ARBEITSKREISE DER GLDV

AK - Korpora

Korpus basierte Lexikographie

Arbeitskreis Korpora tagt am 8.6.95 an der Univ. Stuttgart, Institut für maschinelle Sprachverarbeitung - Computerlinguistik (IMS-CL)

Robert NEUMANN begrüßte in seiner Eigenschaft als Leiter des Arbeitskreises "Korpora" innerhalb der GLDV die Referenten und die Gäste und dankte dem Institut für maschinelle Sprachverarbeitung der Universität Stuttgart für die Ausrichtung des Treffens. Er stellte erfreuliche Kontinuität in der wissenschaftlichen Diskussion des Arbeitskreises fest, die sich in diesem sechsten Treffen unter seinem Vorsitz manifestiert. Es folgte die Begrüßung durch den Gastgeber Ulrich HEID, der in groben Zügen sein Institut vorstellte, das im Jahre 1987 als Forschungszentrum für zentrale Fragen der theoretischen Linguistik und der maschinellen Bearbeitung von natürlicher Sprache gegründet wurde. Es besteht aus den Abteilungen "Computerlinguistik" (Prof. Christian Rohrer), "Experimentelle Phonetik und Phonologie" (Prof. Grzegorz Dogil), "Logik und Sprachphilosophie" (Prof. Hans Kamp) und" Theoretische Computerlinguistik" (Prof. Mats Rooth).

Robert NEU MANN berichtete in seinem Vortrag "MECOLB (Multilingual Environment for Corpusbased Lexicon Building): Ein Schritt auf dem Weg zur korpusbasierten Lexikographie" über das EU-Projekt COSMAS-II, dessen Projektkoordinator er selbst ist, und das Konzept der virtuellen Korpora und ihrer Multidimensionalität. Im einzelnen beschrieb er die fachlichen, innovatorischen, statistischen, informatischen und linguistischen Dimensionen von COSMAS und die Möglichkeiten, die das erweiterte System für die Verifizierung und Falsifizierung von linguistischen Hypothesen bietet.

Cyril BELICA ("Statistische Analyse von Zeitstrukturen in Korpora") hat am Institut für deutsche Sprache im Zusammenhang mit dem Lexikographie-Projekt "Wendewortschatz" eine Methode zur Berechnung, Analyse und Auswertung zeitrelevanter statistischer Parameter von Texten entwickelt und implementiert, die er sehr anschaulich vorstellte. Auf einem konkret festgesetzten Signifikanzniveau werden sprachliche Einheiten isoliert, deren zeitliche Verteilung im Text einer durch den Sprachwissenschaftler angenommenen "natürlichen" Verteilung widerspricht. Mit diesen mathematisch-statistischen Aussagen über einen Sachverhalt von Zeichenketten im Text können sehr leicht bereits mehr oder weniger stark frequentierte, aber auch potentielle Neologismen in den Korpustexten - das Analysekorpus umfaßte 4 Millionen laufende Wörter - ermittelt werden. Es werden dadurch Evidenzen bei der synchronischen und diachronischen Sprachanalyse aufgezeigt; die Signifikanz solcher Auffälligkeiten kann somit besser beurteilt werden, und explizite Hinweise auf die Interpretation dieser Auffälligkeiten werden möglich.

Der Vortrag machte klar, daß das vorgestellte Verfahren eine Analyse der zeitlichen Entwicklung von Sprachphänomenen ermöglicht. Es kann für das automatische

Filter von Texten für Monitorkorpora, als Programm zur Überwachung externer Textquellen und zur automatischen Aktivierung der Aufzeichnungseinrichtungen von Nachrichtendiensten angewendet werden.

