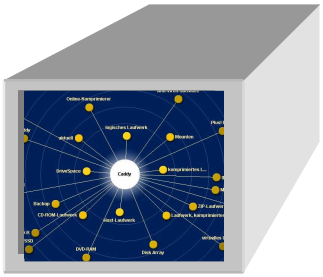
Lexikonbasierte Wissenserschließung: natürlichsprachliche Suche und 3D-Wissensnavigation

Abschlussbericht zum Förderprojekt 01C5885 „LeWi“

„Man sollte die Frage ‚ist der Himmel blau?’ eintippen können und darauf eine Antwort erhalten […] Das ist eine der großen Aufgaben der Informatik.“

Ein anderes Marathon-Projekt: Die Suche nach besseren Benutzerschnittstellen. Es scheint intuitiv richtig zu sein, dass dreidimensionale Oberflächen Bildschirmmaterial klarer präsentieren könnten, so Gates. Microsoft habe an dem Problem seit zehn Jahren gearbeitet. „Und wie viele Prototypen wir da gebaut haben. Alle wurden sie durch Usability-Tests geschickt. Und keiner funktionierte.“

Interview mit Bill Gates, aus: Gates: IT wird völlig unterschätzt, Technology Review 5.3.2004,
http://www.heise.de/tr/aktuell/meldung/print/45276



Inhaltsverzeichnis

1 Vorwort ... 3
2 Die Aufgabenstellung ... 4
3 Ontologie der deutschen Sprache ... 7
3.1 Software und Datenmodell ... 7
3.2 Datenquellen für die Ontologie ... 8
4 Textdatenbank, Thesaurus, Klassifikation und Indexierungssoftware ... 9
4.1 Textdatenbank des Projektes: Brockhaus multimedial 2004 ... 9
4.2 Thesaurus, Klassifikation und Indexierungssoftware ... 10
4.2.1 Thesaurus des Brockhaus multimedial ... 10
4.2.2 Brockhaus-Klassifikation ... 11
4.2.3 Indexierungssoftware ... 12
5 Dialogmodul für natürlichsprachliche Fragen zur Suche in Volltextdatenbanken ... 13
5.1 Fragenanalyse ... 14
5.1.1 Funktion und Struktur der Fragentypologie ... 14
5.1.2 Morphosyntaktische Fragenanalyse ... 15
5.2 Suchen in der Datenbank ... 16
5.3 Ergebnisaufbereitung ... 17
5.3.1 Relevanzberechnung ... 18
5.3.2 Rückfragen ... 19
5.3.3 Fortführung des Dialogs ... 21
6 Visualisierung von Dokumenten aus Volltextdatenbanken ... 24
6.1 Inhaltliche Konzeption ... 24
6.1.1 Herstellung assoziativer Zusammenhänge ... 24
6.1.2 Sachsystematische Einordnung von Dokument-Lemmata ... 25
6.1.3 Interaktionskonzept ... 26
6.2 Technische Umsetzung ... 27
6.2.1 Visualisierungsengine ... 27
6.2.2 XML-Schnittstelle ... 27
6.2.3 Umsetzung der Visualisierung ... 28
6.2.4 Hard- und Software Voraussetzungen ... 28
7 Usabilitytest und Fokusgruppenbefragung an Schulen und Universitäten ... 29
7.1 Die Konzeption der Fokusgruppenbefragungen ... 29
1. Kognitiver Bereich: Verstehen von funktionalen Elementen ... 29
2. Emotionaler Bereich: persönlicher Bezug zur grafischen Oberfläche ... 29
3. Handlungsbereich: Funktionalität und virtuelles Raumverständnis ... 29
4. Erwartungsbereich: Qualität und Quantität des zu Erfahrenden ... 29
7.2 Das Ergebnis der Fokusgruppenbefragungen ... 29
7.3 Eingesetzte Technik zur Durchführung der Fokusgruppenbefragung ... 30
8 Ausblick / wirtschaftliche Verwendung ... 30
8.1 Bibliographisches Institut & F.A. Brockhaus AG / Brockhaus Duden Neue Medien GmbH ... 30
8.2 Krieger, Zander & Partner GmbH ... 31
8.3 interActive Systems GmbH (iAS) ... 31
8.4 Joanneum Research (JR) ... 32
8.5 Intelligent Views (i-views) ... 32
8.6 Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. (IAI) ... 33
8.7 MediaSupervision Software Consulting GmbH (MSSC) ... 33

1 Vorwort

Am 26.2.04 konnte ich zum ersten Mal einen Prototyp der LeWi-Software ausprobieren. Als ich die Frage „Welches Tier hat zwei Höcker?“ eintippte, und der Prototyp als Antwort den Artikel „Kamele“ auf den Bildschirm zauberte, da war mir klar, dass hier gerade ein Meilenstein der technologischen Entwicklung gehauen wurde.

Die Ergebnisse des Projektes stoßen eine Tür auf zu einer neuen Generation von Suchmaschinen, die den Benutzern nicht mehr spezielles Datenbank-Know-how abverlangt und versierten Umgang mit booleschen Operatoren, sondern natürlichsprachige Anfragen so akzeptiert, wie sie von einem Menschen gestellt würden. Mehr noch, das Beispiel zeigt, dass auch Fragen, bei denen die eigentlichen Suchbegriffe unbekannt sind, „vernünftige“ Resultate erzeugen.

Forschungsprojekte, auch solche der produkt- und anwendungsnahen Forschung, unterscheiden sich von Produktentwicklung dadurch, dass man zu Beginn nicht sicher weiß, ob bzw. inwieweit man die anvisierten Ziele erreichen kann. Umso erfreulicher ist es, wenn sich Hoffnungen nicht nur erfüllen, sondern sogar weit übertroffen werden. Das ist hier geschehen.

Im Jahre 2005 wird Brockhaus sein 200jähriges Firmenjubiläum feiern. Dabei wird man sicher auf die glorreiche Tradition des Verlags zurückblicken. Es erfüllt mich aber mit besonderem Stolz, dass es nicht nur die Lorbeeren der Vergangenheit, sondern auch führende Informationstechnologie des 21. Jahrhunderts ist, über die man im Zusammenhang mit Brockhaus dann sprechen wird.

Der amerikanische Softwareriese Microsoft forscht nach Aussage seines Vorsitzenden, Bill Gates, noch nach Suchmaschinen, die auf die Frage "Ist der Himmel blau?" eine sinnvolle Antwort liefern. Ich gratuliere allen Beteiligten und danke in besonderer Weise dem Fördergeber BMBF und Projektträger FhG, die es ermöglicht haben, dass einem deutschen mittelständischen Unternehmen eine Entwicklung gelungen ist, die nicht nur auf diese Frage eine Antwort hat.

Alexander Bob, Vorstand Bibliographisches Institut & F.A. Brockhaus AG

Mannheim, im März 2004

2 Die Aufgabenstellung

In der heutigen Wissens- und Informationsgesellschaft wird es immer aufwendiger, aus der Fülle der angebotenen Informationen die relevanten zu finden, zu selektieren und anschließend die Inhalte zu ordnen und Zusammenhänge zu erkennen. Der Zugang zu den Informationen erfolgt schon heute häufig über Volltextdatenbanken und in Zukunft werden mehr und mehr Menschen solche Datenbanken nutzen. Der Umgang mit solchen Volltextdatenbanken ist aufgrund komplizierter Interfaces, die oft nicht intuitiv und benutzergerecht gestaltet sind, schwierig und zeitaufwendig, oft sogar unmöglich. Insbesondere die Abfrage einer Datenbank über die Eingabe einzelner Stichwörter in Verbindung mit booleschen Operatoren bereitet allen Nutzern Probleme, die relevanten Daten überhaupt oder in angemessener Zeit herauszufiltern. Die Qualität der Ergebnisse einer Abfrage krankt zudem grundsätzlich an der eingeschränkten semantischen Intelligenz der klassischen Volltextsuche auf Wortbasis.

Ein weiteres Manko von Wissensdatenbanken ist, dass für den Benutzer häufig die Zusammenhänge zwischen den verschiedenen angebotenen Dokumenten aufgrund der Fülle der Informationen verloren gehen oder, vor allem zwischen entfernteren Sachgebieten, gar nicht erst erkannt werden.

Um der beschriebenen Problematik entgegenzutreten wurden zwei Lösungsansätze verfolgt:

- Der Zugang zu Volltextdatenbanken soll durch die Möglichkeit von natürlichsprachlich formulierten Anfragen erleichtert werden. Durch die Ausnutzung morphosyntaktischer und semantischer Auswertungen der Frage soll die Qualität der Ergebnisse gegenüber einer traditionellen Suche verbessert werden. Als erfolgreiche "Antwort" auf die Frage sollte die Generierung einer Trefferliste gelten, die möglichst viele in Bezug auf die Frage relevante Artikel enthält, idealerweise den Artikel mit der passendsten Antwort zuoberst und automatisch aufgeschlagen. Weitergehende Ergebnisse, wie z.B. das Markieren der betreffenden Textstelle oder die textliche Generierung einer expliziten Antwort war in diesem Projekt (noch) nicht Gegenstand.
- Für einen besseren Überblick über die Zusammenhänge, in denen ein bestimmter Text steht, soll in einer dreidimensionalen Darstellung die inhaltliche Nähe von Artikeln dargestellt werden und zusätzlich das jeweilige Stichwort (Lemma) eines Artikels sachsystematisch eingeordnet werden.

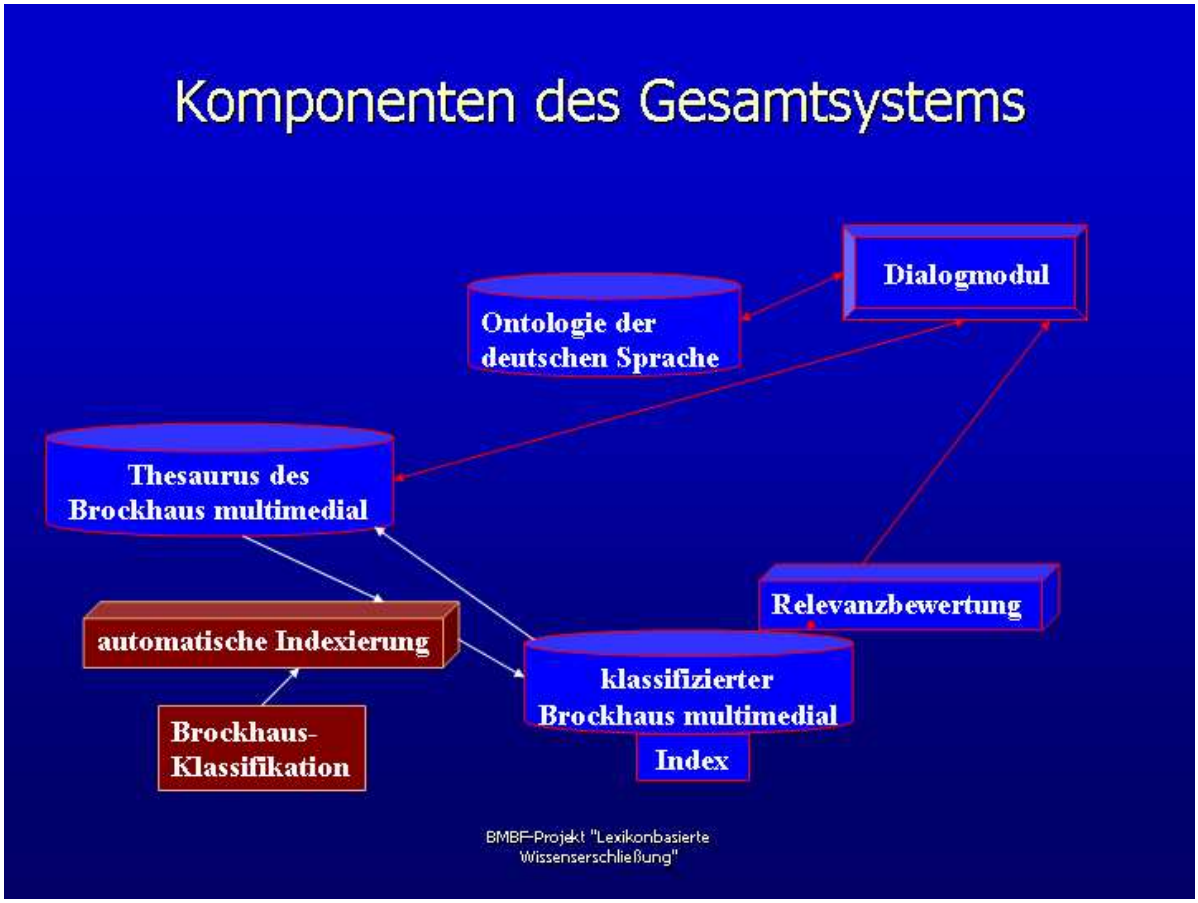
Zur Realisierung dieser Lösungsansätze wurden im Förderprojekt folgende Module entwickelt:

- Eine Ontologie der deutschen Sprache als Unterstützung für die computerlinguistische Analyse natürlichsprachlicher Fragestellungen und für die Darstellung sachsystematischer Informationen in einem 3D-Wissensraum.
- Ein so genanntes »Dialogmodul« zur Analyse und Umsetzung natürlichsprachlicher Fragestellungen mit folgenden drei Funktionen: Fragenanalyse, Suchdurchführung, Ergebnisaufbereitung, bei Bedarf noch Fortsetzung des Dialogs zur Eingrenzung der Ergebnisse bzw. zur Disambiguierung
- Eine für diese Zwecke optimierte Relevanz bewertende Suchmaschine
- Einen Thesaurus der im Projekt verwendeten Textdatenbank, für die Klassifikation von Texten und als Unterstützung bei der Fragenanalyse.
- Eine Indexierungssoftware, die mithilfe des Thesaurus die Texte der Projektdatenbank analysiert und so die Suchdurchführung des Dialogmoduls verbessert.
- Ein dreidimensionaler Wissensraum zur visuellen Repräsentation von Informationen aus der Ontologie und der Textdatenbank und deren jeweiliger Kontext.

Diese Elemente werden im Folgenden ausführlich dargestellt. Die im Rahmen des Projektes entwickelte Software spiegelt unterschiedliche Entwicklungsstadien wieder. In einigen Bereichen des Projekts wird man das erreichte Ergebnis als *Proof of Concept* bezeichnen dürfen (Beispiel Ober-/Unterbegriffsausgabe im 3D-Viewer), dessen Umsetzung für die komplette Lexikonsubstanz z. T. noch umfangreiche Datencodierungsarbeiten erfordern wird. In manchen Bereichen des Projekts kam es gegenüber dem zu Projektbeginn geplanten Vorgehen (vgl. Antragsstudie) zu Änderungen. Diese sind z. T. projektintern und sachlich begründet (Reaktion auf Zwischen-ergebnisse), z. T. extern (die Firma Tanto Xipolis GmbH gab das Projektgeschäft "Online-Nachschlagen / xipolis.net" auf). Im Detail sei hierzu auf die projektbegleitende Berichterstattung verwiesen. Stärkster Faktor für die Abweichungen im Vorgehen war letztlich eine kompromisslose Konzentration auf die beiden Kern-Projektziele, was im Rückblick durch die erreichten Ergebnisse mehr als gerechtfertigt erscheint.

Das eigentlich Bemerkenswerte an Aufgabenstellung und Ergebnis ist jedoch, dass die zentralen Fragestellungen wie z.B. die natürlichsprachige Abfrage gerade nicht nur Ergebnisse im Sinne eines *Proof of Concept* – ein Prototyp, der schon einmal mit 5 Fragen umgehen kann – sinnvoll erarbeitbar waren oder sind. Vielmehr hat sich gezeigt, dass eine diesbezüglich in der Breite funktionierende Technologie von vorneherein nur auf „breiten Echtdaten“ entwickelbar, und nicht als „Laborexperiment im Prinzip“ lösbar war.

