

ARBEITSKREISE DER GLDV

Stand der Korpustechnologie in Deutschland und internationale Entwicklungen

Arbeitskreistreffen KORPORA der Gesellschaft für Linguistische Datenverarbeitung (GLDV) im Februar'94 am IDS, Mannheim.

Am 11. Februar 1994 tagte der Arbeitskreis "Korpora" der Gesellschaft für Linguistische Datenverarbeitung, der von Robert Neumann geleitet wird, im Institut für deutsche Sprache, Mannheim, und beschäftigte sich mit der Thematik "Stand der Korpustechnologie in Deutschland und internationale Entwicklungen - Projekte der Europäischen Union".

Rainer Wimmer, Geschäftsführender Direktor des Instituts für deutsche Sprache, begrüßte die Teilnehmer des Treffens und hob die Wichtigkeit der Korpustechnologie für die germanistische Linguistik hervor.

Robert Neumann eröffnete die eintägige Zusammenkunft und umriß das Ziel der aktuellen Sitzung des Arbeitskreises: einen Überblick über die deutschen und internationalen Korpus-Aktivitäten zu gewinnen und die gesammelten Erkenntnisse in Form mehrerer bibliographischer Zusammenstellungen zu Korpus-Arbeiten in Deutsch in einen Listserver zu geben und allen Interessenten so einen Zugang zu gewähren.

In seinem Eröffnungsvortrag stellte R. Neumann das Institut für deutsche Sprache als zentrale außeruniversitäre Forschungseinrichtung vor und sprach über die Aufgaben der Arbeitsstelle "Linguistische Datenverarbeitung" im IDS.

Die Vortragsreihe eröffnete Wolf Paprotté mit der Vorstellung des Projekts 62050 MULTTEXT (Multilingual Text Tools

and Corpora). Die Diskussion dazu berührte vor allem Fragen der Textauszeichnung (I. Batori), des morphologischen Taggings (A. Storrer), die Definition von Speech-Dateien und Standardisierungsprobleme (R. Neumann).

Wolfgang Teubert berichtete über das Projekt "Network of European Reference Corpora" (NERC), einen Vorschlag für die Erarbeitung nationaler Korpora.

Angelika Storrer und Helmut Feldweg sprachen über ein System zur Wissensrepräsentation mit der Bezeichnung ELWIS (Korpusunterstützte Entwicklung lexikalischer Wissensbasen).

R. Neumann stellte als drittes Projekt in der Europäischen Union "Multilingual Environment for Corpus-Based Lexicon Building" (MECOLB) vor, das für die Abspeicherung und Verwaltung von großen Textmengen geeignete Werkzeuge entwickelt, die den Aufbau und die Wartung von maschinenlesbaren Lexika unterstützen.

In der anschließenden Diskussion gab R. Neumann bekannt, daß die Arbeitsstelle "Linguistische Datenverarbeitung" des IDS einen Listserver für Korpus-Fragen eingerichtet hat, der ein permanentes Informationsbrett und ein Diskussionsforum der Mitglieder des Arbeitskreises sein soll. Die Bibliographie zum NERC-Projekt soll über den Listserver allen Mitgliedern des Arbeitskreises zur Verfügung gestellt werden, und W. Teubert stimmt diesem Verfahren als Urheber der Bibliographie zu.

Winfried Lenders, Vorsitzender der GLDV, sagte, daß neben dem Informationsaustausch über den Listserver die persönliche Kommunikation der Experten

sehr wichtig sei und daß der Arbeitskreis sich zweimal im Jahr zu Tagungen zusammenfinden sollte. Weiterhin sollen die Mitglieder des Arbeitskreises und des Korpus-Workshops miteinander korrespondieren, und es sollten regelmäßig Informationen über die Projekte und über erzielte Ergebnisse und angewandte Methoden ausgetauscht werden. Materialien, die alle Mitglieder interessieren, sollten über die elektronischen Medien zur Verfügung gestellt werden; die verschiedenartigsten Aktivitäten sollten weitergeleitet werden und darüber regelmäßig in der Zeitschrift "LDV-Forum" berichtet werden.

R. Neumann stellte - die Diskussion zusammenfassend - fest, daß vor allem Clearing-Aufgaben zu den wichtigsten Gegenständen der Arbeit des Arbeitskreises gehören, und schlug eine systematische Aufbereitung von Informationen über die deutschen Textkorpora vor, denn alle deutschsprachigen Korpora sollten zumindest im nationalen Rahmen aufgelistet greifbar sein. Die Mitglieder des Arbeitskreises werden aufgerufen, maschinenlesbare Texte - auch phonetische Korpora verfügbar zu machen und an den Listserver zu melden, so daß eine Liste aller vorhandenen Texte erarbeitet werden kann.

Cyril Belica führte zum Abschluß der Veranstaltung das Computersystem COSMAS (Corpus Storage, Maintenance and Access .System) vor, mit dessen Hilfe Textbelege für einzelne Wortformen und Satzzeichen, für Wortformen in verschiedenen Kombinationen und mit unterschiedlichen Abständen gesucht werden können. COSMAS wurde vom Institut für deutsche Sprache in Zusammenarbeit mit der Firma Makrolog, Wiesbaden, entwickelt, um für sprachwissenschaftliche und andere sprachbezogene Forschungsvorhaben die umfangreichen Textkorpora zur geschriebenen und gesprochenen deutschen Sprache zu erschließen und zu pflegen.

Das nächste Treffen des Arbeitskreises "Korpora" wird am 11. Juni 1994 an der Westfälischen Wilhelms-Universität Münster zum Thema "Aktivitäten zu Korpora der gesprochenen Sprache in der Bundesrepublik Deutschland: VERBMOBIL"

stattfinden.

Irmtraud Jüttner, Mannheim

AK - Lexikographie

Der Arbeitskreis Lexikographie möchte in den bevorstehenden Monaten mit folgenden Aktivitäten innerhalb der GLDV in Erscheinung treten:

1. Erstellung einer möglichst vollständigen Bibliographie zur maschinellen Lexikographie (*computationallexicography*) in Deutschland. die Bibliographie soll zentral in Bonn gesammelt werden. Voraussetzung für das Gelingen ist jedoch eine aktive Teilnahme möglichst zahlreicher Beitragender - vornehmlich durch die Bereitschaft, Literaturangaben maschinenlesbar in einem verabredeten Format zur Verfügung zu stellen.
2. Bestandsaufnahme und Bekanntmachung von maschinelesbaren lexikalischen Ressourcen, die für Forschungszwecke verfügbar gemacht werden können. Eine kleine Dokumentation soll Auskunft darüber geben, wo welche Datei wie und unter welchen Bedingungen für Forschungszwecke genutzt werden können.

Dr. Nico Weber (Sprecher) Universität
Bonn
IKP
Poppelsdorfer Allee 47
53115 Bonn
Tel.: 0228/73 56 44
Fax: 0228/73 56 39