. Zudem ist es möglich, Attribute unterschiedlich stark zu gewichten. Tabelle 3 zeigt messbare Attribute und entsprechende Messmethoden für ein Qualitätsmodell unter Berücksichtigung der Qualitätskriterien *Funktionalität* und *Benutzbarkeit*, welches schließlich für die Evaluation von Annotationswerkzeugen für diachrone Korpora verwendet werden soll.

<i>Funktionalität (I): Angemessenheit</i>	<i>Messmethode / Messfrage</i>
(01) Unicode Unterstützung	Anzeige und Speicherung als UTF-8?
(02) Alternative Zeichenkodierung	Anzeige und Speicherung als ASCII, ISO-8859 etc.?
(03) Unterstützte Dateiformate	Formate: *.txt, *.doc, *.xml, *.html, *.sgml etc.?
(04) Modifizierbarkeit der Rohtexte	Modifizierung vor und während der Annotation?
(05) Tokenisierung der Rohtexte	Interne oder externe Tokenisierung mit Möglichkeit der Parametrisierung?
(06) Automatisierbarkeit der Annotation	Automatisierbarkeit durch Lexika oder RegEx?
(07) Skopus der Annotationseinheiten	Singuläre Annotationseinheiten oder gerichtete Relationen zwischen den Einheiten?
(08) Änderung der Schemata	Nachträgliche Einfügung, Änderung und Löschung von Tags?
(09) Validierung gegen Schema	Möglichkeit der internen Validierung?
<i>Funktionalität (II): Interoperabilität</i>	<i>Messmethode / Messfrage</i>
(10) Im-/Export von Stand-off Formaten	Im-/Exportkompatibilität zu anderen Stand-off Formaten?
(11) Im-/Export Inline-Formaten	Im-/Exportkompatibilität zu anderen Inline-Formaten?
(12) Wohlgeformtheit der Annotation	Wohlgeformtheit des XML-basierten Annotationsformats nach Definition des W3C?
(13) Kompatibilität des Annotationsformats	Kompatibilität des Annotationsformats mit mindestens einem anderen gängigen Annotationsformat?
<i>Benutzbarkeit (I): Erlernbarkeit</i>	<i>Messmethode / Messfrage</i>
(14) Dokumentation	Umfang und Qualität der Dokumentation?
(15) Hilfesystem	Umfang und Qualität des Hilfesystems?
(16) Kurzinfo	Umfang und Qualität der Kurzinfos?
(17) Praktische Lernhilfen	Umfang der Lernhilfen in Form von Videos, Tutorials, Übungsdateien, Sekundärliteratur, etc.?
(18) Problemlösungsstrategien	Vorhandensein von zusätzlichen Problemlösungsstrategien wie etwa FAQs, Mailinglists, etc.?
<i>Benutzbarkeit (II): Bedienbarkeit</i>	<i>Messmethode / Messfrage</i>
(19) Installationsaufwand	Technischer Aufwand zur Installation der Software?
(20) Ebenenbezogene Aktionen	Löschen, Ein- und Ausblenden von Ebenen möglich?
(21) Editierung der Schemata	Technischer Aufwand bei der Editierung von Schemata?
(22) Flexibilität der Schemata	Flexibilität der Schemata bei unvorhergesehenen Annotationsphänomenen?
(23) Zugriffsrechte für Annotationsebenen	Vergabe von Zugriffsrechten für unterschiedliche Benutzer?
(24) Suchfunktion für Textdaten	Umfang und Qualität der Textsuchfunktion?
(25) Anpassung der Darstellung	Flexibilität der Darstellungsanpassung durch Skins, etc.?
(26) Anpassung der Funktionalität	Flexibilität der Funktionalitätsanpassung durch Shortcuts, etc.?

