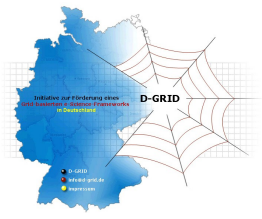
Ontologien und Wortnetze

Version 2008-01-29
Arbeitspaket: AP5
verantwortlicher Partner: Uni Trier

TextGrid

Modulare Plattform für verteilte und kooperative wissenschaftliche Textdatenverarbeitung - ein Community-Grid für die Geisteswissenschaften



Bundesministerium für Bildung und Forschung

Projekt: **TextGrid**
Teil des D-Grid Verbundes und der deutschen e-Science Initiative
BMBF Förderkennzeichen: 07TG01A-H
Laufzeit: Februar 2006 - Januar 2009

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:
Stefan Büdenbender, Uni Trier

Inhaltsverzeichnis

1 Ausgangslage

1.1 TextGrid und Ontologien

Die dynamische Entwicklung der vielfältigen *semantic web*-Initiativen ist eng mit dem Einsatz von Ontologien, ihrer Konzeption, Erstellung und Einbindung verknüpft. Der Terminus ist indes nicht vollkommen eindeutig, die Einsatzfelder sind weit und teils sehr unterschiedlich. Versuche einer allgemeingültigen Definition dessen, was unter „Ontologie“ zu verstehen ist, fallen dementsprechend unscharf aus.

Während der vorliegende Report diese Frage nicht außer Acht lassen kann, muss er die Tiefe des Begriffs unterdessen nicht vollständig ausloten; in TextGrid wurden Ontologien von Anfang an eine relativ klar umrissene Rolle zugemessen, so dass die folgenden Kapitel sich auf diejenigen Teilgebiete konzentrieren sollen, die für das Projekt in seiner jetzigen Form bzw. für schon geplante zukünftige Ausbaustufen relevant sind. Dabei wird die philologisch-konzeptionelle Ebene im Vordergrund stehen; Fragen nach der konkreten programmiertechnischen Implementierung bzw. existierenden Softwareumgebungen werden im folgenden Report 1.6 abgehandelt, eine Zusammenfassung zu einem Gesamtreport soll ebenfalls erfolgen. (Der Titel des vorliegenden Reports wurde entsprechend angepasst.)

Aus dem kombinierten Angebot von ebenso umfassenden wie unterschiedlichen Textdaten und darauf aufbauenden Services ergeben sich für TextGrid zwei potenzielle Einsatzfelder: einerseits kann der Bereich der Analyse, Annotation, und Recherche von Textdaten von der Integration lexikalischer Ontologien profitieren. Zweitens kann die Verwaltung von Textdaten und Services bzw. die Abstimmung beider auf einander durch den Einsatz von Ontologien effizienter gestaltet werden.

Darüber hinaus kann der Datenbestand selbst – insbesondere hochstrukturierte Daten wie etwa die Wörterbücher – zur Grundlage für die Erstellung von Ontologien werden.

Die potenziellen Einsatzfelder sollen in 1.3 näher ausgeführt werden; vor dem oben skizzierten Hintergrund soll zunächst geklärt werden, welche Art von Ontologien für TextGrid in Betracht kommt.

1.2 Ontologien: Definition, klassische Ziele, Fragestellungen und Probleme

Das „Handbook on Ontologies“[1] zählt eine Reihe von Einsatzfeldern auf, von der KI-Forschung über Wissensmodellierung bis hin zum E-Commerce. Entsprechend weit ist der Ontologiebegriff gefasst: die Herausgeber berufen sich, wie viele andere, auf die Definition von Gruber, der Ontologien als „formal explicit specification of a shared conceptualization for a domain of interest“[2] versteht.