Der nicht hoch genug einzuschätzende wissenschaftliche Wert des Ergebnisses liegt darin, dass bisherige Systeme mit vergleichbaren Abfragemöglichkeiten die damit verbundenen typischen Probleme (v.a. Vieldeutigkeit und Kontextabhängigkeit natürlicher Sprache) i.d.R. durch eine grundsätzliche Eingrenzung der Domäne umgingen (Expertensysteme). Das vorgelegte Projektergebnis geht demgegenüber mit dem gesamten enzyklopädischen Wissenskreis um und steht unseres Erachtens damit zurzeit ohne Vergleich da.



Eine Übersicht über die im Projektrahmen entwickelten Softwarekomponenten. Diese und ihr Zusammenspiel werden im Folgenden erläutert.

3 Ontologie der deutschen Sprache

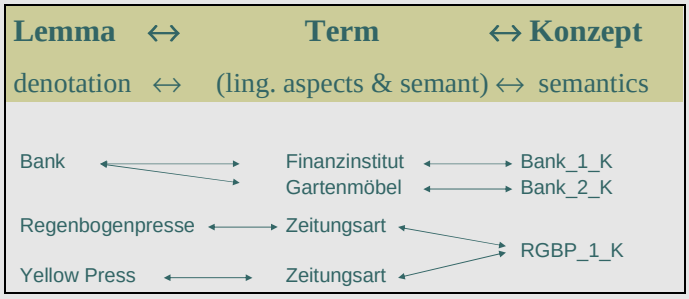
Als Grundlage dieses Projektes wurde eine Ontologie der deutschen Sprache aufgebaut. In ihr sind die grammatikalischen Informationen zu über ca. 250.000 Wörter abgelegt, die Bedeutungen dieser Wörter definiert und Bedeutungen untereinander verknüpft ("Taxonomie"). Ziel war die Repräsentation von sprachlichem und lexikalischem Wissen in einer Datenbank. Die Ontologie war von Anfang an zur Verwendung in beiden Projekt-Modulen vorgesehen: Im Dialogmodul sollen Begriffe aus der Ontologie für die Suche verwendet werden, im 3D-Wissensraum sollen sachsystematische Informationen aus der Ontologie direkt dem Benutzer angezeigt werden (s. u.)

3.1 Software und Datenmodell

Für den Aufbau der Ontologie wurde die Software *K-Infinity* des Unternehmens *intelligent views* in Darmstadt verwendet. *K-Infinity* ist eine objektorientierte Datenbank, deren Struktur für Wissensnetze optimiert ist. Mithilfe von *K-Infinity* ist eine objektorientierte Modellierung des Wissens durch Relationen auf semantischer Grundlage möglich.

Das grundsätzliche Datenmodell der Ontologie, gemeinsam entwickelt von Mitarbeitern der *Brockhaus Duden Neue Medien GmbH*, der *Dudenredaktion*, dem *FhG-IPSI-Institut* und *intelligent views*, stellt eine bedeutende Innovation dar, da es erstmalig in dieser Weise Sprach- und Sach-Aspekte von Begriffen in einem einheitlichen Framework erfasst und als Entität maschinell verarbeitbar macht. Es wurde der wissenschaftlichen Öffentlichkeit auf der LREC-Konferenz präsentiert und im zugehörigen Tagungsband veröffentlicht:

Melina Alexa, Bernd Kreissig, Martina Liepert, Klaus Reichenberger, Lothar Rostek, Karin Rautmann, Werner Scholze-Stubenrecht, Sabine Stoye: The DUDEN ontology. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC '2002). 2002, Las Palmas. Published by European Language Resources Association (ELRA).



Das Datenmodell unterscheidet zwischen Benennung mit den sprachbezogenen Aspekten eines Wortes und Bedeutung, das die konzeptionellen Aspekte repräsentiert: Das Lemma (Stichwort) ist die Benennung, z.B. das Homonym „Bank“, der Term umfasst Benennung und Bedeutung, also "Bank im Sinne von Finanzinstitut" bzw. "Bank im Sinne von Sitzgelegenheit". Das Konzept für diese Benennungen, also „Unternehmen, das Geld- und Kreditgeschäfte betreibt und den Zahlungsverkehr vermittelt“ bzw. „Sitzgelegenheit aus Holz, Stein o. Ä., die mehreren Personen nebeneinander Platz bietet“ wird unter einer Kennung, z.B. Bank_1_K bzw. Bank_2_K abgespeichert. In einem anderen Fall haben die beiden Lemmata „Regenbogenpresse“ und „Yellow Press“ unterschiedliche Terme, deren Konzept aber identisch ist und mit der Kennung Regenbogenpresse_1_K beschrieben wird; die beiden Lemmata sind also synonym zueinander.

Das Wissen wird auf drei Ebenen repräsentiert: Begriffe sind verallgemeinerte physikalische (z.B. Stuhl oder Tisch) und abstrakte (z.B. Verein oder Gefühl) Objekte, aber auch Situationen

und Ereignisse. Relationen und Attribute sind die möglichen Eigenschaften und Zusammenhänge der Begriffe mit anderen Objekten; Beispiele für Relationen sind hat-Teil, hat-Mitglied, verursacht oder gefolgt-von. Durch die hierarchischen Relationen geschieht die Systematisierung zwischen verschiedenen Begriffen. Individuen schließlich benennen konkrete Individuen und ihre realen Zusammenhänge wie der Fußballverein Werder Bremen oder der Komponist Ludwig van Beethoven, aber auch Dinge wie der Stuhl in einem bestimmten Zimmer. Individuen sind zu Begriffen durch die Relation Instanz-von verbunden.

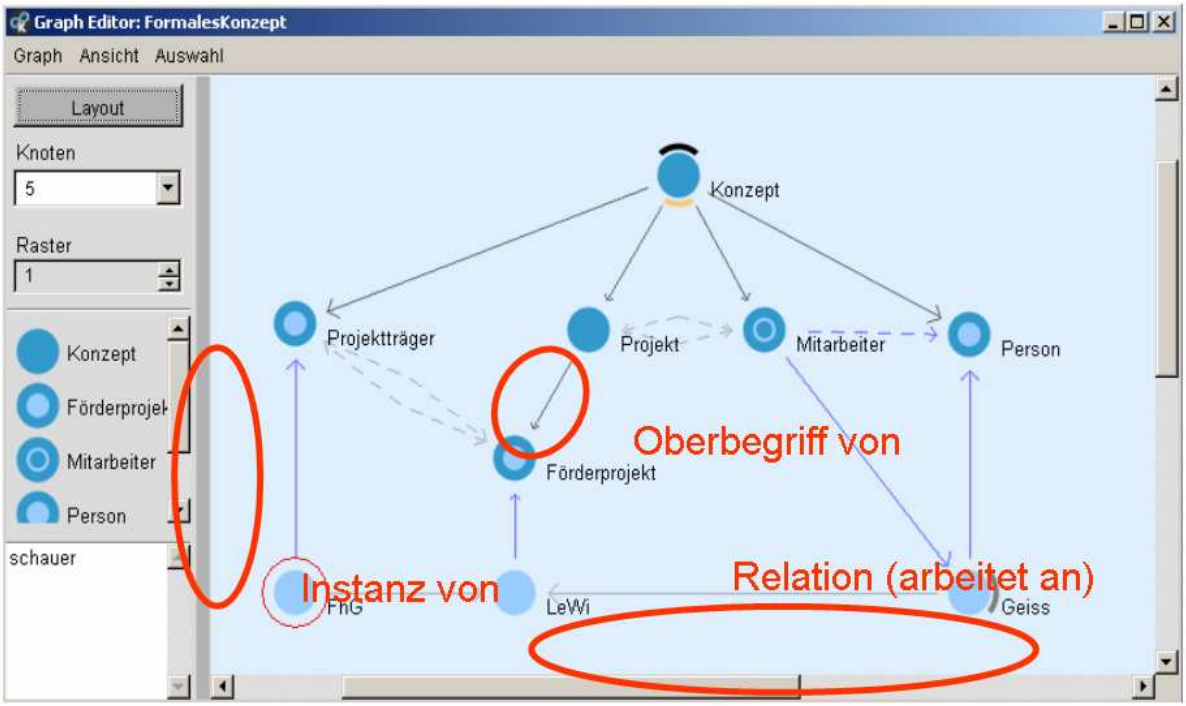
3.2 Datenquellen für die Ontologie

Die Grundlage für die Ontologie ist das Werk *Duden das große Wörterbuch der deutschen Sprache*. Dieses Werk mit seinen etwa 230.000 Stichwörtern mit 250.000 Termen bot sich aufgrund der vielfältigen Angaben zu Bedeutung, Herkunft, Gebrauch und Grammatik und seiner reichen Strukturierung mithilfe einer SGML-DTD als erste Datenbasis an. Die Daten wurden aus der SGML-Struktur extrahiert und in die Datenbank integriert. Den Stichwörtern wurden etwa 380 verschiedene Grammatik- und ca. 1000 Pragmatikklassen zugeordnet. Zur Strukturierung der Daten werden zwischen den Einträgen ca. 30 unterschiedliche Relationen verwendet, welche die Beziehungen zwischen den Konzepten definieren.

Da die Bedeutungsrelationen *Synonym*, *orthografische Variante*, *Oberbegriff* und *Unterbegriff* für die beiden Anwendermodule vorrangig gebraucht wurden, sollten diese über die gesamte Breite der Ontologie automatisch erzeugt werden. Dazu hat das *Fraunhofer Institut für Integrierte Publikations- und Informationssysteme (FhG-IPSI)* die Definitionstexte der Konzepte mithilfe computerlinguistische Methoden analysiert und untersucht, ob sich aus den Definitionen die genannten Relationen zu anderen Begriffen erkennen lassen. Wurden die Informationen erkannt, wurden die durch die jeweilige Relation verbundenen Begriffe in einer Tabelle hinterlegt. *intelligent views* importierte anschließend die Tabellen nach der Prüfung durch Redakteure der *Dudenredaktion*.

Es war von Beginn absehbar, dass die automatische Extraktion von Synonymen aus dem *Wörterbuch der deutschen Sprache* die möglichen sprachlichen Ausdrucksvarianten ("ODER-Suchen", s.u.) nicht ausreichend abdecken würde. Daher wurden von Redakteuren der *Dudenredaktion* die Synonyme aus dem Werk *Duden - die sinn- und sachverwandten Wörter* für den Import in die Ontologie neu bearbeitet und wenn nötig vereindeutigt; die Daten wurden anschließend von *intelligent views* in die Ontologie importiert.

Da das *Wörterbuch der deutschen Sprache* jedoch im fachsprachlichen Bereich nicht so sehr in die Tiefe geht und kaum geografische Namen und Personennamen enthält, wurde die Ontologie mit Daten aus dem *Brockhaus in 15 Bänden* und der *Brockhaus Enzyklopädie* erweitert. Aus den biografischen und geografischen Lexikonartikeln wurde durch das *FhG-IPSI* grundlegende Fakten extrahiert, bei Biografien z.B. Funktionen (englischer Physiker) und Lebensdaten, bei geografischen Artikeln z.B. Verortung (Aschaffenburg liegt in Bayern) und Zahlenwerte (Einwohner, Höhen, Längen). Die Angaben zur Funktion und Verortung werden im dreidimensionalen Wissensraum als sachsystematische Information dem Benutzer angezeigt. Darüber hinaus können die Informationen unter besonderen Umständen in einem natürlichsprachlichen Dialog abgefragt werden. Zur Extraktion der Fakten wurden sowohl die vorliegende SGML-Struktur der Texte analysiert als auch eine spezielle Analysegrammatik angewendet. Diese Methoden könnten in einem vertiefenden Projekt zur Faktenextraktion aus weniger strukturierten Texten weiter entwickelt werden.



Screenshot des eingesetzten Ontologie-Bearbeitungswerkzeugs Graph-Editor der K-Infinity-Suite. Die modellierten Instanz- und Unterbegriff-Relationen kommen in der LeWi-Software u.a. direkt in der 3D-Visualisierungskomponente zum Einsatz (vgl. Abschnitt 6).

In den lexikalischen Texten sind bei vielen Einträgen Abkürzungen und Synonym-ähnliche Begriffe vorhanden, die durch ein eigenes SGML-Element gekennzeichnet sind und sich einfach herauslesen lassen. Die entstandenen Listen wurden redaktionell von Redakteuren der *Brockhausredaktion* und Mitarbeitern der *Brockhaus Duden Neue Medien GmbH* bearbeitet, vor allem bei den Synonym-ähnlichen Begriffen, da in den Synonym-Elementen nicht immer nur Synonyme im engeren Sinne eingetragen sind, sondern auch orthografischen Varianten, bei Personen z. B. Künstlernamen oder bei geografischen Namen z. B. amtliche landessprachliche Bezeichnungen. Nach Abschluss der Arbeiten konnten ca 17.000 Synonym-ähnliche Einträge in die Ontologie importiert werden.

Viele Sachartikel, deren Lemma aus mehreren Wörtern bestehen, z. B. „Christlich Demokratische Union“ oder „Astronomische Einheit“, sind nur in den Lexikontexten vorhanden. Diese ca. 6000 Lemmata wurden aus dem Lexikon extrahiert und durch Mitarbeiter der *Brockhaus Duden Neue Medien GmbH* den vorhandenen Begriffen in der Ontologie wie „Partei“ oder „Längeneinheit“ zugeordnet und anschließend in die Ontologie importiert. Weiterhin wurden ca. 3.000 Abkürzungen und ihre Auflösungen, die aus dem *Brockhaus in 15 Bänden* aufgrund eines eigenen SGML-Elements für Abkürzungen leicht extrahierbar waren, in die Ontologie importiert.

4 Textdatenbank, Thesaurus, Klassifikation und Indexierungssoftware

Für eine sinnvolle Entwicklung der beiden Anwendermodule, dem Dialogmodul und des 3D-Wissensraums, war eine Textdatenbank als Testobjekt sowie zur Aufbereitung der Daten eine Indexierung der Texte notwendig.

4.1 Textdatenbank des Projektes: Brockhaus multimedial 2004

Als Textdatenbank, mit der die im Folgenden beschriebenen Module getestet wurden, diente der Brockhaus multimedial Premium („BMM“) 2004, ein multimediales Lexikon mit 17 Millionen Wörtern in ca. 120.000 Lexikonartikeln, 6.500 ausführlichen Zusatztexten und Quellen und ca. 600 Tabellen.

Die zugrunde liegenden Texte sind in einem SGML-basierten Redaktionssystem erstellt und mit einer tief gehenden Document Type Definition (DTD) strukturiert, die vielfältige Auswertungen von Textelementen erlaubt. Grundsätzlich unterschieden wird auch zwischen verschiedenen Artikeltypen wie Sachartikel, Personenartikel und geografische Artikel, die jeweils ihre eigene DTD-Struktur haben. Die Information über die Artikeltypen wurden sowohl im Dialogmodul bei einigen Suchen als auch bei der Darstellung der Artikel und ihres Kontextes verwendet (siehe unten).

Der Brockhaus multimedial deckt als universelles Lexikon alle Wissensgebiete in annähernd gleicher Tiefe ab. Die beiden Module ließen sich daher für alle Sachgebiete in gleichem Maße testen; insbesondere das Problem von mehrdeutigen Begriffen konnte so anhand vieler Beispiele bearbeitet werden. Der Brockhaus multimedial bot sich für den Test des Dialogmoduls auch an, da aufgrund der großen Textmenge die Wahrscheinlichkeit groß ist, dass viele Fragen auch anhand der umfangreichen Texte beantwortet werden können.