<i>Benutzbarkeit (III): Konformität</i>	<i>Messmethode / Messfrage</i>
(27) Aufteilung der Arbeitsfläche	Aufteilung der Arbeitsfläche in <i>tiled panes, one-widow paging</i> oder <i>multiple windows</i> (Tidwell 2006) ?
(28) Gestaltung der Arbeitsfläche	Stimmigkeit der Farbpalette und Größenverhältnisse sowie Aussagekraft von Piktogrammen und Ähnlichem?
(29) Standardisierte Aktionen	Implementierung von <i>Undo, Redo, Autosave, etc?</i>
(30) Selektion von Annotationseinheiten	Auswahl der Annotationseinheiten durch Doppelklick, Klicken-Markieren-Loslassen, Richtungstasten, etc.?

Tabelle 3: Übersicht zu grundlegenden Attributen und Metriken der in der Evaluationsstudie

Aus Platzgründen können an dieser Stelle die Werteskalen ebenso wenig wie die Gewichtungsregeln aufgeführt werden, vgl. dazu ausführlich Burghardt 2008. Prinzipiell können die Attribute auch für andere Evaluationsstudien im Bereich (diachroner) Annotation herangezogen, und je nach Szenario und Anforderungsprofil mit individuellen Skalen und Gewichtungen versehen werden.

5 Interpretation der Ergebnisse und Ausblick

Bei der Auswertung der Evaluationsergebnisse wird deutlich, dass jedes der vier Tools individuelle Stärken und Schwächen aufweist, insgesamt jedoch das GATE Framework durchweg die besten Ergebnisse erzielt. Unter Berücksichtigung der unterschiedlichen Gewichtung von obligatorischen und optionalen Attributen, und durch Aufrechnung der jeweils akzeptabel erfüllten Attribute für die verschiedenen Qualitätskriterien der ISO 9126 und deren Unterkriterien, soll Tabelle 4 Aufschluss darüber geben, in welchem Maße die vier evaluierten Werkzeuge der Forderung nach Softwarequalität gerecht werden. Eine hundertprozentige Erfüllung bedeutet dabei ein akzeptable Erfüllung aller Attribute des entsprechenden Bereichs.

<i>Kriterium</i>	<i>Werkzeug</i>	<i>CALLISTO</i>	<i>GATE</i>	<i>MMA2</i>	<i>UAM</i>
Qualität der Funktionalität		57,1%	90,5%	81%	61,9%
(I) Qualität der Angemessenheit		56,3%	87,5%	87,5%	56,3%
(II) Qualität der Interoperabilität		60%	100%	60%	80%
Qualität der Benutzbarkeit		70,4%	77,8%	66,7%	63%
(I) Qualität der Erlernbarkeit		85,7%	85,7%	71,4%	100%
(II) Qualität der Bedienbarkeit		60%	73,3%	60%	60%
(III) Qualität der Konformität		80%	80%	80%	20%

Tabelle 4: Übersicht zur Qualität von Funktionalität (Angemessenheit, Interoperabilität) und Benutzbarkeit (Erlernbarkeit, Bedienbarkeit, Konformität)

Die Evaluation zeigt, dass das GATE-Framework in vielen Bereichen Stärken zeigt, und dabei nur wenige Schwächen aufweist. Individuelle Schwachstellen sind zwar vorhanden, können aber in allen

Fällen entweder kompensiert oder vernachlässigt werden. Ursprünglich konzipiert als komplexe Architektur für die Implementierung beliebiger Anwendungen aus dem Bereich des Text Engineering, ist die Verwendung als Annotationswerkzeug nur eines von vielen möglichen Einsatzgebieten der *General Architecture for Text Engineering*. GATE stellt für die Automatisierung von Abläufen ein elaboriertes Konzept auf Basis von endlichen Automaten und regulären Ausdrücken zur Verfügung, welches die regelbasierte Programmierung von Annotationsautomatismen erlaubt. Eine weitere Stärke von GATE besteht darin, verschiedenste Dateiformate lesen zu können. Die optional zuschaltbare Funktion *markup-awareness* erlaubt es darüber hinaus, Dokumente mit bereits bestehendem Markup hinsichtlich der verwendeten Tags und der Dokumentstruktur zu interpretieren. So können etwa teilannotierte Texte in einem beliebigen Markupformat importiert und mitsamt den Annotationen bearbeitet werden.

