Abschlußbericht

2. Februar 1995

Projekt	Forschungsschwerpunktprogramm vom Ministerium für Wissenschaft und Forschung des Landes Baden-Württemberg
Thema	ELWIS. Korpusunterstützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik
Projektleiter	Prof. Dr. Erhard W. Hinrichs
Mitarbeiter	Helmut Feldweg, M.A. Marie Boyle-Hinrichs (bis 1.5.1994) Ralf Hauser, M.A. (ab 15.5.1994)

Inhaltsverzeichnis

1 Ziel, Kenntnisstand, Ausgangsfragestellungen

1.1 Ziel

Im Antrag für die Förderungsphase 1992-1994 wurde das Ziel wie folgt definiert:

> "Ziel des beantragten Projekts ist die Entwicklung von lexikalischen Wissensbasen, die in der Computerlinguistik für eine Vielzahl von Forschungsgebieten benötigt wird, wie etwa für natürlich-sprachliche Dialogsysteme, für die maschinelle Übersetzung und für die Diskursverarbeitung ...
>
> Durch die Verbindung von maschinell lesbaren Lexika mit Ressourcen, die in großen Textkorpora implizit vorhanden sind, bietet sich die Möglichkeit, die automatisierte Erstellung von lexikalischen Wissensbasen signifikant zu verbessern. Es ist daher das Ziel des beantragten Projektes, ein computerlinguistisches Lexikon zu entwickeln, das auf der Verzahnung dieser beiden traditionellen computerlinguistischen Ressourcen beruht. Dabei soll die computerlinguistische Verarbeitung des Deutschen im Vordergrund stehen, obwohl das Augenmerk nicht auf mono-linguale Lexika beschränkt werden soll."

1.2 Kenntnisstand bei der Antragsstellung

Die Verfügbarkeit großer Mengen von qualitativ hochwertigen lexikalischen Daten ist die Voraussetzung für die Entwicklung leistungsfähiger sprachverarbeitender Syteme. Andererseits erfordert die computergerechte Beschreibung dieser Daten ein hohes Maß an linguistischem Know–how, so daß der Lexikonaufbau einen entscheidenden Kostenfaktor bei der Entwicklung solcher Systeme darstellt. Dies ist einer der Gründe für die zunehmenden Forschungsbemühungen, die darauf abzielen, bestehende lexikalische Quellen für den Aufbau von lexikalischen Wissensbasen zu nutzen.

Als Quellen lexikalischen Wissens kommen dabei in Frage:

- **Textkorpora** in maschinenlesbarer Form.
- **maschinenlesbare Wörterbücher**, d.h. für menschliche Benutzer erstellte lexikographische Produkte, die in maschinenlesbarer Form verfügbar sind.
- **Computerlexika**, d.h. Mengen von klassifizierten lexikalischen Daten, die für computerlinguistische Anwendungen kodifiziert worden sind.

Verfahren, lexikalisches Wissen aus solchen Quellen automatisch zu extrahieren, wurden bislang überwiegend an englischem Sprachmaterial erprobt. Dabei läßt sich grob differenzieren zwischen **korpusbasierten Methoden**, die quantitative und statistische Berechnungen über das gemeinsame Vorkommen lexikalischer Einheiten in mehr oder weniger fein annotierten Textkorpora auswerten (z.B. Anick und Pustejovski, 1990; Church und Hanks, 1990; Church et al., 1991; Smadja, 1991) und **wörterbuchbasierten Methoden**, bei denen u.a. Grammatikcodes bzw. Bedeutungserläuterungen von maschinenlesbaren Wörterbüchern in ein für Sprachverarbeitungsprogramme verwertbares Format konvertiert wurden (z.B. Alshavi, 1989; Byrd et al., 1987; Brustkern, 1988; Veronis und Ide, 1991; Montemagni und Vanderwende, 1992). Momentan steht jedoch kein ausreichend großes repräsentatives Korpus der deutschen Gegenwartsprache allgemein zur Verfügung.

1.3 Ausgangsfragestellungen

Im Rahmen des Projektes wurde versucht das Problem des "lexikalischen Flaschenhalses" zu lösen, indem auf bereits an anderen Sprachen erprobten Ansätzen aufgesetzt wurde und deren Übertragbarkeit auf die Beschreibung des Deutschen evaluiert wurde. Gleichzeitig wurden neue eigenständige Ansätze für spezielle Fragestellungen entwickelt. Es wurden zwei Arbeitsbereiche unterschieden:

- Akquisition: Nutzbarmachung vorhandener lexikalischer Wissensquellen (Korpora, maschinenlesbare Wörterbücher, Computerlexika) für den effizienten Aufbau lexikalischer Wissensbasen
- Repräsentation: Adäquate, theoretisch fundierte Beschreibung lexikalischen Wissens für verschiedene Typen computerlinguistischer Anwendungen.

2 Ergebnisse und ihre Bedeutung

Insgesamt ergibt sich aus den verschiedenen im Projekt durchgeführten Aktivitäten eine prototypische Instantiierung des **sog. Wiederverwendbarkeits–Szenarios** (vgl. Heid, 1991): Das in verschiedenen lexikalischen Ressourcen gespeicherte lexikalische Wissen wird expliziert, ggf. umstrukturiert und in einer lexikalischen Datenbank abgespeichert, aus der dann neue lexikalische Quellen in verschiedenen Formaten und Formalismen generiert werden. Die Schwerpunkte liegen dabei im Repräsentationsbereich auf der Modellierung ausgewählter Bereiche in den Repräsentationssprachen DATR; im Akquisitionsbereich steht die Entwicklung einer computer– und korpusunterstützten Beschreibungsmethodologie für Flexion, Wortbildung, Valenzeigenschaften, Phraseolexeme und kollokative Verbindungen im Vordergrund. Die Abbildung 1 zeigt die am Gesamtkonzept beteiligten Komponenten in ihrem Zusammenhang:

- Kernstück ist die **lexikalische Datenbank**: sie bildet einerseits die Zielkomponente für die Forschungen im Bereich Akquisition, andererseits fungiert sie als Datenlieferant für die in verschiedenen alternativen Formalismen formulierten Beschreibungen lexikalischen Wissens.