4.2 Thesaurus, Klassifikation und Indexierungssoftware

Für eine effizientere Suche der Artikel durch das Dialogmodul war es notwendig, alle Artikel und Medien des Brockhaus multimedial thematisch einzuordnen, sie zu klassifizieren. So zielen bestimmte Suchen auf Artikel mit definierten Klassen ab; weiterhin ermöglicht die Zuordnung der Artikel zu thematischen Klassen dem Benutzer die vom Dialogmodul ausgewählten Artikel thematisch zu sortieren (siehe unten). Für die Darstellung von Kontexten im Wissensraum, hier die Relevanz zwischen Artikeln, spielt die Klassifikationsinformation in dem Algorithmus zur Berechnung der inhaltlichen Nähe ebenfalls eine wichtige Rolle. Um die große Zahl der Artikel zu klassifizieren, wurde eine Indexierungssoftware entwickelt.

4.2.1 Thesaurus des Brockhaus multimedial

Der Thesaurus des Brockhaus multimedial (BMM) ist ein Fachthesaurus, der für die Bedürfnisse von Information und Dokumentation entwickelt wurde. Obwohl ein Fachthesaurus lassen sich mit ihm sämtliche Themenfelder einer Enzyklopädie abdecken. Die Besonderheit des BMM-Thesaurus liegt in der Annotierung jedes Begriffs mit einer oder mehreren Brockhaus-Grob- bzw. Feinklassen, in denen dieser schon einmal in einem Artikeltext vorkam. Aufgrund der Gesamtheit dieser Annotationen kann bei einem neu zu klassifizierenden Artikeltext die wahrscheinliche Klassenzugehörigkeit errechnet werden.

Der BMM-Thesaurus besteht aus vier Teilen (wissen.mas, wissen.namen.mas, wissen.countries, wissen.cities). Der Thesaurus enthält insgesamt zurzeit ca. 300 000 Einträge.

Folgende Features sind in einem Thesauruseintrag vorhanden (einige fakultativ):

- string: Stichworteintrag
- c (category): Angabe der Wortkategorie; folgende Werte können vorkommen: „desc“ (Deskriptor), „adj“ (Adjektiv), „verb“ (Verb) bzw. „rename“ (bei Synonymen Verweis auf den Repräsentanten)
- besser: Repräsentant der Synonymgruppe
- lu (lexical unit): bevorzugte Schreibweise des Deskriptors
- fc (facets): Grobklassen entsprechend der Brockhaus-Klassifikation (Grobklassennummer mit vorangestelltem „h“); die Gewichtung ist abgestuft (erster Eintrag wird mit „1“ gewichtet, zweiter Eintrag mit „0,5“, dritter Eintrag mit „0,333“ etc.)
- nace: Feinklassen entsprechend der Brockhaus-Klassifikation (Feinklassennummer mit vorangestelltem „h“); die Gewichtung erfolgt analog der Grobklassengewichtung.
- hyper: zugehöriger Oberbegriff

- s (semantics): zum Stichworteintrag zugehörige semantische Kategorie (siehe im Anhang)
- imp (importance): Beeinflussung der Gewichtung; folgende Werte können vorkommen: „high“ (Eintrag wird mit dem Faktor „3“ gewichtet) bzw. „low“ (Eintrag wird mit dem Faktor auf den Faktor „0,33“ heruntergesetzt)
- country: Länderzugehörigkeit

Der Namensthesaurus ist etwas anders aufgebaut. „string“ ist dort der Familienname. Zusätzlich gibt es das Merkmal „fname“, das den Vornamen enthält.

Ein Eintrag des Thesaurus sieht z.B. folgendermaßen aus:

```
{string=schmelzeinheit,c=desc,lu=Schmelzeinheit,fc=h53,nace=h5361}
```

4.2.2 Brockhaus-Klassifikation

Die Klassifikation wurde im Rahmen des Projektes ausgebaut und differenziert, sodass die Brockhaus-Klassifikation jetzt aus 82 Grobklassen und 2358 Feinklassen besteht. Das Klassifikationsschema unterscheidet grundsätzlich zweiziffrig benannte Grobklassen und vierziffrig benannte Feinklassen. Die Grobklassen orientieren sich in etwa an den klassischen Wissenschaften wie Physik, Philosophie, Musik oder Soziologie. Einige Wissensgebiete sind aber auch schon auf der Ebene der Grobklassen unterteilt, etwa die Technik unter anderem in Elektrotechnik, Bautechnik oder Maschinenbau und die Geschichte in unterschiedlich Epochen wie Vorgeschichte oder Mittelalter. Für geografische Objekte (Städte, Berge, Flüsse etc.) ist für jeden Kontinent eine eigene Grobklasse eingerichtet.

Die Feinklassifikation unterhalb der Grobklassen unterscheidet im kulturellen Bereich neben sachsystematischen Gesichtspunkten (z.B. Glasmalerei) häufig vor allem nationale Aspekte (z.B. französische Malerei). In den anderen Fachgebieten sind die Feinklassen meist nach den Teilgebieten der jeweiligen Wissenschaften unterteilt, ist oft aber auch unterhalb dieser Ebene noch gegliedert. Die Geschichte wird neben sachsystematischen Gesichtspunkten auch nach Zeiträumen unterteilt, die Feinklassen für Tiere und Pflanzen orientieren sich an den großen systematischen Einheiten, in der Geografie hat jeder Staat seine eigene Feinklasse, für Deutschland, Österreich und die Schweiz sogar jedes Bundesland bzw. Kanton. Eigene Feinklassen gibt es in vielen Grobklassen für Personen, Organisationen und Preise, in den Naturwissenschaften für Einheiten.

Grob- und Feinklassifikations-Systematik: beispielhafter Auszug

00 Sprachen, Schriften
- 0002 Organisationen, Institutionen, Preise, Zeitschriften (in den Sprachwissenschaften)
- 0010 Sprachwissenschaft (allgemeine Begriffe)
- 0011 klassische Philologen
- 0012 Sprachwissenschaftler
- 0020 Phonologie, Phonetik
- 0021 Morphologie, Morphemik
- 0022 Syntax
- 0023 Semantik
- 0024 Lexikologie, Wortbildung

[…]

42 Zivilrecht
- 4200 Zivilrecht (allgemeine Begriffe)
- 4203 Juristen (für Zivilrecht)
- 4210 Allgemeiner Teil des Bürgerlichen Gesetzbuches (BGB), Schuldrecht
- 4220 Mietrecht
- 4230 Arbeitsrecht

4240 Sachenrecht
4250 Familienrecht
4260 Erbrecht
4270 Zivilprozess
4271 Zwangsvollstreckung
4280 Sonstiges Zivilrecht

[…]

4.2.3 Indexierungssoftware

Die unter Linux lauffähige Indexierungssoftware (bestehend aus 30 Einzelprogrammen) basiert auf der vom *Institut der Gesellschaft zur Förderung der Angewandeten Informationsforschung e.V. (IAI)* entwickelten Software AUTINDEX. Das im Projekt entstandene BROCKINDEX gliedert sich in fünf Hauptprogrammteile, die hintereinander ablaufen. Nach der linguistischen Analyse und der Evaluierung der Textelemente werden die Wortgruppen durch ein Oberflächenparsing ermittelt und die Klassifikation bestimmt mit anschließender Ergebnisausgabe.

Linguistische Wortanalyse

Der Kern des linguistischen Analysetools MPro besteht aus einem Morphemlexikon sowie Regeln zur Kombinierbarkeit der Morpheme. Die Grundidee von MPro ist, dass sich komplexe Wörter aus einfachen ableiten lassen, sodass ein komplexes Wort automatisch mit einer Paraphrase aus seinen Bestandteilen erklärt werden kann. Das lexikalische Material für das Deutsche (jetzt etwa 90.000 Morpheme, die von etwa 78.000 Allomorphen repräsentiert werden) stammt nicht aus handelsüblichen Lexika, sondern aus verschiedenen Textsorten. Es ist dadurch der realen Alltagssprache näher als etwa in Lexika über einen langen Zeitraum gesammeltes Material, das oft zum großen Teil aus literarischen Quellen stammt.

Das Lexikon ist eine Menge von Einträgen, die als Merkmalbündel kodiert sind. Jeder Lexikoneintrag beinhaltet die Angabe eines Strings, der das jeweilige Allomorph darstellt. Diese Strings werden bei der Zerlegung der Wortformen verwendet.

Jede Wortform ist nach Beendigung der Analyse mit zwei Informationsblocks verknüpft, dem Merkmalbündel für die flexionsmorphologischen Features und der Wortstruktur (Derivation und Komposition). Bei Substantiven und Verben werden die üblichen morphosyntaktischen Merkmale ermittelt, bei Adjektiven wird nur die Endung (e, es, em, en, er) festgehalten, da deren Informationsgehalt je nach den Wünschen der Anwender dargestellt wird.

Die Wortstruktur ist eine syntaktisch-semantische Interpretation der gegebenen Wortform, wenn es sich um ein komplexes Wort handelt. Sie ist identisch mit der Zitatform des Wortes, wenn das Wort einfach ist. In jedem Fall wird auch die Zitatform berechnet ("Lemma").

Ermittlung von Wortgruppen durch syntaktisches Oberflächenparsing

Für diesen Schritt muss eine Bestimmung der Strukturierung des Satzes in Teilgruppen vorgenommen werden, ein so genanntes „Oberflächenparsing“. MPro verwendet hier eine externe Grammatik, die die Erkennung syntaktischer Strukturen des Deutschen beschreibt. Diese Grammatik ist in Teilgrammatiken gegliedert, die sequenziell ablaufende Regeln mit Kontextspezifikationen enthalten. Durch diese Strategie entsteht ein robustes Verfahren, das zwar durch die Nichtverfolgung von Mehrdeutigkeiten eine gewisse Unschärfe in Kauf nimmt, jedoch in der überwiegenden Anzahl der Fälle Ergebnisse liefert, die auch bei partieller Analyse gut brauchbar sind.

Evaluierung der Textelemente

Nach der morphosyntaktischen Analyse werden die Textelemente, die zu den Wortklassen Nomen, Verb und Adjektiv gehören (den so genannten bedeutungstragenden Wortklassen), daraufhin analysiert, zu welchen semantischen Klassen sie mehrheitlich gehören. Wörter (oder Wortteile von Komposita), deren semantische Klasse z.B. nur ein einziges Mal im Text vorkommt, werden dann

von der weiteren Betrachtung ausgeschlossen. Es werden nur die häufigsten semantischen Klassen für die Weiterverarbeitung benutzt.

Ermittlung der Klassifikation mithilfe des BMM-Thesaurus

Nach der morphosyntaktischen Analyse ermittelt BROCKINDEX im Abgleich mit dem BMM-Thesaurus die möglichen Brockhaus-Klassifikationen für den Text. Dies erfolgt unter Berücksichtigung der im Thesaurus für die vorkommenden Wörter annotierten möglichen Klassifikationen, den teilweise schon zu vorhandenen, händisch vergebenen Klassen sowie anhand eines komplexen Regelwerks.

Ergebnisausgabe

Als Ausgabe werden dann folgende Elemente präsentiert: Aus dem Text ermittelte Namen von Personen und Firmen, eine Liste der Deskriptoren (mit Gewichtung), eine Liste der Klassifikationen, eine Liste der Oberbegriffe, eine Liste der nicht im Thesaurus enthaltenen Schlüsselwörter (als Vorschlagsliste) und Ländercodes (falls ein oder mehrere Staaten erwähnt wurden).

Beispiel eines analysierten BMM-Artikels:

Florianópolis,
Hauptstadt des brasilianischen Bundesstaates Santa Catarina, 271 300 Einwohner; Hafenstadt auf der durch eine Brücke mit dem Festland verbundenen Insel Santa Catarina; Erzbischofssitz; Universität; Textil- und Nahrungsmittelindustrie.

Topography: Florianópolis, Hauptstadt, -
Names of persons and organisations: Bundesstaat Santa Catarina; Florianópolis; Insel Santa Catarina
Toponyms: Florianópolis
Descriptors: Florianopolis[75]; Bundesstaat[30]; Hafenstadt[30]; Santa Catarina[28]; Brücke[24]; Insel[20]; Hauptstadt[15]; Festland[6]; Brasilien[4]; Einwohner[4]; Nahrungsmittelgewerbe[4]; Textilindustrie[4]; Universität[4]; katholische Kirche[4]
Special fields: H3611[100] (Brasilien)
Facets: H36[100] (Staaten Amerika); H30[38] (Geographie)
Oberbegriffe: Stadt[99]; Verkehrsinfrastruktur[16]
Countries: Brasilien C3BRAZ[14]
Possible terms: brasilianischer Bundesstaat; verbundene Insel
<manfein=3611> Brasilien </man>
<mangrob=36> Staaten Amerika </man>
<autofein=3611> Brasilien </auto>
<autogrob=36> Staaten Amerika </auto>
<RECALL> 100 </RECALL>
<PRECISION> 100 </PRECISION>

5 Dialogmodul für natürlichsprachliche Fragen zur Suche in Volltextdatenbanken

Das so genannte »Dialogmodul« ist der Kern des entwickelten Systems zur Beantwortung natürlichsprachlicher Anfragen in Volltextdatenbanken. Es hat ganz allgemein die Aufgabe, zwischen dem Informationswunsch des Benutzers und dem gespeicherten Wissensbestand zu vermitteln. Dieses Dialogmodul organisiert zum einen die Kommunikation mit dem Benutzer, zum anderen stellt es die technischen Verfahren bereit, mit der die computerlinguistische Analyse der Frage und die datenbankgestützte Recherche durchgeführt werden können. Im Dialogmodul werden dazu verschiedene in diesem Projekt entwickelte Funktionen zur Beantwortung natürlichsprachlicher Fragen integriert: erst die Fragenanalyse, danach die Suchvorbereitung und –durchführung und schließlich die Ergebnisaufbereitung.

Das in dem Projekt entwickelte Dialogmodul soll eine Art „Gespräch“ zwischen Benutzer und Computer schaffen. Ziel sollte dabei aber nicht sein, dass das Modul ein Feedback wie ein Mensch gibt und am Ende eine finale Antwort mitteilt, sondern vielmehr über normalsprachliche Antworten den Fragesteller zur Übermittlung seiner Intentionen und weiterer Informationen zu bewegen, um mit immer feineren Kriterien passende Texte ermitteln zu können. Eine solche Gesprächssituation soll dem Benutzer nicht nur das Gefühl geben, dass er verstanden wird, sondern auch helfen, die Komplexität der Aufgabenstellung zu bewältigen, da das System das vom Benutzer Gemeinte im ersten Schritt meist nicht komplett erfassen kann und somit weitere Anhaltspunkte für die Suche notwendig sind. Bei der Benutzung des Dialogmoduls sollte auch eine Art "Lernprozess" beim Benutzer eingeleitet werden, der bei ihm ein Bewusstsein zur Notwendigkeit der Präzisierung seines Informationswunsches schaffen soll.

5.1 Fragenanalyse

Die Fragen werden vom Dialogmodul sowohl typologisch als auch morphosyntaktisch analysiert. Die typologische Analyse versucht, die Art der Frage zu bestimmen. Die morphosyntaktische Analyse ordnet die sprachlichen Elemente des Fragesatzes und unterscheidet deren Rolle, z.B. das (gesuchte) Subjekt von einer adverbialen Bestimmung.