Es geht demnach darum, tragende Konzepte eines Interessengebiets in ihrer Abhängigkeit zueinander so zu modellieren, dass das entstehende Kompendium von Zuordnungen und Schlussregeln von Softwareanwendungen ausgewertet und zur Bearbeitung semantischer Problemstellungen

1 Staab, Steffen u. Rudi Studer (Hgg.). „Handbook on Ontologies.“ Berlin, Heidelberg, 2004.

2 Gruber, T. „A Translation Approach to Portable Ontology Specifications.“ In: Knowledge Acquisition 5 (1993), S. 199-220.

herangezogen werden kann. Dass Ontologien domänenspezifisch sein sollten und auf Konvention beruhen, sind dabei Einschränkungen, die in dieser Form nicht immer gesehen worden sind, was gerade in der Frühzeit der Ontologieforschung zu überzogenen Ansprüchen an die Mächtigkeit von universalen Datenmodellen geführt hatte. Interessanterweise waren solche Erwartungen bereits zu einem gewissen Grad vorgezeichnet, wie ein Blick in die Begriffsgeschichte belegt: der Terminus „Ontologie“ ist der Philosophie entlehnt, wo er zunächst kein fertiges Produkt, sondern eine Disziplin meint. Die klassische Ontologie ist die Lehre vom Seienden, und damit auch immer eine Lehre von der richtigen oder zumindest optimalen Einteilung des Seienden in ein konzeptuelles Gerüst. Die Debatte, ob ein einziges, universales Gerüst existiere, und wenn ja, worauf sich seine Verbindlichkeit stützen könne, wurde dabei über Jahrhunderte erbittert geführt und kulminierte schließlich im mittelalterlichen Universalienstreit. Auch wenn seitdem die metaphysischen Aspekte zunehmend zugunsten von nominalistisch-konstruktivistischen Positionen aufgegeben wurden, wirkte der Wunsch nach einem einheitlichen konzeptuellen System, das es zumindest erlaubt, unser gesamtes *Wissen* über die Welt mit Hilfe einer begrenzten Anzahl von Kategorien und Relationen allgemeinverbindlich zu ordnen, bis tief in die Neuzeit hinein.[3]

Der frühe Optimismus der KI-Forschung wirkt wie ein moderner Reflex auf diese Debatte; im Umfeld des *semantic web* hat sich aber mittlerweile weitestgehend die eingangs zitierte pragmatischere Einstellung durchgesetzt.

Größere Verbreitung fand der Terminus in den frühen 90er Jahren zusammen mit einer Vision des *semantic web*, die Tim Berners-Lee skizzierte.[4] Ontologien und verwandte Techniken werden in diesem Zusammenhang als Modelle gedacht, mit deren Hilfe sich einerseits statische Datenbestände in maschinenlesbarer Form kennzeichnen lassen, andererseits aber auch Funktionalitäten und Schnittstellen von internetbasierten Informationsquellen. Auf diese Art soll es möglich sein, Softwareagenten nach semantischen Zielvorgaben selbständig im Internet Aufgaben, etwa Rechercheaufträge, ausführen zu lassen. Dazu sollen sie nicht nur in der Lage sein, mit Informationsangeboten zu kommunizieren, sondern auch untereinander, um so eine Art selbständige digitale Verarbeitungskette zu bilden.

Im Zusammenhang mit TextGrid wird im Folgenden zwischen Ansätzen zu unterscheiden sein, die die Analyse, Annotation und Recherche von Textdaten betreffen und solche, die stärker auf organisatorische und infrastrukturelle Aspekte abzielen. Das jeweilige Einsatzgebiet schlägt sich dabei auf die Anlage der Ontologien selbst nieder: Die Lexikalisierung der abgebildeten Konzepte beispielsweise ist in vielen Zusammenhängen sekundär oder von eingeschränktem Umfang. In zahlreichen Anwendungen etwa der Wissensrepräsentation geht es gerade darum, einer begrenzten Anzahl von Konzepten genau je eine „richtige“ Benennung zuzuordnen, um Ambiguitäten zu vermeiden (= Terminologie, kontrolliertes Vokabular; je nach Art der Hierarchisierung auch Taxonomie).

---

3 So legt noch Diderots *Encyclopédie*, die maßgeblichen Einfluss auf die europäische Aufklärung hatte, eine entsprechende durchgängige Baumstruktur zu Grunde. Aber auch manche modernen Thesauri scheinen von ähnlichen Prämissen geleitet, ohne diese explizit zu problematisieren (s. u.).