Abbildung 1: ELWIS–“Szenario”

- Die lexikalische Datenbank wird aufgefüllt durch Informationen, die mit spezifischen Methoden und Werkzeugen aus **maschinenlesbaren Wörterbüchern, Computerlexika und annotierten Textkorpora** extrahiert werden. Bestimmte Bereiche der maschinell extrahierten Daten müssen anschließend intellektuell überprüft und gegebenenfalls ergänzt werden; hierzu wird eine mit den entsprechenden Grundfunktionen ausgestattete **Benutzerschnittstelle** zur Datenbank sowie zu den Textkorpora entwickelt.
- Die in der Datenbank gespeicherten Informationen können wieder genutzt werden, um die Aufarbeitung der Ressourcen, z.B. die Annotation der Korpora, zu verbessern.
- Die lexikalische Information aus der lexikalischen Datenbank wird schließlich mit Hilfe von speziellen Compilerprogrammen in **lexikalische Wissensbasen** integriert.

2.1 Lexikalische Wissensakquisition

2.1.1 Aufbereitung von Textkorpora

Verfügbare Korpora

Zu Beginn der Förderung des Projektes waren ausreichend große Textkorpora des Deutschen nicht allgemein zugänglich. Daher wurde frühzeitig eine Kooperation mit dem Institut für deutsche Sprache (IDS) realisiert, die dem Projekt den Zugriff auf die dort vorhandenen umfangreichen Datenbestände ermöglichte. Darüberhinaus wurde eine größere Anzahl kleinerer Korpora akquiriert und erschlossen. Schließlich sind 1994 mit der Veröffentlichung der CD-ROM *Multlingual Corpus 1* der *European Corpus Initiative* (vgl. Association for Computational Linguistics, 1994) Zeitungsdaten allgemein für Forschungszwecke verfügbar geworden, so daß heute kein Mangel mehr an Deutschen Rohtextdaten besteht.

Da diese Entwicklung abzusehen war, wurde beim Aufbau eines eigenen Korpus bewußt auf eine Ergänzung dieser überwiegend auf Zeitungstexten basierenden Korpora Wert gelegt. Dies wurde durch den kontinuierlichen Aufbau eines deutschen Newskorpus realisiert. Es handelt sich dabei um eine äußerst umfangreiche Sammlung von Beiträgen zu deutschsprachigen News-Gruppen im Internet. Diese Textsorte deckt inhaltlich ein sehr breites Spektrum ab und bietet linguistisch einen sehr interessanten, selten untersuchten und ergiebigen Forschungsgegenstand, da es sich weitgehend um nicht–editierte schriftsprachliche Äußerungen handelt, in der auch viele Merkmale gesprochener Sprache repräsentiert sind. Die in hohem Maße idiosynkratische Auszeichnung und Orthographie dieser Texte stellt überdies höchste Ansprüche an automatische Verfahren zur Textaufbereitung und –analyse. Erste Ergebnisse einer statistischen Analyse dieses Korpus werden in Feldweg et al., 1995 veröffentlicht.

Korpus Tools

Zur Aufbereitung und Analyse dieser Korpora wurde eine fundierte softwaretechnische Infrastruktur geschaffen. Dabei konnte weitgehend auf vorhandene und verfügbare Software, die sich für die Analyse anderer, meist englischer Korpora bewährt hat, zurückgegriffen werden (CLAN, 1994 und MacWhinney, 1995, Tuthill, 1981, Feldweg, 1993, Zimmermann, 1987, Schulze und Christ, 1994 und Christ, 1994a/1994b, Prinz, 1993, Neumann, 1992). Die meist ursprünglich für die Analyse englischsprachiger Korpora entwickelte Software mußte allerdings in einer Reihe von Fällen modifiziert werden, bspw. zur korrekten Behandlung von 8–Bit–Zeichensätzen, die für eine akzeptable Repräsentation von Sonderzeichen, Umlauten und anderen Diakritika benötigt werden.

Für die Extraktion von Mehrwortlexemen aus Textkorpora wurde im Projekt NGRAM entwickelt. Das in C geschriebene Programm extrahiert aus Textkorpora gewichtete Häufigkeitslisten von Wortfolgen basierend auf einem Fenster von n Wörtern, wobei n vom Benutzer vorgegeben werden kann. Die Gewichtung erfolgt durch Berechnung eines Wertes der *Mutual Information*, der Relation der Einzelfrequenzen der an der Wortfolge beteiligten Wörter zur Frequenz der Wortfolge selbst. Verschiedene Optionen erleichtern die Konfigurierbarkeit des Programmes, bspw. können Wort- und Phrasenbegrenzungssequenzen frei gewählt werden, der Skopus eines n-Grammes kann auf Phrasen begrenzt werden, und die innerhalb eines n-Grammes auftretenden Wörter können permutiert werden (vgl. Feldweg und Hipp, 1993)

Bei der Arbeit mit diesen, im wesentlichen auf Vorkommen einzelner Wortformen im Text basierenden Werkzeugen hat sich gezeigt, daß linguistisch anspruchsvollere Analysen deutscher Texte durch die morphologisch bedingte Formenvielfalt noch eingeschränkter sind, als dies bspw. für das morphologisch ärmere Englisch gilt. Eine Reihe von für das Englische positiven Ergebnissen solcher Verfahren, konnten für das Deutsche nicht bestätigt werden.

Aus diesem Grund wurde bei der Entwicklung von Tools ein Schwerpunkt auf Verfahren zur linguistischen Annotierung von Textkorpora gelegt und ein Tagger zur Annotation deutscher Texte mit Wortarten entwickelt.