5.1.1 Funktion und Struktur der Fragentypologie

Grundlage für die Funktion des Dialogmoduls ist die Interpretation der Fragestellung, also die Durchführung einer Sprachanalyse. Dazu wurde eine Fragetypologie entwickelt, mit deren Hilfe so viele Informationen wie möglich gewonnen werden, um den Austausch zwischen Benutzer und Computer in Gang zu setzen. Ziel war es, für jeden Fragentyp eine eigene Reaktionsweise festzulegen. So wird z.B. bei einer Personenfrage ("Wer erfand das Telefon?") die auf biografische Texte beschränkte Suche (s. u.) in der Relevanzbewertung aufgewertet oder bei einer Geschwindigkeitsfrage („Wie schnell rennt ein Gepard?“) auch mit den Wörtern „Stundenkilometer“ und „km/h“ gesucht.

Die zunächst 600 Beispielfragen, die zu Beginn des Projektes gesammelt wurden, wurden nach Art und Gegenstand der gewünschten Antwort klassifiziert und für jeden Fragentyp eine formale Beschreibung entwickelt. Gleichzeitig wurde geprüft, wie sich die Fragewortverwendung über Verben, Adjektive und erkennbare Objekte differenzieren und den einzelnen Fragetypen zuordnen ließ. So werden zum Beispiel die Was-Fragen in die Frageindikatoren „Was_sein_Name?“ für Namensfragen, „Was_für_ein_geographisches Objekt?“ für Ortsfragen und „Was_sein_Ursache/Grund/Voraussetzung?“ für Ursachenfragen differenziert. Das Ergebnis der Bemühungen sind 22 verschiedene Fragentypen, mit denen sich gut 95% der Testfragen identifizieren und für eine Weiterverarbeitung klassifizieren lassen.

Beispiele für Fragentypen:

– Methoden-/Funktionsfragen

Warum fliegt ein Flugzeug?
Wieso werden Blätter im Herbst bunt?
Aus welchem Grund ist der Abendhimmel rot?
Aufgrund welcher Einflüsse kommt es zum El Nino?
Wie funktioniert ein Ottomotor im Auto?

– Zusammenhangsfragen

Wie hängen hygienische Verhältnisse mit bestimmten Krankheiten zusammen?
In welcher Beziehung stand die griechische Philosophie zur griechischen Kunst?
Wie war der Einfluss der Literatur auf die französische Revolution?
Was verbindet Thomas Mann und Venedig?"

– Namensfragen

Wie heißt das Käppchen, dass Juden tragen?
Wie lautet ein anderes Wort für Refrain?
Wie ist die exakte Bezeichnung des Herrschers im mittelalterlichen Deutschland?
Wie nennt man die Behausung der Biber?
Was ist der Name für das Parlament in Island?

– Ortsfragen

Wo wird das meiste Gold gefördert?
Wo liegt Bad Gandersheim?
Welcher Fluss ist der längste in Europa?
An welchem Fluss liegt Ulm?
Aus welchem Land stammte Mutter Teresa?
In welchem Staat leben die meisten Muslime?

– Bedeutungsfragen

Welchen Stellenwert haben die Menschenrechte in den Kulturen Asiens?
Was für eine Bedeutung hatte das Attentat von Sarajevo auf die österreichisch-serbischen Beziehungen?
Wie wichtig ist der Bundesrat für die Gesetzgebung in Deutschland?
Was bedeutet das Abendmahl für die Katholiken?

5.1.2 Morphosyntaktische Fragenanalyse

Die eingegebenen Fragen werden zunächst einer morphosyntaktischen Analyse unterzogen. Dazu wurde eine spezielle Grammatik zur Analyse von Fragesätzen entwickelt, die auf Besonderheiten von Fragen eingeht. Besondere Aufmerksamkeit wurden dem Umgang mit Fragenwörtern, Inversionen, abgetrennten Verbpräfixen, Homonymen, Nebensätzen, Ausdrücken in Anführungszeichen, Ordinalzahlen und römischen Zahlen gewidmet. Erkannt werden die unterschiedlichen Wortklassen und semantischen Klassen (zurzeit ca. 140), die teilweise übergeordneten semantischen Hauptklassen wie Person, Zeit, Objekt oder geografisches Objekt zugeordnet werden. Anschließend werden Fragetypindikatoren und die ihnen dazugehörigen Fragetypen aus einer speziellen Datei ausgelesen und die Frage einem oder mehreren Fragetypen zugeordnet.

Am Ende der Analyse wird eine so genannte „logische Form“ der Frage und der dazugehörige Fragentyp ausgegeben. Auf die Frage „Was war die Ursache des chinesischen Boxeraufstandes?“ hat diese zum Beispiel folgendes Aussehen:

```
sein< >[past] (( {[Ursache< >[object],noun]}, {[chinesisch< China>,adj],
[Boxeraufstand< boxen;aufstand>,noun]}),([was< >,interr1]))
```

Ursachenfrage: singuläres Ereignis

Bei der Frage „Wann begann der chinesische Boxeraufstand?“ dagegen

```
beginnen< >[past] (( {[chinesisch< China>,adj],[Boxeraufstand< boxen;aufstand>,noun]}),
([wann< >,wh_adv])).
```

Zeitfrage

Die logische Form ist immer gleich aufgebaut. Außer dem Hauptverb des Satzes, welches am Anfang der logischen Form steht, sind alle Begriffe in eckige Klammern gesetzt. Jedem Begriff sind bestimmte Eigenschaften zugeordnet. Dies sind die Ableitungen des Begriffs (in gewinkelten Klammern) sowie durch Komma abgetrennt die Wortart. In manchen Fällen erscheint in einer zusätzlichen eckigen Klammer noch das semantische Feature (PERSON, OBJECT, TIME oder GEO) des Begriffs. Die geschweiften Klammern drücken weitere strukturelle Zusammenhänge wie z.B. Adjektiv-Substantiv-Konstruktionen aus.

Auch eine Indexierung der Frage mithilfe eines Thesaurus wurde im Projektverlauf experimentell unter UNIX realisiert. Wegen der bevorstehenden Einbindung der Ontologie (die in ähnlicher Weise eine Expansion der Suchfrage vornimmt) wurde dieser Ansatz nicht weiterverfolgt. Dies bedeutet, dass diese Funktion in der Abschlussversion von LeWi nicht mehr enthalten ist. Die Realisierung einer DLL-Version unter Windows hätte zudem einen erheblichen Zusatzaufwand bedeutet, der im Rahmen des Projektes nicht geleistet werden konnte. In der Abschlussversion wird also zu jedem erkannten Suchwort noch in der Ontologie optional nach Oberbegriffen und Synonymen mitgesucht.

5.2 Suchen in der Datenbank

Prinzipiell erfolgt im Dialogmodul nicht nur eine Suche, in der nur die Suchwörter aus der Frage vorkommen, sondern mehrere, umso die Chance zu erhöhen, die passenden Artikel zu finden. Außerdem sollten durch verschiedenen Suchtypen die Suchräume eingeschränkt werden, um die Präzision der Antworten zu erhöhen. Insgesamt wurden ca. 360 Suchen definiert, von denen bei einer Frage, aber meist nur zwischen 5 und 20 Suchen auch tatsächlich stattfinden.

Alle Suchen wurden in Form von http-Strings an die Datenbank übergeben, die dort die entsprechenden Suchen auslöste. Die Ergebnisse werden dann im Form von Ergebnislisten zur weiteren Bearbeitung an das Dialogmodul zurückgegeben.

Ein Hauptparameter für die Suche war die in der Frage vorkommenden Wortklassen, ein anderes die jeweilige Analyseform. Daraus leiteten sich die Hauptsuchformen ab.

S0: Suche aller eingegebenen Substantive, Adjektive, Verben, Adverbien, Präpositionen, Namen, Zahlen und Daten in den jeweiligen Originalformen, sofern sie in der Frage vorkommen.

S1: Wenn eine Präposition Teil der Frage ist: Suche von Substantiven, Adjektiven, Verben, Adverbien, Präposition, Namen, Zahlen und Daten in den Grundformen, sofern sie in der Frage vorkommen.

S2: Wenn ein Adjektiv und ein Verb Teil der Frage ist: Suche von Substantiven, Adjektiven, Verben, Namen, Zahlen und Daten in den Grundformen, sofern sie in der Frage vorkommen.

S3: Wenn ein Adjektiv Teil der Frage ist: Suche von Substantiven, Adjektiven, Namen, Zahlen und Daten in den Grundformen, sofern sie in der Frage vorkommen.

S4: Wenn ein Verb Teil der Frage ist: Suche von Substantiven, Verben, Namen, Zahlen und Daten in den Grundformen, sofern sie in der Frage vorkommen.

S5: Wenn ein Substantiv Teil der Frage ist: Suche von Substantiven, Namen, Zahlen und Daten in den Grundformen, sofern sie in der Frage vorkommen.

S6: Wenn ein Substantiv Teil der Frage ist: Nach einer Frequenzanalyse, Suche nach dem am seltensten vorkommenden Substantiv, Name, Zahl und Datum in den Grundformen, sofern sie in der Frage vorkommen.

S7: Wenn ein Kompositum Teil der Frage ist: Suche von Substantiven (Komposita zerlegt), Adjektiven, Verben, Namen, Zahlen und Daten in den Grundformen, sofern sie in der Frage vorkommen. Die Suchen S2 bis S5 werden hier als Untersuchen wiederholt.

S8: Suche mit den Ableitungen von Substantiven, Adjektiven und Verben nach einer morphologischen Analyse, sofern sie in der Frage vorkommen. Die Suchen S2 bis S5 werden hier als Untersuchen wiederholt.

In den Suchstrings werden Partikel, Artikel, Pronomen und Konjunktionen sowie definierte Stoppwörter nicht verwendet. Die in der Ontologie gefundenen Synonyme und orthografischen Varianten werden zu dem jeweils in der Frage vorkommenden Suchwort in einer ODER-Verknüpfung gesucht.

Ein weiterer Parameter in den Suchen sind die verwendeten booleschen Operatoren. Alle Suchen werden grundsätzlich mit dem UND-Operator durchgeführt. Für die Verengung des Suchraums und damit eines präziseren Ergebnisses werden NAHEBEI-Suchen durchgeführt, wobei der Wortabstand mit 25 Wörtern recht eng gewählt wurde; allerdings liefern bereits die UND-Suchen in Abhängigkeit von dem Abstand der Treffervorkommen unterschiedlich hohe Relevanz-Ergebniswerte. Unter bestimmten Umständen werden dagegen auch ODER-Suchen durchgeführt, um den Suchraum wiederum zu erweitern. Die NAHEBEI- oder ODER-Suchen werden aber nicht immer durchgeführt, sondern nur bei bestimmten Wortklassen in der Frage oder bestimmten Fragetypen.

Als weitere Suchen werden Assoziationssuchen definiert, das heißt Suchen, bei denen bestimmte Wörter noch zusätzlich in den Suchstring eingebaut werden, beispielsweise bei Gewichtsfragen, u.a. die Wörter Kilogramm, Gramm und die Abkürzungen kg und mg. Wurden Personen bzw. geografische Namen erkannt, wird nur in biografischen bzw. geografischen Artikeln gesucht (die in der Datenbank als solche gekennzeichnet sind). Bei historischen Fragen wird auch speziell nur in Artikeln gesucht, denen durch die Indexierungssoftware definierte historische Grob- und Feinklassen zugewiesen wurden. Schließlich werden auch noch Wörter aus dem Frageindikator wie zum Beispiel „Jahr“ bei dem Frageindikator „In welchem Jahr …?“ bei einigen Suchstrings eingebaut und bei anderen nicht.

5.3 Ergebnisaufbereitung

Nach einer erfolgreich durchgeführten Suche wird dem Benutzer für jeden gefundenen Artikel wird anschließend ein Relevanzwert ausgegeben. Das Dialogmodul kommuniziert mit dem Benutzer über ein oder zwei Sätze und bietet die Möglichkeit, den Dialog fortzuführen oder eine neue Frage einzugeben.

Testen Sie selbst! – Falls Sie Zugang zu der Software haben, finden Sie nachfolgend eine Auswahl von Fragen, die beim Testen durch besonders schöne Ergebnisse auffielen. Die Software wurde natürlich nicht auf diese Fragen speziell getrimmt; und sie wurde bis zum Schluss immer weiter entwickelt, sodass durch die Veränderungen durchaus auch immer wieder eine Frage schlechter als vorher funktionieren kann. Am meisten Spaß macht es und sei ausdrücklich empfohlen, selbst erdachte Fragen auszuprobieren.

Was ist die Ursache von Ebbe und Flut?
Stimmt es, dass durch einen Sturz aus dem Fenster ein Krieg ausgelöst wurde?
Welcher Naturfarbstoff färbt blau?
Welches Teil am Auto macht, dass man mit weniger Kraftanstrengung lenken kann?
Kann ein Computer Sprache erkennen?
Was ist eine Gesellschaft mit beschränkter Haftung?
Wer gründete der Sage nach Rom?
Welche tödlich giftigen Pilze gibt es?
Aus welchen Pflanzen kann man biologischen Kraftstoff gewinnen?
Welche Organisation setzt sich für die Freilassung von aus religiösen oder politischen Gründen Inhaftierten ein?
Welches Tier hat zwei Höcker?
Kann man von Medikamenten abhängig werden?
Welcher Mensch flog als erster auf den Mond?

5.3.1 Relevanzberechnung

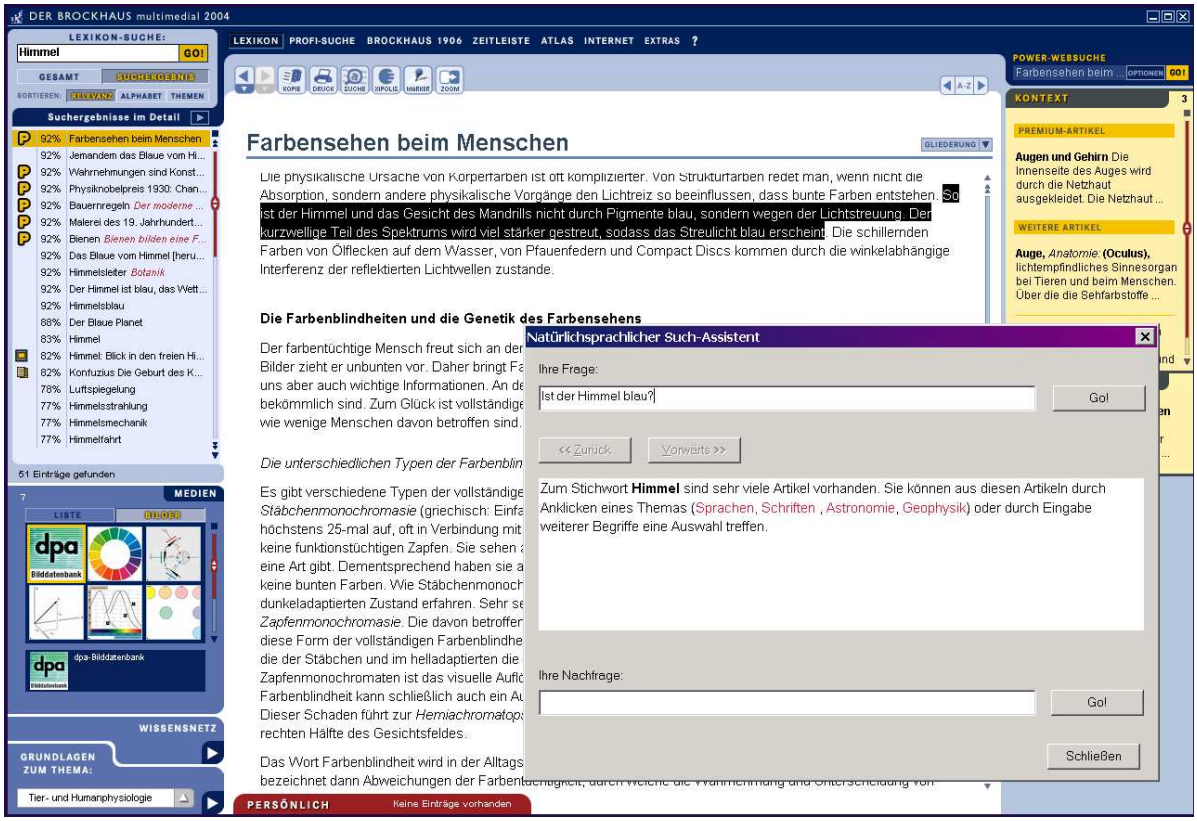
Der „klassische“ Relevanzwert soll dem Benutzer signalisieren, welche Artikel zu seiner Anfrage am besten passen. Hohe Relevanzwerte (maximal 100%) zeigen eine hohe Übereinstimmung des Artikels mit den Suchwörtern der Frage, eine niedrige Relevanz (minimal 10%) eine entsprechend geringere Übereinstimmung. Durch die Höhe des Relevanzwertes und der Abstand in der Relevanz zwischen zwei Artikel kann der Benutzer schon vor dem Lesen eines Artikel einen Eindruck über die Aussagekraft des Artikels in Bezug auf seine Fragestellung gewinnen und ist nicht nur auf die relative Positionierung in der ausgegebenen Reihenfolge angewiesen. Haben etwa fünf Artikel alle Relevanzwerte zwischen 80 und 100%, so ist davon auszugehen, dass alle Artikel Antworten auf die eingegebene Frage geben. Hat dagegen ein Artikel 90%, der nächste 75% und die anderen drei nur 20%, so sind wahrscheinlich nur die ersten beiden Artikel von Interesse.