4 Tim Berners-Lee, James Hendler und Ora Lassila: *„The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities“.* (Vgl. Scientific American, 17.5.2001.)

Während es in TextGrid auch für solche Anwendungen prinzipiell Verwendung gibt, liegt der Fokus innerhalb des Arbeitspakets 5 hingegen explizit auf Ontologien mit lexikalischem Schwerpunkt, die einen semasiologischen Ansatz auf natürlichsprachlichen Texten erlauben. Solche Ontologien bzw. Wortnetze erfassen dabei einen substanziellen Teil des lexikalischen Bestandes einer Sprache und verzeichnen in maschinenlesbarer Form eine Vielzahl von semantischen und lexikalischen Relationen zwischen Wörtern bzw. Wortgruppen (im Unterschied etwa zu Thesauri).

Die Einsatzmöglichkeiten solcher Ansätze in TextGrid sollen im folgenden Überblick zunächst abstrakt erläutert, dann mit Blick auf konkrete Produkte und Projekte näher ausgeführt werden.

1.3 Potenzielle Einsatzgebiete in TextGrid

Für TextGrid als internetbasierte Forschungs- und Arbeitsumgebung ergibt sich prinzipiell ein relativ breites Spektrum für den Einsatz von Ontologien, das jedoch sowohl im Projektantrag wie auch in den daraus hervorgegangenen Arbeitsplänen vorab auf einige Kernbereiche, insbesondere die Annotation von Textdaten, eingeschränkt wurde. Der Report soll darüber hinaus einige Empfehlungen für derzeitig projektierte weitere Ausbaustufen aussprechen, wird sich aber insgesamt am gegenwärtigen bzw. bis Ende 2008 voraussichtlich erreichten Stand orientieren.

TextGrid tritt, wie oben angemerkt, sowohl als Anbieter von Services (= Informationsdiensten und Modulen zur Textdatenverarbeitung) als auch von Content/Daten (= Textdaten) auf. Selbst wenn die Grenze in einigen Fällen fließend ist (etwa das Angebot von Wörterbuchdaten als Service), ergeben sich in Hinblick auf den Einsatz von Ontologien zwei klar getrennte Anwendungsgebiete.

1.3.1 TextGrid als Anbieter von Services

Die von Berners-Lee skizzierte selbständige Interaktion von Services und Schnittstellen über intelligente Softwareagenten ist langfristig auch für TextGrid ein wegweisendes Szenario. Entsprechende Projekte (s. u.) sollen deshalb in Hinsicht auf geplante weitere Ausbaustufen weiter beobachtet werden; im gegenwärtigen Projektrahmen spielen sie aber schon aus pragmatischen Gründen keine Rolle. TextGrid setzt bei der Verwaltung und Verknüpfung der Dienste zunächst auf ein internes Workflowmanagement und eine eigene Service-Registry. Der weitere Ausbau in Richtung *semantic grid* soll im Rahmen der geplanten D-Grid Wissensschicht stattfinden (vgl. insbesondere die Präsentationen von Studer u. Hasselbring auf http://www.d-grid.de/index.php?id=workshops).

Das Arbeitspaket 5 kann hier jedoch auf einer wesentlich elementareren Ebene tätig werden, indem es die zur Koppelung von Textdaten und Services nötigen Metadaten entsprechend optimiert (s. 2.2.2.2).

1.3.2 TextGrid als Anbieter von Content

Die von TextGrid angebotenen Tools und Recherchemöglichkeiten erhalten ihren besonderen Wert aus der engen Verzahnung mit den im projekteigenen StorageGrid gespeicherten Textdaten. Ontologien können bei der Analyse, Annotation und Recherche dieser Datenbestände helfen.