Ulrich HEID sprach über" Texterschließungswerkzeuge als Hilfsmittel für den korpusbasierten Aufbau von Wörterbüchern"; seine Vorbemerkungen betrafen den institutionellen Rahmen zum Projekt "Textkorpora und Erschließungswerkzeuge", die Softwareentwicklung und die Entwickler. Es herrsche Einigkeit über die Notwendigkeit und die Vorteile einer korpusbasierten Lexikographie, aber bisher seien nicht sehr viele Werkzeuge zur Unterstützung der korpuslexikographischen Arbeit entwickelt worden. Eine alternative Vorstellung zum einmal erstellten und nie veränderten Wörterbuchaufbaumodell ist die ständige Verfeinerung der lexikalischen Modellierung der inkrementellen Daten, was eine iterative und zunehmend verbesserte Korpusabfrage bedeutet. Ulrich HEID entwickelte die Arbeitsschritte für die linguistische Korpusanalyse und -annotation sowie die Abfrage mit Hilfe von Konkordanzwerkzeugen. Die zur Zeit verfügbaren Tools sind Tokenizing (Wortformen, Satzgrenzen), eine morphosyntaktische Analyse und Lemmatisierung, ein Part-of-Speech-Tagging, ein partielles Parsing zur Erkennung phrasenstruktureller Konstrukte. Als Anwendungsbeispiel führte der Referent ein Lexikonfragment für lexikalisierte und nichtlexikalisierte Funktionsverbgefüge vor.

Dr. Achim STEIN ("Französische und italienische Korpora") berichtete über die Erstellung der Korpora, ihre Ressourcen und Anwendungsbereiche. Zur Erstellung der Korpora gehören die Erarbeitung eines Grundformenlexikons mit Tags für die Flexionsklassen, die Expansion zum Vollformlexikon (maximales Tagset), die Aufbereitung des Textes und die morphologische Annotierung. Als Hauptprobleme nannte der Referent im Zusammenhang mit dem Französischen und Italienischen die Behandlung von Eigennamen und Abkürzungen, von Mehrwortlexemen und agglutinierenden italienischen Pronomina und problematische Worttrennungen.

Oliver WAUSCHKUHN ("Statistisches Tagging als Vorstufe zur partiellen syntaktischen Analyse deutscher Texte: Einfluß auf die Anzahl der Parse-Ergebnisse") ging der Frage nach, ob der Einsatz eines Taggers als Präprozessor der syntaktischen Analyse zur A-priori-Disambiguierung für Parse-Ergebnisse dienen kann, denn für die linguistische Erschließung von Textkorpora z.B. durch Lexikographen wird zusätzlich zu den Tags eine strukturelle (syntaktische) Annotation benötigt. WAUSCHKUHN untersuchte satzweise die partielle syntaktische Analyse eines kleinen Korpus (aus der "Stuttgarter Zeitung" 3776 Sätze) jeweils in getaggter und in morphosyntaktisch ambig annotierter Form und verglich quantitativ die Anzahl der gelieferten Ergebnisse zwischen getaggter und nicht getaggter Form des Textes Als Schlussfolgerung schlug er entweder eine Verbesserungsmöglichkeit des Taggers oder ein zusätzliches Anbringen von Annotationen bei den zu erwartenden Fehltags und die Festlegung einer klaren Schnittstelle zwischen der Korpusannotation und der syntaktischen Analyse (Grammatik) sowie eine Verbesserung der NP-Grammatikregeln vor.

Petra STEINER gab im Vortrag "Das Münster-Tagging-Projekt (MTP)" einen Überblick über aktuelle Arbeiten am deutschen Textkorpus an der Westfälischen Wilhelms- Universität Münster, über die Aquirierung und Bearbeitung der ca. 100 Millionen Token - vorwiegend aus der "Frankfurter Allgemeinen" und der" Zeit" der Jahre 1990 bis 1992 - und über die Ergebnisse des automatischen Taggings. Ein verhältnismäßig geringer, zahlenmäßig aber beachtlicher Anteil wird mit Hilfe von bereits vorliegenden Informationen automatisch getaggt und mit einem speziellen Editor intellektuell disambiguiert bzw. korrigiert. Der größte Teil des Korpus soll lediglich mit Hilfe von automatischen Verfahren annotiert werden. Das Tagset genügt folgenden Anforderungen: klare Struktur mit mnemotechnischer Qualität, Operationalisierbarkeit und Intersubjektivität, Vorhersagbarkeit, Klasseneinteilung, Genauigkeit, Kompatibilität und Adaptabilität.