Wortartendisambiguierung mit LIKELY

Obwohl statistische Verfahren zur Disambiguierung von Wortarten in Texten spätestens seit Ende der achtziger Jahre bekannt sind und sehr positive Ergebnisse für das Englische vorliegen, lag bislang nur eine Studie zum Deutschen vor (Wothke et al., 1993), die sich auf einen nicht allgemein verfügbaren Tagger bezog. Im Projekt wurde daher ein auf einem Hidden–Markov–Modell (HMM) basierender Tagger in C entwickelt, der aus der Menge der möglichen Wortartkombinationen einer diesbezüglich ambigen Textsequenz die wahrscheinlichste Kombination mittels des Viterbi–Algorithmus auswählt. Eines der Hauptprobleme bestand dabei in der Gewinnung der für HMM–Tagger notwendigen Trainingsdaten. Hier konnte auf Vorarbeiten aus dem SFB 100 zurückgegriffen werden, die dem Projekt vom Institut für deutsche Sprache zur Verfügung gestellt wurden. Es handelt sich dabei um mit dem Saarbrücker Syntaktischen Analysesystem (SATAN) aufbereitete Texte, die unter anderem die Wortart für jedes im Wort auftretende Wort bereitstellen. Die in diesen Daten auftretende hohe Zahl von Fehlzuweisungen konnte im Projekt teils durch (semi-)automatische Verfahren, teils durch intellektuell/manuelle Korrekturen ganz erheblich reduziert werden. Ein Teil dieser bereinigten Daten wurden dann zur Gewinnung der für das HMM notwendigen Parameter benutzt, ein anderer wurde zur Evaluation der durch den Tagger gelieferten Ergebnisse herangezogen. Je nach Qualität des vom Tagger benutzten Vollformenlexikons lag die Fehlerquote bei der Wortartenzuordnung bei 3–7 % (vgl. Feldweg, 1993).

Das bei der Entwicklung dieses Taggers erworbene know–how konnte auch im von der EU geförderten COMPASS–Projekt verwendet werden, in dem ein kontextabhängiges Wörterbuchnachschlagesystem verwendet wird, das Bedeutungen auch mittels eines HMM–Wortartentaggers disambiguiert.

Darüberhinaus kam eine Vereinbarung mit dem Rank Xerox Research Centre in Grenoble zustande, das das deutsche HMM–Modell auch für kommerzielle Anwendungen nutzen wird.

Das POS-Tagset

Die stochastische Wortartendisambiguierung verwendete anfangs ein Wortartenschema oder POS–Tagset, das eine Teilmenge der vom Saarbrücker Lemmatisierungssystem SALEM verwendeteten Wortarten (vgl. Sonderforschungsbereich 100 “Elektronische Sprachforschung”, 1980) enthielt und mit dem auch das Referenzkorpus, das *Mannheimer Korpus I*, analysiert wurde. Es stellte sich heraus, daß sich in diesem Korpus eine nicht zu vernachlässigende Anzahl von fehlerhaften Wortartenzuordnungen befand, die teils durch automatische Verfahren eliminiert werden konnten und teils für eine intellektuelle Analyse zurückgestellt wurden. Eine intellektuelle Analyse wurde jedoch durch die unzureichende Dokumentation des Wortartenschemas erschwert. Dies führte zu einer Überarbeitung des bestehenden POS-Tagsets und der genauen Beschreibung von Richtlinien zur eindeutigen Wortartenzuordnung (Thielen und Sailer, 1994), die gewährleisten sollten, daß auch unterschiedliche Anwender zu denselben Ergebnissen kommen und somit die Konsistenz des Referenzkorpus und daraus resultierender getaggter Korpora erhalten bleibt. Das hier vorgestellte Tagset wurde in Zusammenarbeit mit Mitarbeitern des Instituts für maschinelle Sprachverarbeitung in Stuttgart im TC-Projekt (Textcorpora und Erschließungswerkzeuge) entwickelt, die sich mit ähnlichen Problemen bei der Wortartenannotierung konfrontiert sahen.

Mit diesem gemeinsamen POS-Tagset wird eine weitgehende Übereinstimmung bei der Korpus–Annotation erreicht und sichergestellt, daß die gegenseitige Nutzung bereits durchgeführter Korpusarbeit ohne umständliche Anpassung an unterschiedliche Tagsets möglich ist. Wir betrachten dieses Tagset als Vorschlag zu einer möglichen Standardisierung eines (deutschen) POS–Tagsets, wie sie vor allem von der TEI (Text Encoding Initiative, vgl. Sperberg-McQueen und Burnard, 1994), aber auch von anderen (vgl. EAGLES: Calzolari und McNaught, 1994, Cloeren, 1993, SUC: Ejerhed et al., 1992, Ejerhed, 1995), gefordert wird. Im EAGLES-Projekt (*Expert Advisory Group for Language Engineering Standards*) ist dieser Vorschlag zusammen mit einer morphologischen Erweiterung, die in Stuttgart entwickelt wurde, bereits in die Beschreibung morphosyntaktischer Phänomene für das Deutsche (Monacchini und Calzolari, 1994) eingegangen.

Das Tagset ist hierarchisch strukturiert und orientiert sich weitgehend am "TEI Starter Set Of Grammatical-Annotation Tags" (vgl. TEI AI1W2, 1991). Es umfaßt die Hauptwortarten Nomen, Verb, Artikel, Adjektiv, Pronomen, Kardinalzahl, Adverb, Konjunktion, Adposition, Interjektion und Partikel, die unterschiedlich stark subklassifiziert sind. Insgesamt enthält das Tagset 42 verschiedene POS–*tags*.

Zusätzlich zu den POS-*tags* sind auch *tags* für abgetrennte Wortteile und Satzzeichen vorgesehen. Mehrwortlexeme und fremdsprachliche Wörter werden bisher nicht berücksichtigt. Abkürzungen werden entsprechend ihrer ausgeschriebenen Form getaggt, z.B. *bzw.* als Adverb (ADV).