Neben diesen Anwendungsgebieten, die auf die Integration bestehender Ontologien und Wortnetze abzielen, ergibt sich noch eine weitere Möglichkeit, die der Erstellung von Ontologien. Im Vordergrund stehen dabei mittelfristig allerdings noch keine Verfahren zur automatischen Ontologieextraktion, da es noch offen ist, ob TextGrid in der ersten Projektphase die nötige, extrem breite Grundlage an Textdaten erreichen kann. Statt dessen soll das Verweisnetz des

Wörterbuchverbundes ausgebaut werden. Die entsprechenden Arbeitsschritte werden in 2.2.1 dargelegt.

2 Ontologien in der Praxis

2.1 Existierende Ontologien/Wortnetze und ihre Repräsentation

2.1.1 Ontologien mit lexikalischem Schwerpunkt (deutscher Sprachraum)

Als Prototyp für Ontologien mit breitem, auf die Bearbeitung natürlichsprachlicher Daten ausgerichtetem lexikalischen Fundament darf das englischsprachige WordNet gelten, das 1985 an der Princeton University ins Leben gerufen wurde. Koordiniert von der Global WordNet Association (GWA; vgl. http://www.globalwordnet.org) und dem EuroWordNet Consortium (http://www.illc.uva.nl/EuroWordNet) wurden in der Folge nach seinem Vorbild Wortnetze für zahlreiche andere Sprachen erstellt, die weitestgehend interoperabel sind. Dies betrifft nicht nur die technische Repräsentation, sondern reicht bis zur Strukturierung der Inhalte: so gibt EuroWordNet eine gemeinsame *top-ontology*, d. h. eine Reihe von Basiskonzepten zur Ordnung der oberen Hierarchieschichten vor.

Der deutschsprachige Raum wird dabei von GermaNet abgedeckt, das hier kurz vorgestellt werden soll.

2.1.1.1 GermaNet

GermaNet wurde im Jahre 1996 an der Universität Tübingen initiiert und liegt mittlerweile in Version 5.0 vor (vgl. http://www.sfs.uni-tuebingen.de/lsd). Weitgehende Kompatibilität zu WordNet bzw. den von der GWA und EuroWordNet verwalteten Wortnetzen war eine Leitlinie bei der Konzeption.

GermaNet umfasst ca. 76500 lexikalische Einheiten in über 53000 sog. Synsets, d. h. Gruppen von (Teil)Synonymen, deren Verhältnis zueinander durch ein Netzwerk explizit gekennzeichneter semantischer Relationen markiert ist. Darüber hinaus sind lexikalische Relationen zwischen einzelnen Wörtern verzeichnet. Abgedeckt werden neben Substantiva auch Verben und Adjektive. Im Gegensatz zu rein taxonomisch strukturierten Systemen steht dabei nicht die eindeutige Hierarchisierung im Vordergrund; Kreuzklassifikationen sind möglich, die Vernetzung fällt durch ein breiteres Spektrum von Relationstypen wesentlich dichter aus. So wird die Grundrelation der Synonymie auf Wortebene durch Antonyme und Ableitungen ergänzt, die Synsets sind nicht nur nach Ober- und Unterbegriffe (Hyper- bzw. Hyponomie) geordnet, sondern es werden auch Meronymie (Teil-Ganzes-Relation), Implikation und Kausation verzeichnet.

GermaNet ist als eine Sammlung von XML Dateien erhältlich, die einer eigenen *document type definition* (DTD) folgen. Eine Überführung in Standards zur Ontologierepräsentation wie OWL (s. u.) wurde bisher nur ansatzweise realisiert, allerdings gibt es bereits eine Reihe von Tools, die den Umgang mit den Daten erleichtern.