Bei der Evaluation wird außerdem deutlich, dass trotz des enormen Funktionsumfangs von GATE, und der durchweg guten Testergebnisse aller evaluierten Werkzeuge immer noch etliche Bereiche wie etwa die Koordination eines verteilten Workflows und die konsistente Annotation von sprachlichen Ambiguitäten existieren, die noch von keinem der getesteten Annotationswerkzeuge zufriedenstellend abgedeckt werden. Da GATE aber bereits seit 1995 (Cunningham et al. 2007) kontinuierlich weiterentwickelt wird, und die Community des *open source*-Werkzeugs stetig wächst, scheint GATE gut geeignet, die benannten Desiderata in absehbarer Zeit zu erfüllen.

6 Literatur

- Burghardt, Manuel (2008).** Annotationswerkzeuge für diachrone Korpora. Klassifikation und Evaluation von Annotationswerkzeugen. Magisterarbeit, Informationswissenschaft, Universität Regensburg, online verfügbar auf dem Volltextserver der Universität Regensburg unter: urn:nbn:de:bvb:355-opus-10769.
- Cunningham, Hamish et al. (2007).** *Developing Language Processing Components with GATE Version 4 (a User Guide)*. For GATE version 4.0 (July 2007). [<http://gate.ac.uk/sale/tao/index.html>] – Zugriff am 10.06.2008.
- DIN 66272 (1994).** Bewerten von Softwareprodukten. Qualitätsmerkmale und Leitfaden zu ihrer Verwendung (Identisch mit ISO/IEC 9126: 1991). In: *DIN-Taschenbuch 354 – Software-Ergonomie*. Berlin: Beuth Verlag.
- Dipper, Stefanie et al. (2004).** Simple Annotation Tools for Complex Annotation Tasks: An Evaluation. In: *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*. Lisbon, S. 54–62.
- Dipper Stefanie (2005).** XML-based Stand-off Representation and Exploitation of Multi-level Linguistic Annotation. In: *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, S. 39–50.
- Dybkjær, Laila et al. (2001).** *Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data*. ISLE Natural Interactivity and Multimodality Working Group Deliverable D11.1. [<http://isle.nis.sdu.dk/reports/wp11/D11.1-14.2.2001.pdf>] – Zugriff am 10.06.2008.
- Halliday, M. A. K. & J. R. Martin (eds.) (1981).** Readings in Systemic Linguistics. London: Batsford Academic & Educational.
- Heilemann, Michael (2008).** *Informationsstrukturierung für die syntaktische Annotation eines diachronen Korpus des Deutschen*. Magisterarbeit, Informationswissenschaft, Universität Regensburg, online verfügbar auf dem Volltextserver der Universität Regensburg unter: urn:nbn:de:bvb:355-opus-10778.

- Ide, Nancy & Chris Brew (2000).** Requirements, Tools, and Architectures for Annotated Corpora. In: *Proceedings of the EAGLES/ISLE Workshop on Data Architectures and Software Support for Large Corpora*. Paris: European Language Resources Association, S. 1–5.
- Kroymann, Emil et al. (2004).** Eine vergleichende Analyse von historischen und diachronen digitalen Korpora. Technischer Bericht 174 am Institut für Informatik. Berlin: Humboldt-Universität.
- Tidwell, Jennifer (2006).** *Designing Interfaces. Patterns for Effective Interaction Design*. Sebastopol: O'Reilly.
- Tognini-Bonelli, Elena (2001).** Corpus linguistics at work. Amsterdam: John Benjamins.
- Trippel, Thorsten, Thierry Declerck & Ulrich Heid (2005).** Sprachressourcen in der Standardisierung. LDV-Forum 20(2) (2005), 17-30.
- Wolff, Christian (2004).** *Systemarchitekturen. Aufbau texttechnologischer Anwendungen*. In L. Lemnitzer & H. Lobin (Eds.), *Texttechnologie. Perspektiven und Anwendungen* (pp. 165-192). Tübingen: Stauffenburg.