Holger TREBBE erhellte in seinem Beitrag "Das Hidden-Markov-Modell und seine Entwicklungsmöglichkeiten" die programmtechnische Seite der automatischen Verfahren. Innerhalb des "Münster Tagging Project" existieren momentan zwei Tagger, ein Counting-Rate-Tagger und ein HMM-Tagger. Beide Modelle beruhen auf der Voraussetzung zweier gekoppelter stochastischer Prozesse, die die Wortformenfolge/ Ambiguitätsklassenfolge und die Wortartenfolge beschreiben. Die Ereignisse der Wortformenfolge sind nicht direkt voneinander abhängig, sondern nur über die Wortartenfolge. Dabei wird auf der Basis der Bigramme gearbeitet. Da der Counting-Rate-Tagger mit relativen Häufigkeiten arbeitet und der HMM- Tagger auf dem Maximum-Likelihood-Konzept beruht, läßt sich die Korrektheitsrate erhöhen, wenn als Initialwerte des HMM- Taggers die ausgezählten relativen Häufigkeiten benutzt werden. Als Weiterentwicklung soll das VQHMM implementiert werden, das zusätzlich zur obigen Abhängigkeitsstruktur noch die Abhängigkeit der Wortformenfolge auf Basis der Bigramme unterstellt. Als Zusatzinstrument wurde eine Smoothing-Funktion vorgestellt, die verhindert, daß zu schätzende Parameter den Wert Null annehmen. Weiterhin wurde eine Gütefunktion dargestellt, die es ermöglicht, die Effektivität von Tag-Algorithmen trotz unterschiedlicher Tagsets und Testkorpora zu beurteilen.

Dr. Folker CAROLI sprach über "Korpusgestützte Grammatikentwicklung im Projekt LS-GRAM (Large Scale Grammar Development)", ein Projekt des Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes (IAI) und zeigte anhand von Beispielen, wie bestimmte Methoden der Korpusuntersuchung die Grammatikuntersuchungen unterstützen. Die Eigenschaften der Anwendungsorientiertheit, der Erweiterbarkeit und der Wiederverwendbarkeit gelten für die in beiden Bereichen behandelten sprachlichen Phänomene - wie beispielsweise die Behandlung von Appositionen und Parenthesen.

Zum Abschluß des Arbeitstages führte Oliver CHRIST die Datenbank und das Recherchesystem des Instituts für maschinelle Sprachverarbeitung vor.

Irmtraud Jüttner, Mannheim

AK - Hypermedia

Das Treffen des Arbeitskreises Hypermedia der GLDV am 18.9.1995 in Mannheim

Der Arbeitskreis Hypermedia fand sich am 18. September 1995 bei freundlichem Spätsommerwetter zu einem ersten, konstituierenden Treffen im Institut für deutsche Sprache in Mannheim ein. Das Treffen diente vornehmlich dazu, sich kennenzulernen, dabei gemeinsame Interessen und Themen zu erkunden und über künftige Aktivitäten zu beraten.

Bei der Vorstellungsrunde wurden folgende Projekte angesprochen: Das am Institut für deutsche Sprache durchgeführte Pilotprojekt Grundlagen eines grammatischen Informationssystems GRAMMIS, in dem grammatische und lexikalische Datenbestände des IDS als Hypermedia Anwendungen präsentiert und mit verschiedenen Benutzergruppen getestet werden (A. Storrer); das am GMD-Institut für Integrierte Publikations- und Kommunikationssysteme (IPSI) angesiedelte, EU-geförderte Projekt Editors Workbench, in dem im Rahmen des Aufbaus eines wissensbasierten Publikationssystems eine Hypermedia-Kunstenzyklopädie entsteht (W. Möhr); das EU-geförderte Projekt LOTOS Learning Tools, bei dem u. a. an der Universität Tübingen ein modular aufgebautes Programmsystem für das computergestützte Fremdsprachenlernen konzipiert und entwickelt wird, das ein Lerner-Lexikon, eine pädagogische Grammatik und Werkzeuge zur automatischen Fehlerdiagnose umfaßt *(K.* Krüger-Thielmann); das LINGUA-Projekt GRECOTERM, in dem die Firma ZERES multimediale Lehrmaterialien für deutsche und griechische Touristikfachkräfte entwickelt sowie ein ebenfalls bei ZERES angesiedeltes Projekt, in dem multimediale Produktbeschreibungen für exportorientierte Weinkooperativen im