Aus inhaltlicher Sicht wäre GermaNet eine äußerst wichtige Ressource für TextGrid; einzig der immer noch vergleichsweise kleine lexikalische Bestand würde die Nutzbarkeit zu einem gewissen Grad einschränken. Wesentlich stärker wird sie de facto durch die restriktiven Lizenzbedingungen eingeschränkt. Im Gegensatz etwa zu WordNet ist GermaNet nicht frei erhältlich und die ursprünglich

geplante permanente Einbindung in TextGrid über einen dedizierten *ontology manager* wäre mit erheblichem finanziellen Aufwand verbunden, der im gegenwärtigen Projektrahmen nicht vertretbar ist. In 2.2.1 soll – auch mit Blick auf geplante weitere Ausbaustufen – dennoch nicht darauf verzichtet werden, einige Einsatzszenarien zu skizzieren, wobei der Fokus auf Anwendungen gelegt wird, die im Rahmen einer nicht-kommerziellen Lizenz zu bewerkstelligen sind. Parallel dazu wird zu klären sein, an welchen Stellen ggf. Alternativen zum Einsatz kommen können. Als solche bietet sich in erster Linie der OpenThesaurus an.

2.1.1.2 OpenThesaurus

OpenThesaurus (http://www.openthesaurus.de) ist gleichermaßen der Name für ein Projekt wie auch für sein nicht-kommerzielles Produkt, bei dem es sich, wie der Name ebenfalls bereits andeutet, zumindest im Sinne der *semantic web* Forschung um keine vollwertige Ontologie handelt. OpenThesaurus beschränkt sich zwar einerseits nicht auf eine Auflistung von Wortfeldern, sondern strebt eine vollständige taxonomische Struktur an, vermerkt dabei aber außerhalb der starren hierarchischen Ordnung (Hyper- bzw. Hyponomie) keine weiteren semantischen Relationen. Z. Zt. (Jan. 2008) beinhaltet er ca. 44000 Synonyme, wird aber ständig und unter reger Einbeziehung der Nutzergemeinschaft weiterentwickelt und hat bereits Verwendung in verschiedenen open-source Programmen wie etwa *Open Office* gefunden.

Der Transparenz und Offenheit stehen neben der bisher kleinen lexikalischen Basis und eingeschränkten Vernetzung allerdings für manche Einsatzzwecke grundsätzliche Probleme entgegen. So wurden zwar die sieben Synsets der obersten Hierarchieebene von WordNet übernommen, die weitere Taxonomie scheint jedoch teilweise unausgewogen und nicht immer stringent. Ein weiterer kritischer Bereich ist die Qualitätsprüfung; im Gegensatz zu GermaNet sind die Strukturen offensichtlich nicht lückenlos philologisch geprüft. (Vgl. auch http://www.danielnaber.de/publications/gldv-openthesaurus.pdf.)

Dennoch kann der OpenThesaurus für einige Anwendungsbereiche durchaus eine Alternative zu GermaNet bieten (s. u.).

2.1.2 Serviceverwaltung

Dieser Punkt, wie auch der folgende, soll an dieser Stelle nur der Vollständigkeit halber genannt werden. Wie oben dargelegt, soll TextGrid auf dem Feld einer ontologiebasierten Steuerung und Verkettung von Webservices erst im Rahmen der D-Grid Wissensschicht aktiv werden.

Als wegweisend kann hier das EU-Projekt OntoGrid (http://www.ontogrid.net) gelten, das an einer entsprechenden Referenzarchitektur (S-OGSA) arbeitet; technische Einzelheiten bleiben Report 1.6 vorbehalten.

2.1.3 Repräsentation von Ontologien

Auch in Bezug auf Fragen der technischen Umsetzung der eigentlichen Ontologien soll Report 1.6 zu bestehenden Softwarelösungen nicht vorgegriffen werden, hier gilt es mit Blick auf das folgende Kapitel nur eine kurze Empfehlung auszusprechen.

Während in weiteren *semantic web*-Kontext eine große Vielfalt von Standards und Beschreibungssprachen existiert, hat das W3C-Konsortium in Bezug auf Ontologien nicht nur klare

Empfehlungen für RDF(S) und OWL ausgesprochen, sondern ist auch aktiv an deren Weiterentwicklung beteiligt (vgl. http://www.w3.org/2001/sw/BestPractices/WNET/tf.html).