Im Rahmen des LeWi-Projekts kommt der Relevanzberechnung insofern eine nochmals erhöhte Bedeutung zu, als die grundlegend neuen und verbesserten Retrievalmöglichkeiten nur dadurch zustandekommen, dass zunächst eine deutlich größere Ergebnismenge erzeugt wird (Oder-Verknüfung von alternativen Suchtermen; Durchführung mehrerer verschieden aufgebauter Suchen), und danach die daraus resultierende große Treffermenge sehr sorgfältig sortiert wird. Dazu wurde ein Algorithmus von der *Brockhaus Duden Neue Medien GmbH* entworfen und beschrieben, der dann von *Joanneum Research* implementiert wurde.

Grundlage für die Ermittlung der Relevanzwerte ist der von allen Texten des *Brockhaus multimedial* erstellte Index. Für den Index werden alle Texte mit dem Analysetool Mpro des *IAI* morphosyntaktisch analysiert, sodass auch eine Suche nach Komposita, flektierten Formen oder z.T. auch abgeleiteten Wörtern möglich ist. Weiterhin werden verschieden Elemente der DTD-Struktur der Texte ausgewertet sowie Zahlen und Daten berücksichtigt. Zudem werden die Positionen der Wörter in einem Artikel bestimmt, sodass bei den Suchen mit mehr als einem Suchwort die Nähe zwischen den Wörtern ermittelt werden kann.

Für die Berechnung des Relevanzwertes wird ein Algorithmus verwendet, der die Worthäufigkeit, mögliche Abweichungen der gefundenen Wörter von Grundform (z.B. Komposita oder Ableitungen) und das Vorkommen der Suchwörter in bestimmten Artikelkontexten (z.B. zählt der Treffer mehr, wenn die Textpassage "Lemma" oder "Unterstichwort" ist, als wenn er nur im Volltext liegt) berücksichtigt. Bei Suchen mit mehreren Wörtern wird auch die Entfernung bzw. Nähe der Wörter im Text zueinander und das Vorkommen in einer Phrase berücksichtigt.

Die beschriebene Berechnung der Relevanzwerte findet für jede vom Dialogmodul initiierte Suche statt. Diese Rohwerte werden für jede einzelne Suche mit einem bestimmten Faktor multipliziert. Höher bewertet werden "spezifischere" Suchen, d.h. solche, die mit vielen Wortklassen, nur für bestimmte Klassen oder Artikeltypen durchgeführt wurden oder solche, in denen die Suchwörter durch einen Nahebei-Operator verbunden wurden. Abgewertet durch die Faktoren werden dagegen die Suchen mit zerlegten Komposita und Ableitungen sowie Oder-Suchen der Suchwörter. Da die beschriebenen Parameter in vielen Kombinationen auftreten, sind die Faktoren sorgfältig justiert worden.



„Ist der Himmel blau?“ Vorbei an anderen Artikeln, die z.T. die gesuchten Begriffe direkt als Stichworttreffer enthalten, sortiert die Software den Artikel „Farbensehen beim Menschen“ nach oben, welcher die gewünschte Antwort enthält.

5.3.2 Rückfragen

Wie im vorhergehenden Abschnitt dargestellt, erhält der Benutzer als Feedback eine relevanzbewertete Ergebnisliste. Hier kann er prüfen, ob die Textauswahl seinem Informationswunsch entspricht. Zur Unterstützung der weiteren Interaktion werden dem Benutzer darüber hinaus durch Rückfragen ergänzende Informationen zur Textrecherche und zum Status der programmtechnischen Auswertung seiner Frage gegeben. Die Rückfragen sollen zum einen über das Ergebnis informieren und zum anderen Hilfen für die Fortsetzung des Dialogs geben, um dem Benutzer die Möglichkeit zu geben, durch Folgeeingaben das Ergebnis weiter zu präzisieren. Dazu gehört auch die Nachfrage nach nicht verstandenen Begriffen und zur Korrektur fehlerhafter Schreibweisen.

Das im Projekt entwickelte komplexe System der natürlichsprachlichen Rückfragen ermöglicht für die verschiedenen Ergebniszustände bei der Auswertung einer Frage unterschiedliche, problemadäquate Feedbacks. Grundlage der Rückfragenausgabe ist die so genannte »Statusermittlung« zu einer Frage. Auf der Basis der im Folgenden dargestellten Parameter wird festgestellt, in welcher Weise die Frage vom System "verstanden" und welches Ergebnis dazu ermittelt werden kann. Die Parameter dafür sind graduell abgestuft und können so situationsgerecht Antwortsituationen abbilden. Jedem der entsprechenden Statuszustände sind unterschiedliche Reaktionen zugeordnet und mit entsprechenden Rückfragentexten versehen. Darüber hinaus definieren die Parameter des Rückfragesystems die Struktur für eine zielgerichtete Interaktion mit dem Benutzer: Es wird vom System je nach Art der Benutzereingabe und der Qualität der Auswertung versucht, durch Rückfragen von den Statuszuständen mit geringerem Verständnis zu Zuständen mit höherem Verständnisgrad zu kommen.

Parameter zur Bestimmung des Rückfragetyps

A. Status morphosyntaktische Analyse:

"Can not analyze"=0
"Can analyze"=1

B. Fragetyperkennung:

Kein Fragetyp assoziiert=0
Mehr als zwei Fragetypen assoziiert=1
Zwei Fragetypen assoziiert=2
Ein Fragetyp assoziiert=3

C. Worterkennung:

unknown und nach Stringsuche in BMM-DB nicht gefunden=0
unknown (in der Satzanalyse sind ein oder mehrere 'unknown')=1
bekannt (es kommen keine 'unknown', aber mehrdeutige Begriffe vor)=2
bekannt (keine 'unknown'/mehrdeutige Begriffe in der Satzanalyse)=3

D. Ergebnismenge:

kein Ergebnis=0
ein Ergebnis=1
gering=2 (2-5 Ergebnisse)
mittel=3 (6-20 Ergebnisse)
groß=4 (ab 21 Ergebnissen)

E. Relevanz:

geringe Relevanz=1
mittlere Relevanz=2
große Relevanz=3

Die verschiedenen Kombinationsmöglichkeiten dieser Parameter definieren die Statuszustände. Zu jeder Kombination ist ein eigener Rückfragetext im System festgelegt, sodass sich für unterschiedliche Szenarien differenzierte Feedbacks ausgeben lassen. Im System sind ca. 750 solcher Zustände mit eigenen Rückfragen abgebildet.

Bespiele für Rückfragen:

- Frage: Halten Bären einen Winterschlaf?

Status: MES_A1B3C3D4E3

Rückfrage: Zu den Stichworten **Winterschlaf** und **Bär** sind sehr viele Artikel vorhanden. Sie können aus diesen Artikeln durch Anklicken eines Themas (Tiere, Pflanzen, Pilze, Astronomie) oder durch Eingabe weiterer Begriffe eine Auswahl treffen.

MES_A1B3C3D4E3_QUEST10 bedeutet, dass die Frage als ein formal korrekter Satz erkannt wurde (A1), dass der Fragentyp eindeutig ist (B3, Entscheidungsfrage), dass alle Worte erkannt wurden (C3), dass es sehr viele Ergebnisse gegeben hat (D4), dass die Relevanz hoch ist (E3) und eine eigene Antwort für die Entscheidungsfrage vorgesehen ist (QUEST10).

- Frage: Welche Staaten sind Dempkratien in Asien?

Status: MES_A1B3C1D4E3

Rückfrage: Nicht alle Begriffe Ihrer Anfrage nach einem geographischen Begriff sind bekannt (unbekannt: **Dempkratien**). Es wurden aber sehr viele Artikel zum Stichwort **Asien** gefunden. Neue Eingabe oder ein Sachgebiet für die gefundenen Artikel wählen (geografische Objekte Asiens, Neuere Geschichte (1789 - 1945), Zeitgeschichte (ab 1945))?

Rückfrage zu fehlerhaften Eingaben mit Suche nach erkannten Begriffen:

- Frage: Was ist Mocha?

Status: MES_A1B2C3D1E3

Rückfrage: Sie möchten eine Definition oder eine Sachinformation zum Stichwort **Mocha**. Dazu wurde ein Artikel gefunden. Für weitere Informationen dazu geben Sie bitte eine neue Frage ein.

Beispiel für Rückfrage zu unklarem Frageobjekt/Fragetyp:

- Frage: Wie entsteht Krebs?

Status: MES_A1B3C2D4E3

Rückfrage: Sie interessieren sich für eine funktionale Erklärung. Es wurden dazu sehr viele Artikel ausgewählt. Einige Stellen in Ihrer Fragestellung sind jedoch nicht eindeutig. Welche Bedeutung von **Krebs** meinen Sie: Medizin und Pharmazie, Pharmakologie oder Tiere?

Beispiel für Rückfrage zu mehrdeutigen Begriffen mit Fragentyp:

- Frage: Wo steht das Pantheon?

Status: MES_A1B3C3D4E3

Rückfrage: Ihre Frage zielt auf eine geographische Angabe zum Thema **Pantheon**. Dazu wurde eine große Anzahl von Artikeln ermittelt. Möchten Sie die Auswahl reduzieren? Klicken Sie dazu das gewünschte Sachgebiet an (Religionen, Architektur , Alte Geschichte (ca 3000 v.Chr. bis ca 476 n.Chr.)) oder geben Sie weitere Stichworte zu Ihrer Frage ein.

5.3.3 Fortführung des Dialogs

Das Dialogmodul hat - wie oben skizziert - die Funktion, zwischen der Fragestellung des Benutzers und dem elektronisch gespeicherten Wissensbestand zu vermitteln. Dazu wird ein spezieller Kommunikationsprozess organisiert, mit dem die Komplexität der natürlichsprachlichen Anfrage bewältigt werden kann. Dieser beinhaltet neben dem Feedback an den Benutzer (also die Wiedergabe des "Systemverständnisses" der eingegebenen Frage und der Qualität der Suchergebnisse - siehe Erläuterungen zum Rückfragensystem weiter oben) auch die Möglichkeit zur Fortsetzung des Dialogs. Der Benutzer kann dabei, wenn ihm der Umfang und Qualität der bisherigen Ergebnisse nicht ausreichend erscheinen, durch Nachfragen die Artikelauswahl durch eine Präzisierung und Erweiterung der Ausgangsfragestellung weiter spezifizieren. Die Anzahl der Nachfragen ist dafür auf vier Nachfragen beschränkt. Nach dem Ende des definierten Fragezyklus wird davon ausgegangen, dass in diesem Zyklus keine Verbesserung mehr zu erwarten ist. Der Benutzer kann dann wieder eine neue Frage eingeben, eine Möglichkeit, die ihm im Dialogmodul jederzeit neben den Nachfragen zur Verfügung steht.

Ziel des Folgedialogs ist es zusätzliche Informationen sowie eine eventuelle Fehlerkorrektur zu ermitteln. Die Gestaltung der Dialogfortführung umfasst dazu die Aufbereitung der Nachfrage und Feedback an den Benutzer sowie die Ermittlung und Darstellung neuer Suchergebnisse.

Die Möglichkeiten zur Eingabe von Nachfragen und deren Aufbereitung zu Suchanweisungen sind im Dialogmodul identisch mit jenen für Ausgangsfragen. Es sind komplette Fragen ebenso zulässig wie einzelne Worte; für die daraus ermittelten Suchbegriffe werden wiederum entsprechende Einzelsuchen (S1-S8, siehe oben) gebildet und durchgeführt.

Mit dem zugehörigen Feedback soll nun dem Benutzer ermöglicht werden, das Ergebnis zu optimieren. Er wird dazu bei unvollständigen, fehlerhaften oder nicht erkennbaren Fragestellungen über die festgestellten Mängel und den Status der Fragenauswertung informiert und bei der Ermittlung neuer Suchbegriffe aus der Nachfrage über die Qualität des neuen Ergebnisses.

Entsprechend unterscheidet das Dialogmodul folgende Feedbacksituationen:

1. Bei der Ersteingabe wurden nicht alle Parameter erkannt (logische Form, Fragentyp, verwertbare Begriffe). Wird dieser Zustand durch die Nachfrage behoben, erfolgt die im Standard-Rückfragensystem vorgesehene Antwort und Darstellung der Ergebnisse. Ist dieser Zustand durch die Nachfrage noch nicht behoben, so wird der Benutzer darüber informiert.

2. Bei der Ersteingabe wurden bereits alle Parameter erkannt (logische Form, Fragentyp, verwertbare Begriffe). Die Rückantwort und die Ergebnisdarstellung zur Nachfrage erfolgt dann in Abhängigkeit von der Art des neuen Suchergebnisses (zu dessen Bildung siehe unten), je nachdem, ob Artikel mit allen Suchbegriffen von Erst- und Folgeeingaben vorhanden sind oder ob Artikel dominieren, die nur für die Erst- oder die Nachfrage ein passendes Ergebnisses darstellen.

3. Bei Sonderfällen sind entsprechende Meldungen an den Benutzer vorgesehen, zumeist verbunden mit der Möglichkeit weiterer Eingaben. Sonderfälle liegen vor, wenn Begriffe der Nachfrage unbekannt oder mehrdeutig sind, keine neuen Suchbegriffe durch Folgeeingabe hinzugekommen sind, Schreibfehler vorliegen, Nonsenseingaben (z.B. "Das Ergebnis gefällt mir nicht."), das Ende des Fragenzyklus sowie die zusätzliche Suche mit Ober- und Unterbegriffen aus der Ontologie.