Rioja für das WWW erstellt werden (L. Lemnitzer); den von der Gruppo DIMA in Turin entwickelten NL-Parser DIMACHECK, für den am IAI Lingware zum Deutschen (Morphologie, Lexikon, Grammatik) entwickelt wurde, der als Syntaxchecker in eine Textverarbeitungssystem oder in Systeme für die computerunterstützte Fremdsprachenvermittlung integriert und mit Hyperdokumenten zu Grammatik und Lexikon des, Deutschen verbunden werden kann (S. Rieder, U. Reuther); das am wissenschaftlichen Zentrum der IBM in Heidelberg angesiedelte Projekt COALA (Computer Aided Language Acquisition), in dem u. a. an einer elektronischen Grammatik für das Deutsche gearbeitet wird, die mit einem als Hypertext organisierten Grammatiktutorial und einem Hypertext- Wörterbuch vernetzt wird (B. Harriehausen-Mühlbauer) .

Trotz der Verschiedenheit der Anwendungen (Publikationssysteme, Informationssysteme, Systeme zum computergestützten Fremdsprachenlernen, Grammatikprüfsysteme) ergaben sich erfreulich viele gemeinsame Fragestellungen, die im Rahmen des Arbeitskreises weiter diskutiert werden sollen: Wie kann grammatisches und lexikalisches Wissen mit Hypermedia modelliert und vermittelt werden? Wie lassen sich multimediale Gestaltungsmittel sinnvoll einsetzen? Wie können" traditionelle" Texte in Hypermedia-Anwendungen umgestaltet werden?

Ein weiterer Interessensschwerpunkt betrifft die Verbindung von lexiko- und grammatikographischen Hypermedia-Anwendungen zu Grammatikprüfsystemen und Systemen zum computerunterstützten Fremdsprachenlernen. Daneben besteht auch ein großes Interesse an Informations- und Erfahrungsaustausch im Bereich der Werkzeuge, mit denen sich Hypermedia Applikationen erstellen lassen. Ein Anliegen des Arbeitskreises wird es deshalb sein, Informationen über einschlägige Tools, Projekte, Diskussionslisten, Workshops und Tagungen zu sammeln und weiterzugeben. Als Sammelstelle hat sich das Institut für deutsche Sprache angeboten, die auch die WWW-Seite des AK (http://www.ids-mannheim.de/grammis/ak.html) weiter pflegt und darüberhinaus eine sogenannte Informationsbörse mit WWW-Adressen zum Thema anbietet.

Zum Abschluß des Treffens wurde die aktuelle Version des GRAMMIS-Systems vorgeführt, die neben grammatischen Informationen zu den Wortarten und den kommunikativen Funktionen des Deutschen auch ein elektronisches Glossar grammatischer Termini anbietet und über Schnittstellen zu einer Datenbank der deutschen Funktionswörter und zu einer Valenzdatenbank verfügt.

Als nächste größere Aktivität ist ein Arbeitstreffen zum Thema "Hypermedia in Grammatikographie und Lexikographie" geplant, das am 21. und 22. März am Institut für deutsche Sprache in Mannheim stattfindet; an der Organisation sind Bettina Harriehausen-Mühlbauer, Wiebke Möhr und Angelika Storrer beteiligt. In diesem Workshop möchten wir uns - aus theoretischer und aus anwendungsbezogener Perspektive - mit den neuartigen Gestaltungsmöglichkeiten befassen, die Hypermedia in den Bereichen Lexikographie, Terminographie und Grammatikographie eröffnet. Dabei interessieren uns konzeptionelle, textlinguistische und informatische Aspekte ebenso wie konkrete Anwendungen, z.B. Grammatiken und Lexika für Sprachlernsysteme, WWW-Lexika, innovative elektronische Wörterbücher oder lexikalische Datenbanken. Weiterhin gibt es Beiträge zu Methoden und Werkzeugen, mit denen vorhandene Publikationen in Hypermedia-Anwendungen umgesetzt werden können, zu Fragen der Evaluierung und Benutzungsforschung in diesem Bereich, zur Konzeption von CALL-Systemen sowie zu Gesichtspunkten der Ergonomie und des Grafik-Designs. Aktuelle Informationen zum Workshop finden sich unter http://www.ids-rnannheirn.de/grarnrnis/work.htrnl.

Angelika Starrer, Heidelberg