RDF (ressource description framework) bzw. RDF-Schema stellen dabei eine grundsätzliche Möglichkeit zur Verfügung, mittels einer Subjekt-Prädikat-Objekt-Struktur Aussagen über Ressourcen (über URIs identifizierbare abstrakte oder physische Entitäten, auf die referenziert werden soll) zu modellieren. OWL ist eine darauf basierende, relativ mächtige Beschreibungssprache, die eigens zum Erstellen und Verwalten von Ontologien geschaffen wurde und im Umgang mit Wortnetzen das Mittel der Wahl darstellt.

Neben einer weit verbreiteten XML Syntax existieren gibt es auch RDF Formate, die für Datenbankumgebungen optimiert sind.

2.2 Ontologien in TextGrid; Anwendungsszenarien

Der TextGrid-Antrag sah für AP5 eine starke Fokussierung auf GermaNet vor, unterstützt durch einen zentralen Ontologiemanager, der es ermöglichen sollte, aus verschiedenen Anwendungen heraus auf GermaNet zu verweisen bzw. Links einzurichten. Von dieser aufwendigen Zentrallösung musste angesichts der restriktiven Lizenzbedingungen jedoch Abstand genommen werden. Der Haupteinsatzzweck von GermaNet, die Erstellung einer semantisch annotierten Metalemmaliste, wurde deshalb in einen eigenen Projektantrag (s. u.) verlagert. Dennoch sollen innerhalb von AP5 spezielle Tools, etwa der Wörterbuch-Linkeditor, die Möglichkeiten ausloten, GermaNet auf Basis einer akademischen Lizenz für einen begrenzten Zeitraum einzusetzen. Zweitens sollen philologisch relevante Ansätze auch im Hinblick auf geplante weitere Ausbaustufen des Projekts sondiert werden, gleichzeitig wird zu prüfen sein, wo und in wie weit OpenThesaurus eine Alternative sein kann.

2.2.1 Nutzung existierender Ontologien im Umgang mit Textdaten: Analyse, Annotation, Recherche, Verknüpfung

Der Einsatz von Ontologien in AP5 zielt vor allem auf erweiterte Recherchemöglichkeiten in Bezug auf Primärtexte ab, was untrennbar einhergeht mit einer entsprechenden Analyse und Anreicherung der Quellen mit Zusatzinformationen, die Inhalte beschreiben. Neben diesen semantischen (bzw. onomasiologischen) Metadaten wird es eingeschränkt auch um solche gehen, die statische Eigenschaften bzw. Symptomwerte der Texte beschreiben (deskriptive Metadaten).

Dabei nehmen die in TextGrid integrierten digitalen Wörterbücher aufgrund ihrer hochstrukturierten Form insgesamt eine Sonderstellung ein, die bei den folgenden Szenarien zu berücksichtigen und ggf. gesondert zu behandeln sein wird.

2.2.1.1 Disambiguierung, Annotation

-Allgemein:

Ein Problem bei der Recherche stellen Homographen und polyseme Ausdrücke dar. Ontologien werden hier zur Lesartendisambiguierung und gezielten Suche eingesetzt, indem der Kontext auf Schlüsselwörter untersucht wird, die eine Einordnung des Begriffs ermöglichen: das Umfeld etwa von „Bank“ würde in einem gegebenen Primärtext auf Ausdrücke untersucht, die entweder dem Wortfeld „Finanzinstitut“ oder „Sitzmöbel“ angehören. Die entsprechende Zusatzinformation wird im Text bzw. in den Metadaten abgespeichert und kann bei der Recherche berücksichtigt werden, etwa indem bei der Eingabe eines mehrdeutigen Suchwortes eine Rückfrage an den Benutzer erfolgt.

-Wörterbücher:

Eine Sonderform dieses Verfahrens stellt die Koppelung von Stichwortansätzen in Wörterbüchern mit einer Ontologie dar, was gleich in mehrfacher Hinsicht eine gezieltere Recherche ermöglicht. Neben der Disambiguierung von Homographen kann durch die Hinterlegung der entsprechenden taxonomischen Strukturen ein onomasiologischer Zugang zu dem enthaltenen Weltwissen geschaffen werden: werden zu einem Stichwort jeweils passende Ober- und Unterbegriffe hinterlegt, kann einerseits nach Kategorien gesucht werden („finde alle Artikel zu Sitzmöbeln“), andererseits können zahlreiche neue Verweise („Bank“ als Sitzmöbel ist verwandt mit „Stuhl“) eingerichtet werden.