Die Ermittlung und Darstellung des Ergebnisses zu einer Nachfrage ist abhängig von der Relevanz der bisherigen Suchergebnisse und derjenigen zur Nachfrage. Die Ergebnisse der zunächst separat durchgeführten Suche mit den Suchbegriffen aus der Folgeeingabe werden mit den Ergebnissen der jeweils letzten Suche (aus Ersteingabe oder vorhergehender Nachfrage) kombiniert und zu einer gemeinsamen Ergebnisliste vereint. Dabei wird der Relevanzwert von Artikeln, die in beiden Listen vorkommen, erhöht (z.Zt. 1,2). Damit soll sichergestellt werden, dass diese offenkundig wichtigsten Artikel an der Spitze des kombinierten Ergebnisses stehen. Kommen Artikel nur in der alten Liste vor, bleibt der Relevanzwert dagegen gleich. Damit soll sichergestellt werden, dass der Ergebniskontext der Grundfrage erhalten bleibt. Bei Artikeln, die nur in der neuen Liste vorkommen, wird der Relevanzwert um einen Faktor vermindert (z.Zt. 0,8). Damit soll sichergestellt werden, dass diese Artikel nicht das Basisergebnis vorheriger Suchen im aktuellen Zyklus verfälschen.

	Frage 1	**Rückfrage**
Eingabe	Welchen Stellenwert haben die Menschenrechte in den Kulturen Asiens?	Ihre Frage zielt auf den Stellenwert eines Begriffs oder Ereignisses zu den Themen **Asien** und **Menschenrecht**. Dazu wurde eine große Anzahl von Artikeln ermittelt. Möchten Sie die Auswahl reduzieren? Klicken Sie dazu das gewünschte Sachgebiet an (Öffentliches Recht, geografische Objekte Asiens, Alte Geschichte (ca 3000 v.Chr. bis ca 476 n.Chr.)) oder geben Sie weitere Stichworte zu Ihrer Frage ein.
Nachfrage 1	[Anklicken von geografische Objekte Asiens]	Sind in den ausgewählten Artikeln zum Thema geografische Objekte Asiens die gewünschten Informationen enthalten? Sie können auch eine

		erneute Anfrage durchführen.
Nachfrage 2	Werden die Menschenrechte in den Dempkratien Asiens stärker beachtet?	Ist Ihre Fragestellung korrekt eingegeben (Korrekturvorschläge: Demokratie – keine Korrektur)?
Nachfrage 3	[Anklicken von Demokratie]	Die Suchergebnisliste enthält zusätzlich neue Artikel unter Berücksichtigung des Stichworts **Demokratie**. Sie können die Anfrage mit weiteren Eingaben fortsetzen.

Nach der Berechnung dieser Werte wird daraus eine neue integrierte Ergebnisliste gebildet und nach der Relevanz sortiert. Je nachdem, ob in dieser Gesamtliste Artikel aus beiden Suchen oder aus der alten bzw. der neuen Suche dominieren (also ob die Nachfrage das Ergebnis präzisiert hat, es unverändert gelassen oder neue Aspekte aufgedeckt hat), erfolgt eine entsprechende Information an den Benutzer. Dieser Vorgang wird (bis zur maximal möglichen Anzahl) bei einer neuen Nachfrage wiederholt, wobei die gebildete kombinierte Liste dann die Basisliste (also die "alte" Liste) darstellt.

Ein Sonderfall der Dialogfortführung ist die Einbeziehung von Ober- und Unterbegriffen zu den aktuellen Suchbegriffen. Diese soll erfolgen, wenn das Ergebnis wenig aussagekräftig ist, weil entweder zu viele oder zu wenige Ergebnisse zu einer Frage ermittelt wurden. Z.T. liegt die Antwort auf eine objektbezogene Frage direkt in dessen Oberbegriffsartikel (z.B. "Ist Argon reaktionsträge?" -> Antwort im Artikel "Edelgase" enthalten).

Die Verwendung von Oberbegriffen als zusätzliche Suchbegriffe – sofern sie in der Ontologie gefunden werden - ist durch einen Parameter in der Initialisierungsdatei geregelt. Wenn dieser Parameter positiv besetzt ist, erfolgt unter bestimmten Bedingungen (kein Artikel mit 100% Relevanz und Suchwörter im BMM vorhanden) eine zusätzliche ODER-Suche mit Oberbegriffen. In der theoretischen Konzeption des Fragemoduls wurde kurz vor Projektende auch noch die in der anschließenden Tabelle dargestellte zusätzliche Suchlogik angedacht, die bei speziellen Ergebniskonstellationen eine noch feinere Suchlogik auslöst; dies wurde jedoch im Abschluss-Prototypen nur noch teilweise realisiert.

Artikel zahl			
0	1. übergeordnete Begriffe in ODER-Suche ohne Rückfrage an Benutzer wenn Suche ohne Ergebnis 2. untergeordnete Begriffe in ODER-Suche ohne Rückfrage an Benutzer		
	1. Artikel mit Relevanz ≤30%	1. Artikel mit Relevanz ≥30% und ≤70%	1. Artikel mit Relevanz ≥70%
1	1. übergeordnete Begriffe in ODER-Suche nach Rückfrage an Benutzer wenn Suche ohne Ergebnis 2. untergeordnete Begriffe in ODER-Suche ohne Rückfrage an Benutzer	---	1. Rückfrage, ob jetzt auch nach dem übergeordneten Begriff xyz gesucht werden soll; nach Bestätigung durch Benutzer ODER-Suche wenn Suche ohne Ergebnis 2. untergeordnete Begriffe in ODER-Suche nach Rückfrage an Benutzer
2-5	1. übergeordnete Begriffe in ODER-Suche nach Rückfrage an Benutzer wenn Suche ohne Ergebnis	---	1. Rückfrage, ob jetzt auch nach dem übergeordneten Begriff xyz gesucht werden soll; nach Bestätigung durch Benutzer ODER-Suche

	2. untergeordnete Begriffe in ODER-Suche ohne Rückfrage an Benutzer		wenn Suche ohne Ergebnis 2. untergeordnete Begriffe in ODER-Suche nach Rückfrage an Benutzer
6-20	---	---	---
21-40	1. übergeordnete Begriffe in UND-Suche und Multiplikation der entsprechenden Artikel mit dem Wert 1,1 wenn Suche ohne Ergebnis 2. untergeordnete Begriffe in UND-Suche und Höherstufung der entsprechenden Artikel	---	---
≥ 41	1. übergeordnete Begriffe in UND-Suche und Multiplikation der entsprechenden Artikel mit dem Wert 1,1 wenn Suche ohne Ergebnis 2. untergeordnete Begriffe in UND-Suche und Höherstufung der entsprechenden Artikel	---	1. übergeordnete Begriffe in UND-Suche und Multiplikation der entsprechenden Artikel mit dem Wert 1,1

6 Visualisierung von Dokumenten aus Volltextdatenbanken

Neben der Suche von Dokumenten war die Visualisierung von Wissensfeldern und -zusammenhängen ein wichtiges Ziel dieses Projektes. Dafür wurde ein dreidimensionaler Wissensraum entworfen und programmiert, in dem für ein aktuell aufgerufenes Dokument zweierlei Arten von Bezügen dargestellt werden. Zum einen werden dem Nutzer inhaltlich benachbarte Dokumente gezeigt und zum anderen eine sachsystematische Einordnung des Wissens präsentiert.

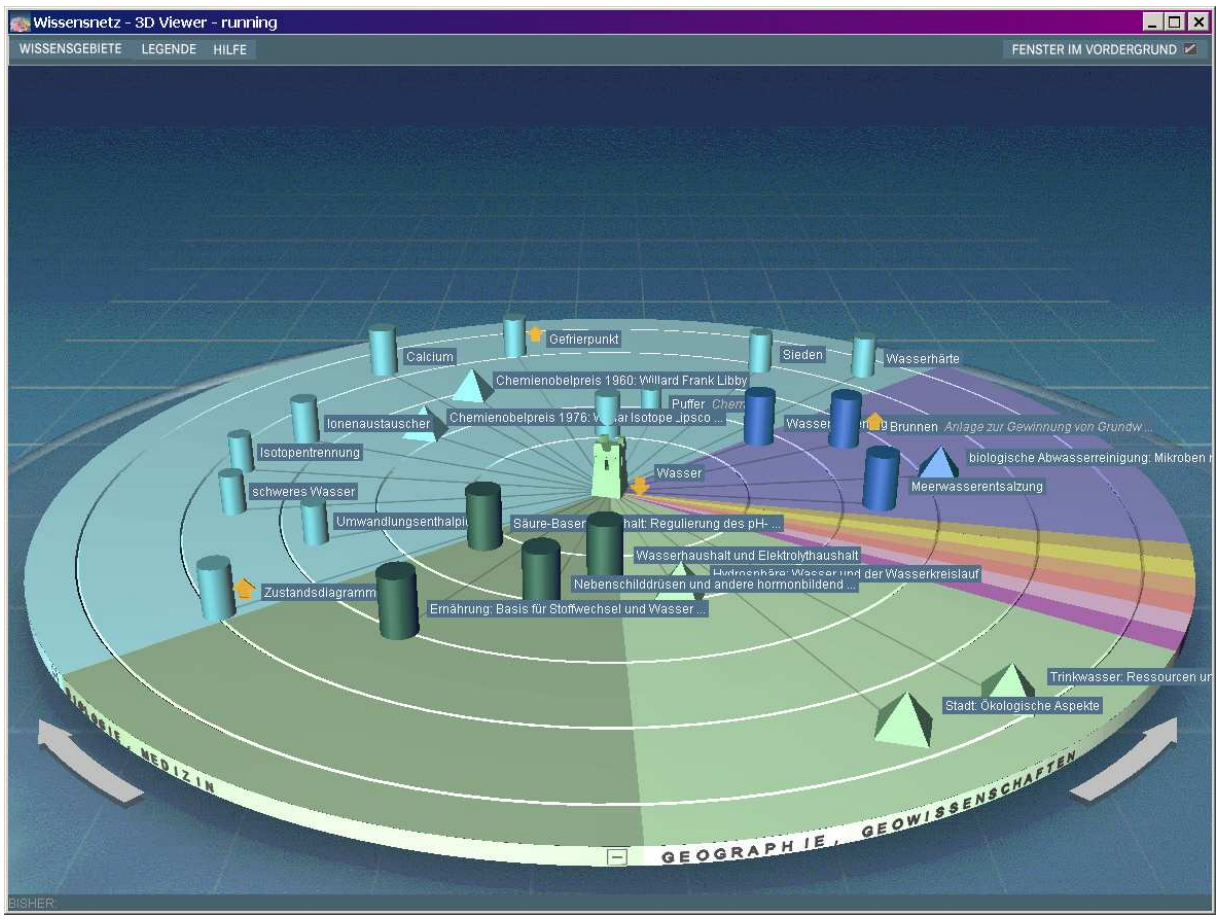
6.1 Inhaltliche Konzeption

6.1.1 Herstellung assoziativer Zusammenhänge

Um für alle Dokumente einen eher assoziativen Zusammenhang visualisieren zu können, wurde zwischen allen Dokumenten des *Brockhaus multimedial* ein Relevanzwert ermittelt. Der dafür verwendete, von *Joanneum Research* entwickelte Algorithmus berücksichtigt gemeinsam zwischen zwei Artikeln verwendete Wörter, redaktionelle Verweise, die Grob- und Feinklassen der jeweiligen Artikel sowie gemeinsam genutzte Medien. Für jeden Artikel werden die 25 höchst relevanten Artikel berechnet und in einer Tabelle abgelegt. In dieser Tabelle ist auch kodiert, welchem Artikeltyp er zugehört (Sachartikel, biografischer oder geografischer Artikel sowie so genannte Premium-Artikel für umfangreiche, nichtlexikalische Texte). Außerdem ist für jeden Artikel eine Wertigkeit berechnet, die sich aus der Länge der Artikel und der Zahl der zugeordneten Medien ergibt. Eine Wertigkeit von 5 entspricht einem sehr langen Artikel und/oder einem mit vielen Medien, die Wertigkeit 1 einem kurzen medienlosen Artikel. Zudem ist jeder Artikel einer der großen Themenklassen, wie Naturwissenschaften oder Kunst zugeordnet. Diese 10 Themenklassen (auch "Superklassen" genannt) ergeben sich aus der Zusammenfassung von mehreren Grobklassen.

All diese Informationen werden im Wissensraum auf einer waagerecht im Raum angeordneten Scheibe dargestellt: Das aktuelle Dokument ist im Zentrum der Scheibe platziert und die zugeordneten Artikel sind von innen nach außen mit abnehmender Relevanz um diesen Artikel angeordnet. Jedes Dokument ist durch ein Objekt dargestellt, wobei jeder Artikeltyp durch einen eigenen Objekttyp (Tonne für Sachartikel, Männchen für Personenartikel, Turm für geografische Artikel und Pyramide für so genannte Premium-Artikel) und die Wertigkeit der Artikel durch die

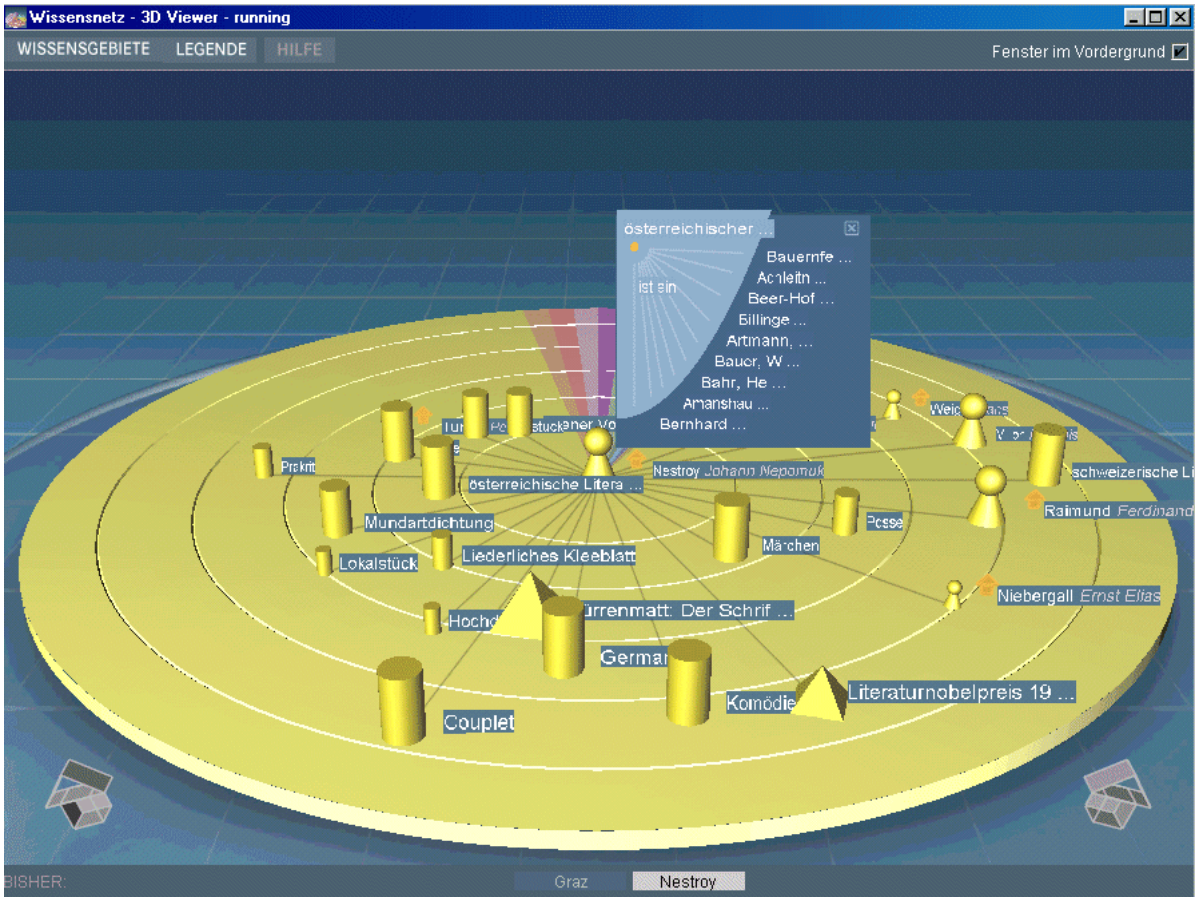
Größe der Objekte symbolisiert wird. Die Artikel werden je nach Themenklasse in einem farbigen Segment eingeordnet.