Die *tags* können jederzeit durch weitere linguistische Merkmale (z.B. morphologische) erweitert werden, sofern dies für eine Anwendung erwünscht oder notwendig ist.

Ein Überblick über das Tagset findet sich in der Tabelle 1, Seite 8.

Extraktion von Mehrwortlexemen

Für die Extraktion von Mehrwortlexemen (MWLs) aus beliebigen Textkorpora wurde ein Programm entwickelt, das aufgrund statistischer Beziehungen der Auftretenshäufigkeiten von Wortgruppen und deren einzelner Elemente Wortcluster mit signifikant hoher *mutual information* (MI) lokalisiert (vgl. Feldweg und Hipp, 1993). Mit diesem Verfahren allein ist es zwar nicht möglich, die so gewonnenen Daten als MWLs zu identifizieren und zu typisieren (Phraseolexem, Kollikation, Funktionsverbgefüge), denn ihre Einteilung erfolgt aufgrund von syntaktischen, semantischen oder auch kontrastiven Kriterien und das kann ein statistisches Verfahren nicht leisten, zumal die Übergänge zwischen den Gruppen meist fließend sind. Dazu müßte wenigstens ein geparstes Korpus vorliegen, wie das *Penn Treebank Corpus* (Marcus et al., 1993), das auch die Möglichkeit gibt, MWLs zu erkennen, deren Elemente in variierender Wortstellung vorkommen (vgl. Breidt, 1993b).

Dennoch bietet dieses Verfahren eine große Erleichterung sowohl bei der Erstellung großer lexikalischer Wissensbasen und bei lexikographischen Arbeiten als auch bei der Untersuchung und Klassifizierung von MWLs, da eine große Datenmenge potentieller MWLs leicht verfügbar gemacht wird.

2.1.2 Aufbereitung von Wörterbüchern

Wörterbücher liegen bislang für das Deutsche maschinenlesbar meist nur in der Form von Satzbanddateien vor, die beim Druck der Wörterbücher von der Satzmaschine erzeugt werden und die neben den eigentlichen Wörterbuchdaten all die Informationen enthalten, die ein

Tag–Name	Tag–Erklärung	Beispiele
N	**Nomina**	
NN	'normales' Nomen	*(der) Hut, (das) Singen, (das) Gute*
NE	Eigenname	*Peter, Bonn, Mercedes*
V	**Verben**	
VFIN	finites Verb	*sagt, hat, will*
VINF	Infinitiv	*sagen, haben, wollen*
VIZU	Infinitiv mit "zu"	*einzuatmen*
VPP	Partizip Perfekt	*gesagt*
ART	**Artikel**	*der, eine*
ADJ	**Adjektive**	
ADJA	attributives Adjektiv	*schöner (Baum), Wiener (Oper)*
ADJD	prädikativ und adverbial gebrauchtes Adjektiv	*(ist) schön, (singt) schön*
P	**Pronomina**	
PPER	stets irreflexives Personalpronomen	*ich, du*
PPERRF	reflexives und irreflexives Personalpronomen	*mich, dich*
PRF	reines Reflexivpronomen	*sich, einander*
PPOSS	substituierendes Possessivpronomen	*(das ist) meins*
PPOSAT	attribuierendes Possessivpronomen	*mein (Haus)*
PROS	substituierendes Demonstrativ–/Indefinitpronomen	*man (sagt)*
PROAT	attribuierendes Demonstrativ–/Indefinitpronomen	*keine (Menschen)*
PROAV	Pronominaladverbien	*damit (wird behauptet)*
PWS	substituierendes Interrogativpronomen	*wer (geht?)*
PWAT	attribuierendes Interrogativpronomen	*welche (Meinungen?)*
PWAV	adverbiales W-Pronomen	*wieso (wird gelacht?)*
PRELS	substituierendes Relativpronomen	*(der Mann,) der (gelacht hat)*
PRELAT	attribuierendes Relativpronomen	*(die Frauen,) deren (Meinungen)*
	Sonderklassen der Pronomina	
PNFL	nicht flektierendes Pronomen	*manch, solch, welch*
PALL	Formen von 'all–'	*(ihr) alle, alle (diese Menschen)*
PBEID	Formen von 'beid–'	*(meine) beiden (Brüder)*
PVIEL	nicht flektierte, nicht attributive Formen von *viel, wenig, zuviel, zuwenig, mehr, weniger*	*(er ißt) viel/wenig/zuviel*
CARD	**Kardinalzahlen**	*1, eins, 2, drei, 1/8, 25348*
ADV	**Adverbien**	*oft, heute, bald, vielleicht*
KO	**Konjunktionen**	
KOUI	unterordnende Konjunktion mit Infinitiv	*ohne (zu fragen)*
KOUS	unterordnende Konjunktion mit (finitem) Satz	*weil, als, daß, ob*
KON	nebenordnende Konjunktion	*und, oder, denn*
KOKOM	vergleichende Konjunktion	*(so groß) wie, (größer) als*
AP	**Adpositionen**	
APPR	Präpositon	*in, auf, um*
APPRART	Präposition mit Artikel	*im, aufs, unterm*
APPO	Postposition	*(des Geldes) wegen*
APZR	Zirkumposition rechts	*(um der Ehre) willen*
ITJ	**Interjektion**	*ach, oh, weh*
PTK	**Partikeln**	
PTKZU	Infinitiv–'zu'	*zu (sagen)*
PTKNEG	Negationspartikel	*nicht*
PTKVZS	abgetrennter Verbzusatz von Partikelverben	*(ich sehe Peter) zu*
PTKANT	Antwortpartikel	*ja, nein, bitte, danke*
PTKA	Partikel bei Adjektiv oder Adverb	*zu/allzu (schnell), am (besten)*
TRUNC	**Kompositions-Erstglied**	*Ein- (und Ausgang), be- (und entladen)*
$	**Interpunktion**	
$,	Komma	*,*
$(	Klammerzeichen	*( ) [ ]*
$.	sonstige Satzzeichen	*. ; ? ! :*
$\Sigma = 47$		