Die Vorteile einer solchen Anreicherung zeigen Projekte wie *Krünitz Online* (www.kruenitz1.uni-trier.de), eine Enzyklopädie, im Zuge der Retrodigitalisierung manuell per DDC (Dewey-Dezimalklassifikation) annotiert wurde. Im Gegensatz zur – ebenfalls kostenpflichtigen – DDC hätten lexikalische Ontologie wie GermaNet oder OpenThesaurus den grundsätzlichen Vorteil, dass sich zumindest ein lexikalischer Kernbestand (teil)automatisch mappen ließe. Aus philologischer Sicht wäre GermaNet als Quasi-Standard hier vorzuziehen; im Rahmen des AP5 sollen aber auch Tests mit OpenThesaurus durchgeführt werden.

TextGrid stellt hier mit dem Lemmasuch-Service und dem geplanten Wörterbuch-Linkeditor eine leistungsfähige Infrastruktur zur Verfügung. Die Erstellung einer philologisch geprüften, onomasiologisch annotierten Metalemmaliste über den gesamten Bestand des Wörterbuchnetzes soll flankierend im Rahmen des Antrags „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“ betrieben werden. Der Antrag wurde am 30.07.07 beim BMBF eingereicht und ist in die engere Auswahl gelangt; am 28.01.08 fand eine Evaluation vor einem internationalen Gutachtergremium statt, mit einer Entscheidung wird Mitte Februar gerechnet.

2.2.1.2 automatische Verschlagwortung:

Eine Reihe von Verfahren zur inhaltlichen Analyse von Texten beruht auf statistischen Ansätzen, die Dokumente nach Wortformen durchsuchen, die signifikant häufiger vorkommen als, gemessen an einem Vergleichskorpus, zu erwarten. Solche, aus dem Bereich des Textretrievals stammenden, Verfahren (Termfrequenz – Inverse Dokumentfrequenz) kommen im Rahmen von TextGrid innerhalb des Wörterbuchnetzes bereits zum Einsatz und können durch den Einsatz von Ontologien stark verbessert werden, indem der Text lemmatisiert, und inhaltlich verwandte Begriffe gruppiert werden. Um die Zuweisung von Schlagworten möglichst zu standardisieren, wäre auch hier GermaNet am besten geeignet.

2.2.1.3 Verknüpfung verwandter Dokumente

Das oben skizzierte Verfahren muss nicht mit der Zuweisung von Schlagworten abgeschlossen werden; es kann ebenfalls zu Markierung einer nicht näher spezifizierten inhaltlichen Verwandtschaft oder Ähnlichkeit verwendet werden, was das Primärziel der Wörterbuchverknüpfung ist, bei der Hyperlinks zwischen Artikel nach inhaltlichen Gesichtspunkten und unabhängig vom Artikelstichwort eingerichtet werden.

Allgemein sind so aber Verknüpfungen zwischen beliebigen Dokumenten (z. B. auch Wörterbucheinträge – Primärtext) bzw. textsortenübergreifende Vergleiche möglich. Da hier die

Benennung von Inhalten sekundär ist, wäre zum Gruppieren verwandter Wortformen auch OpenThesaurus gut geeignet.

2.2.1.4 Recherche

Die Rolle des Recherchetools wurde implizit bereits angesprochen; wo Textdaten mit semantischen Informationen angereichert wurden, muss die Suchinstanz auf das gleiche Metadatenset bzw. die gleiche Ontologie zurückgreifen können, um Anfragen gezielter gestalten zu können. (Im Bereich der deskriptiven Metadaten ist dies bereits der Fall, auch wenn unter 2.2.2.2 noch Verbesserungspotenzial aufgezeigt werden soll.) Hier wäre im semantischen Bereich eine Zentrallösung nach wie vor wünschenswert und soll unabhängig von lizenzrechtlichen Fragen in Report 1.6 geprüft werden.