Der 3D-Viewer mit dem Wissensnetz des Artikels „Wasser“

6.1.2 Sachsystematische Einordnung von Dokument-Lemmata

Liegen sachsystematische Informationen vor, wird das über Pfeile an den Lemmata der Artikel signalisiert. Nach einem Klick des Benutzers erscheint eine senkrechte Fläche, in der ein übergeordneter Begriff gezeigt wird, sowie dazu untergeordnete Begriffe, die also mit dem Ausgangsbegriff sachsystematisch auf einer Ebene stehen. Diese sachsystematischen Informationen wurde für biografische und geografische Einträge über eine mit computerlinguistischen Methoden durchgeführte Faktenextraktion aus lexikalischen Artikeln des Verlages gewonnen, die anschließend in die Ontologie importiert wurden. Die auf den senkrechten Flächen auch angezeigten Ober- und Unterbegriffe für Sacheinträge wurden dagegen aus den Konzeptdefinitionen des großen *Duden-Wörterbuchs* gewonnen.



Das Wissensnetz von Johann Nestroy mit geöffnetem Ontologie-„Segel“. Das Oberbegriff-Segel benennt in seiner Titelzeile die Art des zugehörigen Objekts (hier: Österreichischer Schriftsteller), die Relation „nach unten“ (hier: „ist ein“) und weitere Relationsziele, d.h. in diesem Fall weitere österreichische Schriftsteller, die größtenteils nicht Teil des assoziativen Wissensnetzes auf der Scheibe sind. Damit können Objekte durch Verweis auf Gleichartiges erklärt werden, z.B. „Eine Limone ist eine Zitrusfrucht, wie auch die Orange, die Zitrone und die Pampelmuse eine sind“. – Ein Pfeil nach unten würde Unterbegriffs-Segel öffnen, z.B. bei Berlin die Relation „hat Stadtteile“ mit den Relationszielen „Zehlendorf“, „Pankow“ usw. und dadurch das Objekt durch die Darstellung seiner Teile weiter erläutern.

6.1.3 Interaktionskonzept

Der Zugang zu den Artikeln im *Brockhaus multimedial* soll so schnell und einfach wie möglich sein, z.B. durch einen einfachen Mausklick auf ein Objekt, das einen Artikel repräsentiert. Daher wurde die Interaktion des Benutzers mit dem 3D-Wissensraum so gestaltet, dass er keinerlei ungewohnte Aktionen ausführen muss, die unter Umständen eine erfolgreiche Benutzung der Software behindert hätten. Sämtliche Aktionen innerhalb des 3D-Viewers, die zur Bedienung notwendig sind, konnten mit normalen Interaktionen wie linker Mausklick und linker Mausklick und Ziehen der Maus mit gedrückter linker Maustaste realisiert werden.

Der Nutzer kann Scheibe drehen, um weiter hinten, zunächst nur undeutlich zu sehende Objekte, besser sichtbar zu machen (Prinzip der Informations-Fokussierung/Defokussierung). Außerdem kann er einzelne Segmente mit den jeweiligen Artikeln ausblenden, um so sich noch besser auf die Verbindungen in einem ihn interessierenden Segment konzentrieren zu können.

In einer oberen Leiste kann der Benutzer unter einer Schaltfläche die Wissensgebiete und den ihnen zugeordneten Farben erkennen und einzelne Wissensgebiete abschalten, d.h. das entsprechende Kreissegment ausblenden. In einer Legende werden ihm die Objektsymbole und die Objektgrößen erläutert. Daneben steht eine Hilfe-Funktion, die in die Hilfe des Brockhaus multimedial integriert

ist, zur Verfügung. In einer unteren Leiste ist eine History-Funktion implementiert. Für jedes neu aufgebaute Wissensnetz erscheint ein Button mit dem jeweiligen zentralen Artikel. Diese Buttons sind anklickbar, sodass der Benutzer direkt eines der Wissensnetze aufrufen kann, die er zuvor schon besucht hat.

6.2 Technische Umsetzung

Der dreidimensionale Wissensraum wurde von *interActive Systems* mit Java 3D programmiert. Die angezeigten Daten werden in einer Tabelle abgespeichert, die über eine XML-Schnittstelle zwischen dem Wissensraum-Modul und der Datenbank des *Brockhaus multimedial* abgefragt wird.

6.2.1 Visualisierungsengine

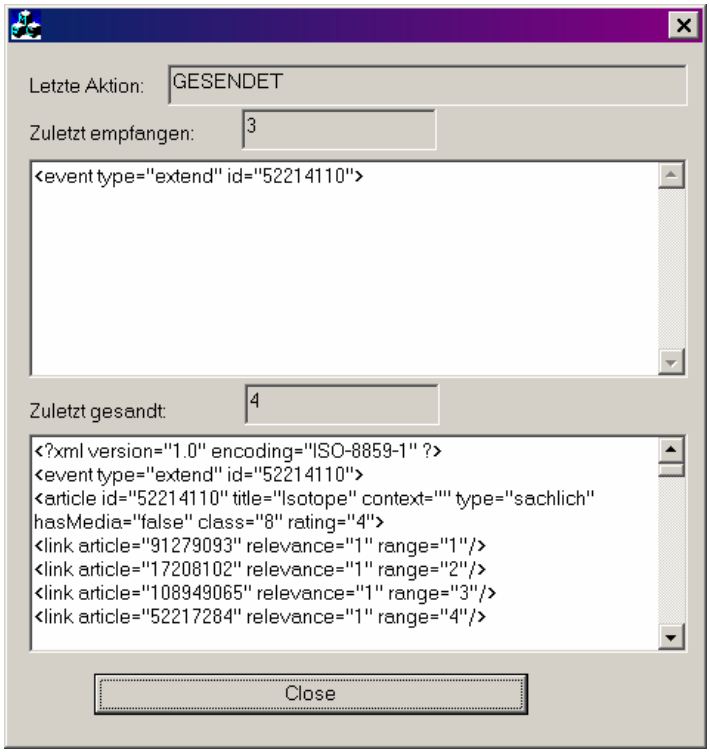
Der Eindruck einer interaktiven Echtzeit 3D-Anwendung wird durch zyklisches Berechnen von Veränderungen und Darstellen neuer Koordinaten in einem Bildschirmbereich erzeugt. Dieser Mechanismus wird von Renderengines verwaltet. Nach einem Vergleich verschiedener möglicher 3D-Visualisierungsengines (Java 3D, Shockwave 3D, Cult 3D) wurde der dreidimensionale Wissensraum in Java 1.4 mit Java 3D-Unterstützung programmiert. Wichtige Entscheidungskriterien waren die Plattformunabhängigkeit von Java 3D und Lizenzgebühren, die z.B. bei Cult 3D anfallen wären. Zusätzlich zum dreidimensionalen Wissensraum wurden zweidimensionale Navigations- und Informationskomponenten wie z.B. Menüs, Legende und eine History in Java 1.4 programmiert.

Der 3D-Wissensraum wurde technisch so konzipiert, dass es mit relativ geringen Anpassungen für Offline- und auch Online-Produkte verwendet werden kann. Die Technologie Java 1.4 mit 3D-Unterstützung ist eine webfähige Technologie. Für Online- und Offline-Version müssen jedoch unterschiedliche Schnittstellen realisiert werden.

6.2.2 XML-Schnittstelle

Die im Wissensraum anzuzeigenden Daten sowie notwendige Befehle und Anfragen werden über eine XML-Schnittstelle zwischen dem Wissensraum-Modul und dem *Brockhaus multimedial* (einer C++-Anwendung) ausgetauscht. Hierbei wird eine Schnittstellenbibliothek (LeWiExchange.dll) verwendet, die plattformunabhängig implementiert und somit auch für andere Betriebssysteme kompilierbar ist.

Die Datenschnittstelle zwischen dem so genannten LeWi-Viewer und der Runtime-Application BMM.exe erfolgt im XML-Format unter Einhaltung einer spezifizierten Dokumenttyp-Definition (DTD). Diese Definition gibt Auskunft über die zulässigen Übergabewerte und deren Parametrisierung. Es wird genau festgelegt, in welchem Format Daten z.B. über die Ontologie-Informationen oder den computergenerierten Kontext von BMM-Artikeln übermittelt werden. Im Laufe des Projektes wurde die DTD in einem iterativen Prozess mehrfach von *iAS* und *Joanneum Research* an den jeweiligen Projektentwicklungsstand und die daraus resultierenden Spezifikationen angepasst.



Debug-Ansicht des XML-Dialogs zwischen der Java-Wissensraum-Anwendung und der C++-Lexikonsoftware

6.2.3 Umsetzung der Visualisierung

Für die Umsetzung der Visualisierung wurden zunächst 3D-Entwürfe der Scheibe, anderer 3D-Objekte und der Perspektive angefertigt. Diese Entwürfe wurden zunächst mithilfe von 2D-Grafikprogrammen und einer 3D-Software (Softimage) ausgeführt. Anschließend wurde die dynamischen 3D-Objekte (Scheibe, Links, 3D-Buttons, senkrechte Fläche für sachsystematische Informationen, Text) in Java generiert. Es folgte eine Konvertierung und der Import der 3D-Modelle (Turm, Pyramide etc.) aus Softimage in Java und die Implementierung des dynamischen Mesh (Wissensnetz) sowie der dynamischen 2D- und 3D-Texte. Darin integriert wurden die Effekte für Perspektive, Hintergrund und Licht.

6.2.4 Hard- und Software Voraussetzungen

Für die Offline-Version des 3D-Wissensraums wurden folgende Mindestanforderungen definiert:

- Multimedia-PC der Pentium-IV-Klasse (ab 1 GHz) oder höher
- Grafikkarte: 32 MB
- Betriebssystem: Windows 98/ME, Windows NT 2000/XP; mindestens 64 MB RAM, je nach Betriebssystem.
- Farbtiefe: 65000 Farben oder mehr
- Pixelgröße des Applet-Fensters: Die Anzeige des Wissensraums ist auf Vollbildschirm zu vergrößern. Außerdem ist das Fenster skalierbar. Die erforderliche Minimalgröße des Fensters ist 800 x 600 Pixel, normalerweise1024 x 768 Pixel.
- Schnittstellensprache: XML

7 Usabilitytest und Fokusgruppenbefragung an Schulen und Universitäten

Die Umsetzungen und Visualisierungen zur 3D-Wissensnavigation wurden im Projektverlauf von zwei Usabilitytests begleitet, die in der Form von Fokusgruppenbefragungen durchgeführt wurden.

In der Zusammenarbeit mit dem Pädagogischen Medienzentrum Mannheim und den Schülern des Lessing-Gymnasiums Mannheim, sowie mit den Studierenden und Lehrkräften der Universität Heidelberg führte die *Media Supervision Software Consulting GmbH (MSSC)* diese Befragungen durch.

Die Befragungsrunden fanden im Dez. 2002 und im Dez. 2003 / Jan. 2004 statt mit insgesamt 39 Schülern des Mannheimer Gymnasiums und 56 Studierenden sowie 7 Lehrkräften der Universität Heidelberg.

7.1 Die Konzeption der Fokusgruppenbefragungen

Die thematische wie inhaltliche Herausforderung bestand darin, nicht nur die Konzeption einer neuen Oberflächenentwicklung zu verifizieren, sondern die Besonderheiten einer neuartigen dreidimensionalen Navigation zu berücksichtigen. Es ging um nicht viel weniger als um die Frage: Wie bekommt man in einer Benutzerbefragung heraus, ob und wie ein „durchschnittlicher“ Benutzer mit einer dreidimensionalen Navigation im Gegensatz zu einer herkömmlichen zweidimensionalen Darstellung „zurechtkommt“. So sollten die Ergebnisse der ersten Befragungsrunde die Weichenstellung für die weitere Entwicklung liefern, die zweite Befragungsrunde gleichwohl die Verifikation des abschließenden Entwicklungsmodells auf Benutzbarkeit offen legen.

Die Konzeption der Befragungen konzentrierte sich im Wesentlichen auf vier grundlegende Aspekte bzw. Bereiche:

1. Kognitiver Bereich: Verstehen von funktionalen Elementen
2. Emotionaler Bereich: persönlicher Bezug zur grafischen Oberfläche
3. Handlungsbereich: Funktionalität und virtuelles Raumverständnis
4. Erwartungsbereich: Qualität und Quantität des zu Erfahrenden

7.2 Das Ergebnis der Fokusgruppenbefragungen

Die Verifikation auf Benutzbarkeit des Entwicklungsmodells fand auf der Grundlage der dreidimensionalen Modellierung mit einer Darstellung einer zweidimensionalen Scheibe in der zweiten Befragungsrunde statt (vgl. nebenstehende Grafik).

Die Ergebnisse lassen sich wie folgt zusammenfassen: Der Erstzugang war durch Skepsis geprägt, jedoch gewährleistete die interessante Oberfläche eine positive Erwartungshaltung auf neue Erkenntnis-

möglichkeiten. Die Metaphorik wurde als eindeutig eingestuft und erkannt. Signifikante Ergebnisse gab es bei der Zuordnung von Farben zu den einzelnen Wissensgebieten. Hier gab es konkrete Änderungsvorschläge.

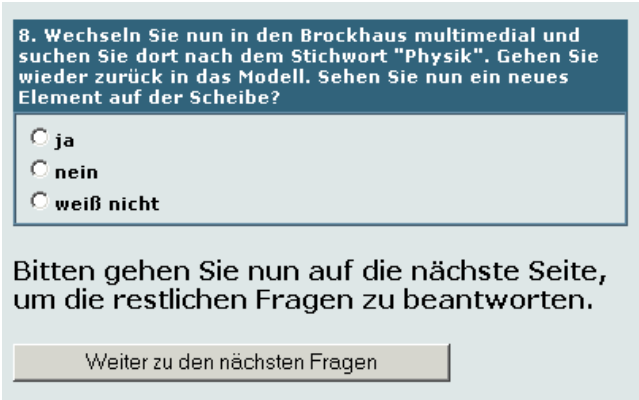
Die Drehfunktion der Scheibe war erkennbar und wurde positiv aufgenommen, ein Informationsmehrwert war dadurch allein nicht ersichtlich. Gut erkannt wurde eine klare Abfolge der Funktion: Scheibe drehen –> Symbol anklicken –> neue Information. Überwiegend wurde der Zusammenhang zwischen Datenbasis (Brockhaus multimedial) und Wissensnetz (Scheibe) erkannt und durch die dargestellten Elemente und farbigen Segmente richtig zugeordnet. Zusatzinformationen („Info-Segel“) wurden als hilfreich empfunden, aber ebenso als überarbeitungsbedürftig.

Überwiegend kann man das kommunikative Verhalten in der Mensch-Maschine-Situation als zielorientiert ansehen. Weniger wird der Proband (User) durch die „unechte“ räumliche Darstellung in den Bannkreis der Virtualität gezogen, um dort zu versinken. Die Trennung zwischen Mensch und Maschine (Computer) bleibt erhalten, im Gegensatz zu manchen Anwendungssituationen bei Computerspielen. Die Anzahl der Informationsmetaphern (Figuren) wurde als ausreichend erfahren. Mehr wäre hier zu viel.