Tabelle 1: Tagset

Photobelichter braucht, um den Text im gewünschten Format wiederzugeben. Für die automatische Extraktion der lexikalischen Informationen, die im Wörterbuch zu den einzelnen Stichwörtern gegeben werden, müssen diese Satzbanddateien in eine strukturiertere Form überführt werden, in der die artikelinternen und –externen Strukturen des Wörterbuchtextes explizit dargestellt sind. Der Prozeß einer solchen Konvertierung wird im allgemeinen als "Wörterbuchparsing" bezeichnet (vgl. Neff et al., 1988; Neff und Boguraev, 1990; Bläsi und Koch, 1992).

Die dabei entstehende strukturierte Repräsentation des Wörterbuchtextes ermöglicht dann den gezielten Zugriff auf die im Mikrostrukturprogramm des Wörterbuchs verzeichneten Angabeklassen und ist insofern die Grundlage für die automatische Überführung der im Wörterbuch enthaltenen lexikalischen Informationen in die lexikalische Wissensbasis.

Für die entsprechenden Forschungen im ELWIS–Projekt hat die Duden–Redaktion freundlicherweise Satzbanddateien mit Ausschnitten des Stilwörterbuchs (Duden–2) und des Bedeutungswörterbuchs (Duden–10) zur Verfügung gestellt. Diese Wörterbücher erschienen für unsere Zwecke deshalb interessant, weil sie eine qualifizierte lexikographische Beschreibung lexikalischer Information enthalten, die auch im Vordergrund der ELWIS–Forschungen stehen: Das Stilwörterbuch enthält Satzmusterangaben zu Verben, beschreibt das Kollokationspotential und die kombinatorischen Eigenschaften der Stichwörter und verfügt über einen gut bearbeiteten phraseologischen Teil. Das Bedeutungswörterbuch (Duden–10) legt den Schwerpunkt auf die Beschreibung der Bedeutungseigenschaften und der Wortbildungsmöglichkeiten der erfaßten Stichwörter.

Um diese Satzbanddateien in eine strukturierte Form zu überführen, wurde ein Wörterbuchparser entwickelt, der den Satzbanddateien relativ zu einer vorgegeben Grammatik eine Strukturbeschreibung zuweist und diese in einer vom Benutzer definierbaren Form ausgibt (z.B. in hierarchischer Attribut–Werte–Notation oder als SGML–Struktur). Der Parser verfügt über einen eleganten und mächtigen Grammatikformalismus, der speziell auf die Erkennung von Artikelstrukturen anhand bestimmter Steuerzeichen zugeschnitten ist (vgl. Hauser, 1993 und Hauser und Storrer, 1993). Die Beschreibungen der hierarchischen Artikelstruktur in den Grammatiken für die beiden Wörterbücher basiert auf einer Theorie lexikographischer Textdeskription, wie sie in Wiegand, 1989a/1989b/1989c/1990 entwickelt wurde; das methodische Vorgehen bei der Entwicklung dieser Grammatiken basiert auf der eben dort skizzierten Methode der funktional–positionalen Segmentation von Artikeltexten, modifiziert und erweitert im Hinblick auf die speziellen Erfordernisse des Wörterbuchparsings. Aus den in strukurierter Form vorliegenden Wörterbuchdaten können nun gezielt bestimmte Angaben herausgegriffen und in die lexikalische Datenbank (vgl. Abschnitt 2.2) überführt werden.

Der Wörterbuchparser findet zur Zeit seinen Einsatz in dem LRE-Projekt der EU, COMPASS, in dem die Sprachrichtung Deutsch-Englisch des Collins Großwörterbuchs (Collins, 1991) für die Ansprüche eines *comprehension assistant* geparst wird.

2.1.3 Aufbereitung von Computerlexika

Im Gegensatz zu Wörterbuchdaten handelt es sich bei Computerlexika um lexikalische Ressourcen, die nicht zum direkten Gebrauch durch einen menschlichen Benutzer bestimmt sind, sondern von Sprachverarbeitungssystemen benutzt werden. Derzeit existieren nur sehr wenige Computerlexika mit einer nennenswerten Abdeckung der deutschen Lexik. Das Projekt hatte Zugriff auf zwei solcher Lexika: SDW und CELEX. Für beide Lexika wurde jeweils untersucht, ob sie sich fuer eine Wiederverwendung und Integration in die ELWIS-Datenbank eignen.

Bei SDW handelt es sich um das "Saarbrücker deutsche Wörterbuch", einem weiteren Produkt des SFB 100, besser bekannt unter dem Namen der kleineren Version, des "Saarbrücker Deutschen Analysewörterbuches (SADAW)". Das im SFB für das Lemmatisierungssystem SALEM und den Parser SATAN verwendete Lexikon wurde am IDS fortgeschrieben. Bedingt durch die langfristige Pflege durch wechselndes Personal ist das Lexikon in einigen Bereichen inkonsistent. Der Zugang wird weiterhin durch die heute altertümlich erscheinende, noch auf Lochkartentechnologie und extrem kleine Speicherkapazitäten basierende Datenstruktur erschwert. Trotz dieser Hindernisse konnten zentrale Teile dieses Lexikons in die ELWIS–Datenbank abgebildet und geladen werden.