Ungeachtet dessen ist eine Query Expansion vor der Suche, d. h. eine Suche, die auf Wunsch verwandte Suchbegriffe mit einbezieht, durch die Integration von OpenThesaurus möglich und soll zumindest im Bereich der Wörterbuchrecherche als Option zur Verfügung stehen.

2.2.2 Eigene Ansätze

2.2.2.1 Vorhandene Daten/Ressourcen (Ontologieerstellung/-extraktion)

Verfahren zur automatischen Ontologieextraktion erfordern einen Korpusumfang, von dem TextGrid z. Zt. noch weit entfernt ist.

Einen Sonderfall bilden auch hier die Wörterbücher, die aufgrund ihrer Systematik potenziell geeigneter sein könnten. Vor allem aber bildet die Summe der Querverweise zwischen inhaltlich verwandten Einträgen ein Wortnetz von Artikelstichwörtern, das für die Forschung schon deshalb von großem Interesse ist, da es auch Sprachstufen und Varietäten mit einbezieht, die bisher nicht Gegenstand ontologischer Erschließung waren. Hier soll die Möglichkeit einer manuellen Qualitätsprüfung (im Falle automatisch eingerichteter Verweise) und Annotation von semantischen/lexikalischen Relationen geschaffen werden. Zumindest ein kleiner Teil der Verweise dürfte sich auch durch einen teilautomatischen Abgleich mit GermaNet bzw. OpenThesaurus einrichten lassen.

2.2.2.2 Deskriptive Metadaten: Verknüpfung von Ressourcen

Für die Auffindbarkeit von Textdaten spielen nicht nur semantische, d. h. inhaltsbeschreibende Metadaten eine wichtige Rolle, sondern auch deskriptive Metadaten, die statische Eigenschaften bzw. Symptomwerte von Texten beschreiben. Standardisierte Metadatenschemata wie z. B. Dublin Core stellen zwar keine Ontologien im engeren Sinne dar, können aber durch die Hinzufügung von Interferenzregeln entsprechend ausgebaut werden. Insofern gibt es Berührungspunkte zu den im Rahmen von AP5 zu erstellenden *metadata application profiles*, auf die hier abschließend noch kurz eingegangen werden soll.

So kann es für die Zuweisung von Textdaten auf der einen Seite und Tools bzw. Services auf der anderen beispielsweise nötig sein, Eigenschaften wie Sprache, Sprachstufe oder etwa Textgattung zu kennen. In der Praxis können dabei verschiedene Probleme auftreten, beispielsweise, wenn Daten importiert werden sollen, die nach einem anderen Standard markiert sind. TextGrid sieht hier ein Mapping auf ein internes Metadatenset vor, das in seinen Kategorien interoperabel zu internationalen Standards (DC, zvdd, TEI) sein soll. Ein zweiter kritischer Punkt sind die Werte, die diese Kategorien

annehmen können; eine direkte Zuordnung wird hier oft dadurch erschwert, dass existierende Standards von unterschiedlicher Granularität oder unvollständig sind.

Hier können Ontologien im Sinne von hierarchisch geordneten Sets von Metadaten Fallbacklösungen zur Verfügung anbieten, um gleichwertige oder zumindest geeignete Entsprechungen zu finden.

So kann einem Text, der innerhalb von TextGrid nach ISO 639-3 als „schwäbisch“ gekennzeichnet ist, zwar kein unmittelbar passendes Wörterbuch zur Verfügung gestellt werden; durch ein Nachschlagen in der Hierarchie der Sprachcodes kann aber festgestellt werden, dass es sich um einen neuhochdeutschen Dialekt handelt und andere Wörterbücher aus diesem Bereich angeboten werden. Andererseits kann etwa für das Elsässische Wörterbuch, für den ISO 639-3 keinen passenden Sprachcode zur Verfügung stellt, als Interimslösung ein eigener Code definiert werden, der dann am entsprechenden „Ast“ der Isonorm eingehängt und somit als alemannischer Dialekt gekennzeichnet wird.