Die Ergebnisse aus den Fokusgruppenbefragungen vermitteln den Gesamteindruck, dass die Entwicklungen zur mehrdimensionalen Wissensnavigation auf einem guten Weg sind und in diesem Sinne weiter vorangetrieben werden sollten.

7.3 Eingesetzte Technik zur Durchführung der Fokusgruppenbefragung

Die technische Durchführung der Befragung beruhte auf dem Konzept einer Online-Befragung. Im Mix mit einerseits präsenten Befragungsmoderatoren und andererseits einem Online-Befragungsbogen zur Erhebung der Daten konnten die Befragungen vollständig papierlos am Bildschirm erfolgen. Die Vorteile dieser Befragungstechnik liegen im bruchlosen Ablauf selbst: Der Proband empfindet in der Fensterumschaltung zwischen Befragungsgegenstand und elektronischem Fragebogen keine Mediendiskrepanz bzw. Unterbrechung, was wiederum dem Befragungsfluss zugute kommt.



Die durch die Online-Befragung gesammelten Daten wurden auf dem Serversystem sogleich in einer Datenbank erfasst und zeitnah ausgewertet. Die eingesetzte Plattform bot das Open-Source-System Zope, welches als Web-Applikationsserversystem diente.

8 Ausblick / wirtschaftliche Verwendung

Durch den modularen Gesamtaufbau der entwickelten Software Projektergebnisse ergibt sich eine Vielzahl von Verwertungsmöglichkeiten bei den am Projekt beteiligten Firmen, die im Folgenden aufgelistet sind. Dabei stammt der Text jeweils direkt von der zugehörigen Institution.

8.1 Bibliographisches Institut & F.A. Brockhaus AG / Brockhaus Duden Neue Medien GmbH

Während des Berichtszeitraums fasste der Verlag *Bibliographisches Institut & F.A. Brockhaus AG* den Entschluss, zum 200jährigen Brockhaus-Jubiläum im Jahre 2005 eine neue Ausgabe (21.

Auflage) der Brockhaus-Enzyklopädie herauszugeben. Diese Ausgabe wird zum ersten Mal von vorneherein nicht als reines Buchprodukt konzipiert (und allenfalls im Nachhinein auf den Bildschirm transferiert), sondern von vorneherein als für mehrere Medien konzipiertes Werk. Infolge der beeindruckenden Zwischenergebnisse des LeWi-Projekts beauftragte der Verlagsvorstand die *Brockhaus Duden Neue Medien GmbH*, die beiden Module (d.h. das Projektergebnis) für die digitale Offline-Version der neuen Enzyklopädie zur Marktreife zu bringen. Die Forschungsergebnisse werden also ihren Platz an denkbar prominenter Stelle finden.

Die Verfügbarmachung der Ergebnisse und Technologien für Dritte, insbesondere der Wissenschaft, unter der Randbedingung der Wahrung der wirtschaftlichen Interessen des Verlages (was eine direkte Veröffentlichung von Programmsourcen und Inhalte-Daten ausschließt), war im Berichtszeitraum Gegenstand von Überlegungen und Maßnahmen. Es wurden daher die wichtigen Bestandteile des Software-Systems als einbindbare, über Standardschnittstellen (http, XML, DLL/shared library) kommunizierende Softwaremodule implementiert. Diese können Dritten (Wirtschafts- und Forschungseinrichtungen) zusammen mit einer API-Dokumentation verfügbar und damit benutzbar gemacht werden, wobei das „intellektuelle Eigentum“ geschützt bleibt. Insbesondere wurde für die Kernkomponente „Ontologie“ eine Laufzeit-Schnittstelle („run time query“) geschaffen. Durch diese können auch die Nutzdaten (Sprachwissen) in computerlinguistischen Systemen Dritter verwendet werden, ohne dass ihnen die Gesamtquelldatenbank zugänglich gemacht werden muss; vielmehr werden jeweils nur die benötigten Dinge zur Laufzeit abgefragt („Liefere ein Synonym zu ‚Haus’!“). Der Verlag *Bibliographisches Institut & F.A. Brockhaus AG* hat dadurch ein aktives (wirtschaftliches) Interesse, das Modul an Dritte zu lizenzieren.

Die Umsetzung der erreichten Ergebnisse in ein kommerzielles Produkt wurde mithin eingeleitet; die Verhandlungen mit den Entwicklungspartnern darüber laufen z. Zt. An vielen Stellen wurden jedoch auch neue Grenzen gesichtet, deren Überwindung wiederum eigene Forschungsvorhaben erfordern würden. Ziel des Verlages ist zunächst die Konsolidierung des Erreichten. Die Option, nach Projektende in Nachfolgeprojekten den Erfolg weiter auszubauen, wird nicht ausgeschlossen, kurzfristig aber nicht aktiv forciert.

8.2 Krieger, Zander & Partner GmbH

Die *Krieger, Zander & Partner GmbH* kann die Erfahrungen und Ergebnisse des LeWi-Projekts sowohl für den Geschäftsbereich der elektronischen Literaturprodukte als auch für Weiterentwicklungen in den Bereichen »E-Learning« und »Knowledge-Management« nutzen. Im Hinblick auf die Recherchetechniken unserer Offline-Produkte für literarische Verlage werden wir auf die im LeWi-Projekt entwickelten, hoch differenzierten Datenbankabfrage-Methodiken zur Textermittlung zurückgreifen, um damit unsere eigenen Abfrage- und Aufbereitungsstrategien zu verfeinern. Darüber hinaus können wir hier – wie bei allen anderen elektronischen Wissenstiteln – die Analyse- und Systematisierungstechniken zur Steuerung des Benutzerverhaltens für die Optimierung des Interface und der Benutzerkommunikation einsetzen.

Für unsere E-Learning-Produkte ermöglichen verschiedene Analyseverfahren des LeWi-Projekts zur Informationsvermittlung eine differenzierte Aufschlüsselung elektronisch gesteuerter Lernprozesse. Darüber hinaus können wir die Methodik der Konzeptentwicklung im LeWi-Projekt, insbesondere die systematische Umsetzung in testfähige Schritte, zur Optimierung für unsere eigene E-Learning-Produktentwicklung nutzen. Dies gilt auch für Knowledge-Management-Projekte, die wir auf der Basis der LeWi-Verfahren zur Erfassung und Operationalisierung von Informationsbedürfnissen verstärkt vorantreiben wollen.

8.3 interActive Systems GmbH (iAS)

Die iAS strebt eine verstärkte Integration von 3D-Wissensräumen jeglicher Art in verschiedene Softwareprodukte im Bereich Knowledge-Management an. Dabei können einzelne Java-Module und

Komponenten aus der Entwicklung des LeWi 3D-Wissensraumes integriert bzw. modifiziert und adaptiert werden. iAS sieht die Entwicklung von Tools zur Visualisierung von Wissen als eine der vordringlichen zukünftigen Aufgaben im Bereich Knowledge-Management und Retrieval an und plant weitere auf den Ergebnissen von LeWi aufbauende Entwicklungsarbeiten durchzuführen, z.B. im Rahmen des vom BMBF geförderten Projektes IKAR/OS.

Da sich die iAS auf Kunden aus dem Bereich Medizin und Pharmazie spezialisiert hat, sollen die 3D-Wissensräume in für diesen Bereich speziell entwickelten Knowledge-Management-Tools z.B. in Kundennetzwerken (Intranet oder Internet) eingesetzt werden. Aber auch ein Einsatz im Bereich von Datenbanken von Bildungs- oder Forschungsinstituten sowie auf Lernplattformen ist möglich.

8.4 Joanneum Research (JR)

Die Aufgabe von *Joanneum Research* in LeWi war die Integration der verschiedenen Software-Komponenten. Dabei war insbesondere gefordert, eine effiziente Schnittstelle zwischen der von *Joanneum Research* für den Brockhaus multimedial (BMM) entwickelten Datenbank, die die Dokumente und Metadaten des BMM verwaltet, und den Analysekomponenten einerseits, und den Anzeigemodulen andererseits zu verwirklichen. Die Erfahrungen, die dabei gewonnen wurden, werden in zukünftige Projekte einfließen. Im Zuge der Arbeiten wurde auch eine wesentliche Verbesserung der Retrieval-Engine der Datenbank erreicht, die zukünftigen Projekten zugute kommen wird. So konnten durch Verwendung von genauen Wortpositionen im Index Verbesserungen bei der Trefferbewertung und neue Möglichkeiten bei der Aufbereitung der Daten (z.B. Integration von synonymen Schreibweisen für Zahlen wie "13" und "dreizehn") erzielt werden.

Joanneum Research positioniert sich als Anbieter von Softwarelösungen, die „Mehrwert für digitale Inhalte“ bieten. In diesem Zusammenhang ist es besonders wichtig das Potenzial der heute verfügbaren Analyse- und Kombinationsverfahren zu kennen und es durch sorgfältiges Finetuning der vielfältigen Optionen so weit wie möglich auszuschöpfen. In diesem Prozess werden auch die Grenzen des Möglichen deutlich, was im Hinblick auf Verwertungen und zukünftige Kooperationen ebenso bedeutsame Information darstellt.

8.5 Intelligent Views (i-views)

Wichtige Alleinstellungsmerkmale des Produktes K-Infinity von *intelligent views* sind eine höhere Intelligenz in der Informationsverteilung und Suche sowie die konsequente Benutzerfreundlichkeit des Systems. Daher ist die natürlichsprachliche Suche als Feature von K-Infinity und den damit aufgebauten Wissensmanagement-Lösungen für *intelligent views* sehr interessant.

Eine Schwierigkeit, mit der nahezu alle Unternehmen konfrontiert sind, die Wissensmanagement-Projekte aufsetzen, ist die Motivation der Mitarbeiter ihr Wissen zu teilen und z.B. in eine Wissensdatenbank einzugeben. Viele Unternehmen setzen darauf, dass die Mitarbeiter den Nutzen des Wissensmanagements direkt erfahren - dadurch, dass sie ihre eigenen Fragen beantwortet und die eigenen Probleme dort gelöst bekommen. Hier stehen das Wissensmanagement-Projekt dann aber oft vor einem "Henne-Ei-Problem": Solange noch keine Inhalte eingestellt sind, ist die Wissensdatenbank nicht nützlich und kann niemanden überzeugen, seinerseits sein Wissen dort einzustellen. Wir sehen daher einen hohen Bedarf bei unseren Kunden an externen, z.B. enzyklopädischen Inhalten für Wissensmanagement-Lösungen. Das LeWi-Projekt gibt eine Maximalstrukturierung von Inhalten vor, die es möglich macht, zu vielen Themen eines Unternehmenswissensnetzes passende Strukturen in der BIFAB-Ontologie/Thesaurus zu identifizieren. So können Inhalte gezielt abgerufen werden, ohne dass man für jedes Unternehmenswissensnetz neu klassifizieren muss.

intelligent views wird weiterhin prüfen, ob die Techniken der natürlichsprachlichen Suche und Interpretation von Fragen, beispielsweise die Fragentypologie, übertragbar sind auf Probleme des Textmining. Vor allem die Extraktion von Fakten aus natürlichsprachlichen Texten verspricht großen Nutzen für Wissensmanagement-Anwendungen, erfordert aber noch erheblichen Forschungs-

und Entwicklungsaufwand. intelligent views möchte zur Auslotung und Nutzung dieses Potenzials auch über die Projektlaufzeit hinaus mit den Projektpartnern zusammenarbeiten.

Schließlich wird intelligent views im Markt testen, ob die in LeWi entwickelte 3D-Visualisierung einen Zusatznutzen für Kunden und Interessenten darstellt.

8.6 Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. (IAI)

Auswahl an Verwertungen der Projektergebnisse LeWi vom Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes

In folgenden Bereichen sind im Rahmen des LeWi-Projekts für das IAI verwertbare Ergebnisse erarbeitet worden:

1. Klassifikationen und Thesauri: Die im Rahmen von LeWi entwickelten Arbeitstechniken für die Erstellung von Thesauri werden in künftigen Projekten des IAI verwendet. Die Erfahrungen, die bei der Erstellung des BMM-Thesaurus gemacht wurden, insbesondere bei der schwierigen Frage der gleichmäßigen Abdeckung von Themenbereichen, fließen in geplante Thesaurus-Vorhaben ein.

2. Automatische Indexierung: Die Weiterentwicklung der BMM-Klassifikation hat grundsätzliche Probleme bei der Bildung von Klassifikationen aufgezeigt, die in neuen Projekten künftig vermieden werden können. Die im Projekt entwickelten Methoden zur automatischen Indexierung bzw. Klassifizierung von Texten jeglicher Art nach einer vorgegebenen Klassifikation werden in Zukunft in vielen anderen Kontexten, z.B. bei der Verschlagwortung von Texten in Verlagen, Bibliotheken etc. eingesetzt. Gleichzeitig finden die Ergebnisse weitere Verwendung in einem neuen DFG-Projektantrag des IAI, bei dem es um die Verbesserung automatischer Indexierung durch das systematische Auflösen semantischer Mehrdeutigkeiten geht.

3. Morphosyntaktische Analyse von Texten allgemein: Durch das LeWi-Projekt konnten im Bereich der morphosyntaktischen Analyse von Texten Fortschritte in Bezug auf Präzision und Genauigkeit der Analyse erreicht werden. Insbesondere bei der Auflösung von Zusammensetzungen sowie der Erkennung von abgetrennten Verbteilen, Eigennamen und festen Adjektiv-Substantiv-Verbindungen wurden Verbesserungen erzielt. Diese Verbesserungen werden künftige Analysekomponenten des IAI (vor allem in Rechtschreib- und Grammatikkontrolle) optimieren.

4. Morphosyntaktische Analyse von Fragesätzen: Die im LeWi-Projekt speziell entwickelte Fragenanalyse ist richtungsweisend für zukünftige Projekte. Die gewonnenen speziellen Erkenntnisse über die Analyse von Fragesätzen des Deutschen erleichtern den Aufbau solcher Systeme in Zukunft außerordentlich.

5. Fragentypologie und Fragenklassifikation: Die erarbeitete Fragentypologie sowie die speziell entwickelte Fragengrammatik werden in zukünftige Dialogsysteme einfließen. Der Mechanismus der Fragentyperkennung kann als Best-Practice-Beispiel für derartige Vorhaben herangezogen werden.

8.7 MediaSupervision Software Consulting GmbH (MSSC)

Für die MSSC ergeben sich aus den im Projekt aufgeworfenen Fragestellungen schwerpunktmäßig zwei weiterführende Verwertungsbereiche: erstens in der Konzeption und Spezialisierung von Fokusgruppenbefragungen mit mehrdimensionalen Navigationsräumen; zweitens in der Verfeinerung von Online-Befragungsverfahren.

Usabilitytesting mit Fokusgruppenbefragungen konzentrierte sich in der Vergangenheit weitgehend auf zweidimensionale Navigationsformen. Mit welchen Befragungskonzeptionen aber lassen sich dreidimensionale Navigationsräume angemessen validieren und verifizieren? Hier sind neue Erhebungsinstrumente gefragt, die bereits zu einem guten Teil im Projekt umgesetzt und angewendet werden konnten, zugleich aber auf eine breitere Basis gestellt werden müssen.

Ebenso konnte die MSSC im Projekt ein Befragungsverfahren entwickeln, dass im Mix von Befragungsmoderation und Online-Datenerhebung effizienzsteigernde Resultate hervorbrachte. Für die weitere Entwicklung und den Ausbau der Verfahrenstechniken von Online-Fokusgruppenbefragungen kann die MSSC hier weiterführende Impulse in Zukunft konstruktiver aufgreifen und vorantreiben.