Das "Centre for Lexical Information (CELEX)" am MPI in Nijmegen bietet in einer Datenbank lexikalische Informationen für die Sprachen Deutsch, Englisch und Niederländisch, die dem Projekt online und offline zur Verfügung standen. Im Gegensatz zum SDW sind diese Daten technisch gut erschlossen. Allerdings ist Deutsch unter den drei zur Verfügung stehenden Sprachen die qualitativ und quantitativ am schlechtesten erschlossene. CELEX bietet detaillierte und im Allgemeinen exakte Angaben zur Orthographie, Phonologie, Morphologie und syntaktischen Klassenzugehörigkeit der beschriebenen Lexeme und Wortformen. Keine Angaben finden sich hingegen in den deutschen CELEX-Beständen zur Subkategorisierung. Auch scheint die Auswahl der 50000 Lemmata mitunter zufällig, die 300000 Wortformen wurden aus diesen – nicht immer unter Berücksichtigung von Belegungsrestriktionen – generiert. Die Angaben zu Orthographie und Flexionsmorphologie des deutschen CELEX–Bestandes konnten vollständig in die ELWIS–DB abgebildet und geladen werden.

2.2 Die Lexikalische Datenbank

Derzeit umfaßt die lesikalische Datenbasis drei Datenbanken gleicher Struktur aber mit unterschiedlichem Inhalt. Dies sind zum einen die zwei Datenbanken, in die die Computerlexika CELEX und SADAW überführt wurden und zum anderen eine Datenbank, in der exemplarische Einträge für Nomina, Verben und Adjektive erfaßt wurden, die ein möglich großes Spektrum der Flexionsmorphologie und Wortbildungsregularitäten des Deutschen abdecken. Mit Hilfe dieser 'Testdaten' wurde ein Verfahren (LEXPAND) entwickelt, das aus den Grundformen und ihren zugehörigen Flexionsparadigmen die entsprechenden Vollformen generiert und ihre morphosyntaktischen Merkmale (Wortart, Genus, Numerus, Kasus) angibt.

Die Trennung dieser Datenbanken erfolgte aus praktischen Gründen, nämlich um ein Testen der einzelnen Verfahren zu erleichtern. Diese drei Datenbanken, die aus verschiedenen lexikalischen Quellen gewonnen wurden bzw. manuell erstellt wurden, sollen jedoch in einer deskriptiven Datenbank, der ELWIS–DB, zusammengeführt werden, die wiederum die Basis für weitere Repräsentationsformen bildet.

Aus pragmatischen Gründen haben wir uns entschlossen, die ELWIS-DB zunächst in Form einer relationalen Datenbank mit dem DBMS EMPRESS zu realisieren. Zu diesen Gründen gehören:

- Vorhandenes Know-how.
- Verfügbarkeit geeigneter DBMS-Software.

- Ausgereiftheit und Zuverlässigkeit der Software.
- Relative Flexibilität beim Datenbankdesign.

Diesen Vorteilen eines relationalen DBMS stehen einige für die Modellierung lexikalischer Wissensbasen nicht unerheblichen Nachteile gegenüber:

- Keine rekursiven Datenstrukturen.
- Keine Unterstützung für Vererbungsmodelle.

Datenbankstruktur

Die Struktur der ELWIS-Datenbank spiegelt Makro- und Mikroebenen wieder, wie sie sich beispielsweise in Wörterbüchern finden lassen: Entitäten wie "Lemma", "Flexion", "Bedeutung" finden sich in der Datenbank in Form von Tabellen wieder. Aus der Analyse der strukturellen Beziehungen solcher Einheiten läßt sich ein sogenanntes *entity relationship model* erstellen, das dann wiederum als Basis für die Modellierung von Tabellen in einem relationalen Datenbanksystem dient. Die Struktur der ELWIS-Datenbank ist in Abbildung 2 in der Form eines vereinfachten ER–Modells wiedergegeben.