4. Zwischenbilanz und Ausblick

Für die Erfassung von semantischen Feldern zu bestimmten inhaltlichen Konzepten (institutionelle Akteure, Europa), die aus politikwissenschaftlicher Sicht interessierten, eignete sich die gegenwärtige Version von DISCO nur bedingt, weil lediglich ähnlich verwendete und kookkurrierende Wörter (zu "Deutschland" bspw. "Frankreich", zu "Europa" "Asien") erfasst werden, nicht aber semantische Untergruppen wie "Bundesregierung", "Auswärtiges Amt" etc. Eine weitere Einschränkung von DISCO ist, dass es lediglich tokenbasiert arbeitet und daher keine Mehrwortausdrücke gesucht werden können. Auch ist mit DISCO nicht unmittelbar festzustellen, ob durch ein Wort im Korpus stets der gerade gesuchte Akteur benannt wird. Beispielsweise bezieht sich ein Vorkommen von "Bundeskanzler" nicht immer auf Deutschland als außenpolitischen Akteur. Um sicherzustellen, dass sich bestimmte Vorkommen einzelner Wörter regelmäßig auf das gesuchte semantische Feld bzw. den dadurch bezeichneten Akteur beziehen, ist eine Untersuchung der einzelnen Belegstellen notwendig, was derzeit nur durch eine manuelle Konkordanzanalyse geleistet werden kann.

Im Text Mining wie bei der Inhaltsanalyse besteht eine grundsätzliche Kluft zwischen der konzeptuellen Ebene einerseits und der textlich-lexikalischen Ebene andererseits. Existiert eine manuell erstellte Ontologie aus interessierenden Konzepten, d.h. abstrakten semantischen Einheiten, müssen diese in konkret formulierten sprachlichen Ausdrücken wiedergefunden werden. Texte müssen also erst einmal mit Konzepten annotiert werden. Dabei treten die schon aus dem Information Retrieval bekannten Probleme der Lesartenambiguität und Paraphrase auf, d.h. ein Korpus mit einer vorhandenen Ontologie zu annotieren ist keineswegs trivial. Wir erwarten, mit einem korpusgetriebenen Ontologieaufbau mittels DISCO dieses Problem entschärfen zu können.

Ein alternativer Ansatz besetzt darin, textbasiert vorzugehen und aus relevanten Phrasen im Text einen "Themenbaum" aufzubauen. Hier wird versucht, von der konkreten Formulierung im Text zu einer semantischen, möglichst abstrakten Darstellung zu gelangen. Die Schwierigkeit ist dabei, einen hinreichenden Abstraktionsgrad zu erreichen, um gleichbedeutende, aber unterschiedlich ausgedrückte Inhalte überhaupt aufeinander beziehen zu können. Außerdem ist zu erwarten, dass es sich bei den automatisch extrahierten Themen nicht unbedingt um Konzepte der Art handelt, die einer intellektuell erstellten Ontologie entsprechen. Wir hoffen, durch die kombinierte Verwendung der verschiedenen Ansätze die Kluft zwischen konzeptueller und textueller Ebene überbrücken zu können.

Literatur

- P. Cimiano, J. Völker und R. Studer (2006): Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. *Information, Wissenschaft und Praxis*, 57(6-7), S. 315-320.
- M. Hearst (1992): Automatic acquisition of hyponyms from large text corpora. *COLING 1992*, Nantes, Frankreich, S. 539—545.
- P. Kolb (2008): DISCO: A Multilingual Database of Distributionally Similar Words. *Tagungsband der 9. Konferenz zur Verarbeitung natürlicher Sprache – KONVENS 2008*, Berlin.
- R. Rapp (1999): Automatic Identification of Word Translations from Unrelated English and German Corpora. *Proceedings of ACL*, College Park, Maryland, S. 519–526
- J.J. Rocchio (1971): Relevance Feedback in Information Retrieval. In G. Salton (Hrsg.): *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall.
- G. Salton (1971): *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall.
- H. Schmid (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*, S. 47—50.
- F. Sebastiani (2002): Machine learning in automated text categorization. *ACM Computing Surveys*, 34, S. 1—47.