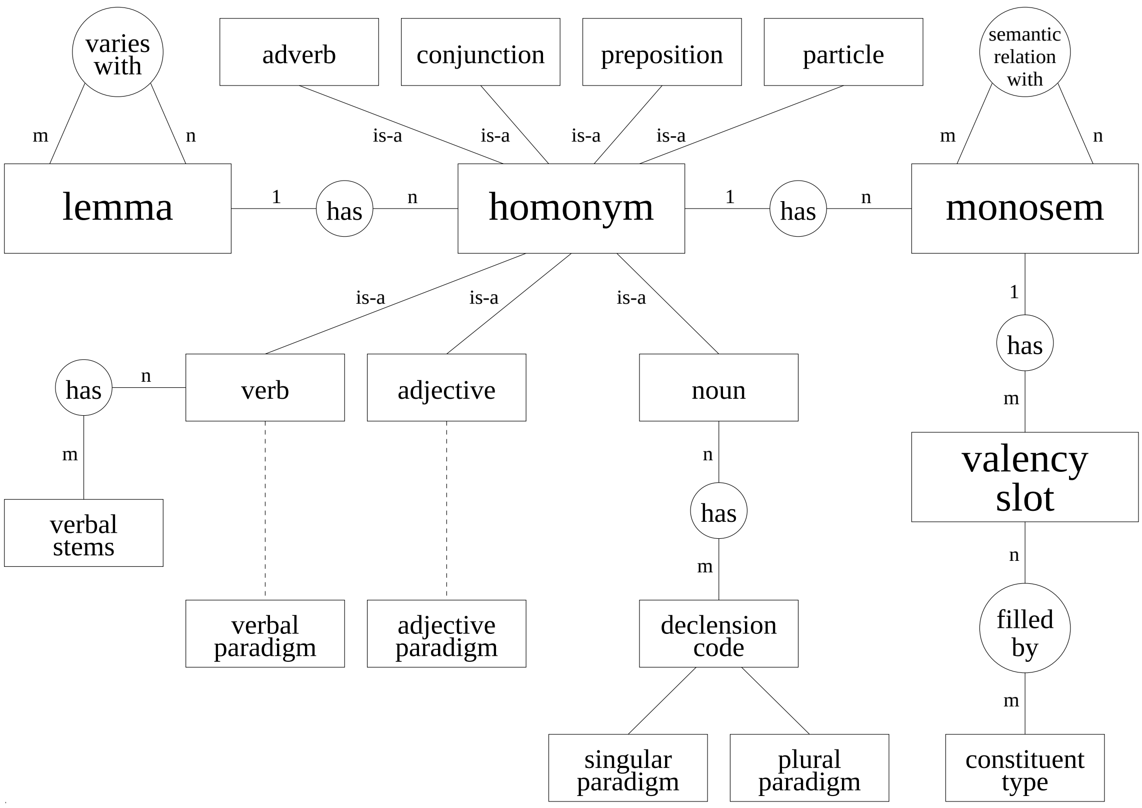

Abbildung 2: Konzeptuelles Datenbankschema

Benutzerschnittstelle

Um die Daten in der Datenbank zu sichten und gegebenenfalls neue Daten einzufügen oder bestehende Daten zu ändern, wurde eine Benutzerschnittstelle entwickelt. Damit wird der einfache Zugriff auf die verschiedenen Tabellen über Menüs ermöglicht, ohne daß dabei Kenntnisse einer Datenbankabfragesprache wie SQL erforderlich sind. Abbildung 3 zeigt die Bildschirmmaske zur Tabelle Nomina mit dem ausgewählten Lemma 'Band' (*Buchband*, Homonymnr. 3). Hier finden sich Informationen zur Grammatik, wie Genus, Stammänderung und Flexionsendungen.

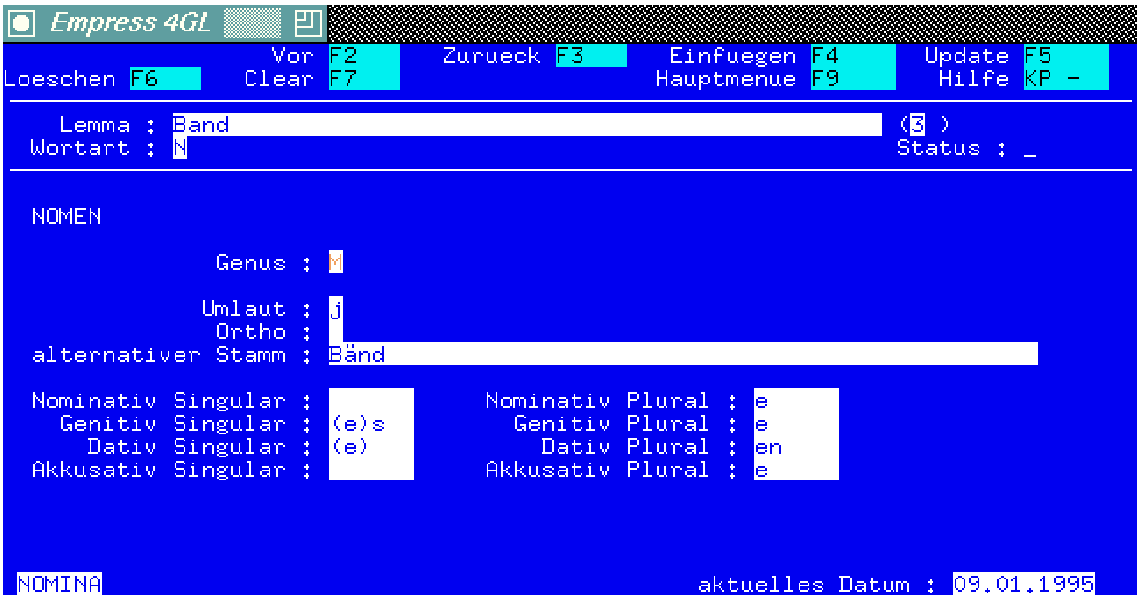

Abbildung 3: Auschnitt aus der Graphikschnittstelle zur ELWIS-Datenbank

2.3 Repräsentation Lexikalischen Wissens

Das in der lexikalischen Datenbank gespeicherte Wissen wird durch sogenannte Compilerprogramme in eine lexikalische Wissensbasis übersetzt, d.h. in eine durch eine Wissensrepräsentationssprache formulierte theorieabhängige Beschreibung des jeweiligen lexikalischen Gegenstandbereichs.

Ein erster Prototyp eines solchen Compilerprogramms (INSYST) wurde zur Überführung von lexikalischen Einheiten in den DATR-Repräsentationsformalismus entwickelt und mit mehreren alternativen DATR-Theorien zur deutschen Nominalflexion getestet (vgl. Light et al., 1995).

2.4 Technologie- und Datentransfer

Die im Laufe des Projekts entwickelten Programme und Daten konnten zum Teil auch in anderen Projekten, bzw. Arbeiten eingesetzt werden.

So wurde der Wörterbuchparser LexParse der Forschungsstelle der Heidelberger Akademie der Wissenschaften für die Überführung des Deutschen Rechtswörterbuchs in eine

Datenbank zur Verfügung gestellt. Zur Zeit wird LexParse am Seminar für Sprachwissenschaft in dem vom Rank Xerox Research Centre in Grenoble koordinierten LRE-Projekt *COMPASS: Adapting Bilingual Dictionaries for Online Comprehension Assistance* für das Parsen des Collins Großwörterbuchs (Deutsch-Englisch) eingesetzt.

Der Wortarten–Tagger LIKELY wird ebenfalls im COMPASS–Projekt eingesetzt. Aus dieser Zusammenarbeit ergab sich die Erstellung einer Parametermenge für ein Hidden–Markov–Model (HMM) zur Wortarten–Disambiguierung des Deutschen, das von Rank Xerox für die Entwicklung von kommerziellen Spell- und Grammarcheckern eingesetzt werden soll. Der LIKELY–Tagger und das für Rank Xerox entwickelte HMM verwenden das ELWIS-Tagset (siehe Abschnitt 2.1.1), das auch als Standardisierungsvorschlag in die Ergebnisse des EAGLES-Projekts der EU eingegangen ist. Außerdem wurde das Tagset im Rahmen einer Studienarbeit von Bernd Krengel zur Satzerkennung auf Basis des Bi- bzw. Trigrammodells an der Technischen Universität Dresden verwendet. Weiterhin wurde LIKELY auch am Institut für Deutsche Sprache (IDS) in Mannheim zur Analyse deutscher Textkorpora eingesetzt.

Das Korpus–Tool NGRAM zur Extraktion von Mehrwortlexemen wurde außer in den Arbeiten von Elisabeth Breidt, einer Stipendiatin des Graduiertenkollegs Tübingen, auch an der Universität Münster von Lothar Lemnitzer in seinen Arbeiten zur Extraktion rekurrenter Textmuster aus großen Textkorpora angewendet.

Die im Laufe des Projekts angesammelten, umfangreichen Korpora wurden für verschiedene Arbeiten als empirische Grundlage zur Verfügung gestellt. Tilo Weber (Linguistics Department, University of Colorado at Boulder) wertete die beiden Korpora MCI und das Deutsche Newskorpus für seine Untersuchungen zu "Genusschwankungen im Deutschen" aus. Susanne Riehemann untersuchte vor allem das Mannheimer Korpus hinsichtlich der Bildung von *bar*–Adjektiven und für die Untersuchungen zum Sprachgebrauch in deutschen Newstexten (Feldweg et al., 1995) war das Deutsche Newskorpus die empirische Grundlage.

3 Vergleichbare Arbeiten außerhalb des ELWIS-Projektes

Um einen Überblick über den Stand vergleichbarer Arbeiten im Bereich der Wiederverwendbarkeit lexikalischer Ressourcen und Methoden zur linguistischen Erschließung des Deutschen zu gewinnen, wurde von Mitarbeitern des ELWIS–Projekts am 17. und Februar 1994 ein Arbeitstreffen in Tübingen veranstaltet. Darüberhinaus sollte mit diesem Workshop eine bessere Zusammenarbeit zwischen den verschiedenen Projekten angeregt werden, um unnötige Mehrfach–Entwicklungen gleichartiger Verfahren, wie z.B. Korpus–Tools und Tagging–Verfahren, zu vermeiden. Der Schwerpunkt lag dabei auf Arbeiten zur Textaufbereitung, die zum eine durch Schnittstellen oder Korpus–Tools unterstützt wird oder mittels statistischer Verfahren, wie das Wortarten–Tagging, automatisch durchgeführt wird. Dabei wurden neben der Beschreibung verfügbarer Korpora auch die Probleme einer notwendigen Standardisierung von Annotationsvorschriften angesprochen. Weiterhin wurden Verfahren zur automatischen Wörterbuchanalyse und existierende Lexikonsysteme vorgestellt.

Die vorgestellten Beiträge werden im Frühjahr dieses Jahres in einem Tagungsband veröffentlicht (Feldweg und Hinrichs, 1995).

4 Verhältnis Arbeitsprogramm/Ergebnisse

Im Verlauf des ELWIS–Projekts konnten umfangreiche lexikalische Ressourcen erworben und Verfahren zur Akquisition und Repräsentation des aus diesen Ressourcen gewonnenen lexikalischen Wissens zum Deutschen entwickelt und angewendet werden. Damit wurde eine Infrastrukur gechaffen, die für den Aufbau großer lexikalischer Wissensbasen eine wichige Voraussetzung ist und so die Entwicklung leistungsstarker computerlinguistischer Systeme unterstützt.

Im Verlauf des Projekts stellte sich heraus, daß die zur Verfügung gestellten Ressourcen erst einmal zeitaufwendigen Voruntersuchungen und Vorverarbeitungen unterzogen werden mußten, bevor die eigentlichen Verfahren zur Akquisition angewendet werden konnten. Besonders die Verarbeitung der Computerlexika bereitete größere Probleme als erwartet, da sie in einem schwer zu verarbeitenden Format vorlagen. Ferner mußte festgestellt werden, daß die Daten in den Computerlexika, den Wörterbüchern und im Referenzkorpus in nicht unerheblichem Maße fehlerhaft waren und für eine weitere Verarbeitung ausgeschlossen werden mußten. Im Austausch mit den jeweiligen Lieferanten dieser Ressourcen hätten diese Fehler zum beiderseitigen Nutzen behoben werden können, doch z.B. auf Seiten des DUDEN–Verlages bestand dazu kein Interesse. So entstanden die in Abschnitt 2.2 beschriebenen voneinander getrennten Datenbanken und wurden bis zum jetzigen Zeitpunkt noch nicht in einer Datenbank zusammengeführt, da die Fehler in den Daten erst einmal beseitigt werden müssen. Außerdem wurden die geparsten Satzbänder der Wörterbücher bisher noch nicht in die Datenbank überführt, da auch sie einen beträchtlichen Anteil von Fehlern aufweisen.

Aufgrund dieser erschwerten Ausgangssituation konzentrierte sich die Arbeit in erster Linie auf die Akquisition syntaktischen Wissens, während die Verarbeitung von semantischen Informationen aus Textkorpora und Wörterbüchern bisher außen vor gelassen wurde. Um bei der Extraktion von semantischer Information aus Textkorpora zu einer befriedigenden Lösung kommen zu können, wäre es notwendig gewesen, zusätzlich zu einem Wortarten–Tagging wenigstens ein NP-Tagging oder Chunk–Parsing (z.B. Abney, 1991) durchzuführen, aber dies hätte den Rahmen des Projektes bei weitem überstiegen.

5 Veröffentlichungen, Arbeitspapiere, Aktivitäten

5.1 Veröffentlichungen

References

Breidt, E. (1993a). Extraction of V–N–collocations from text–corpora: A feasibility study for german. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. 22.6.1993, Ohio State Univesity, Columbo.*

Feldweg, H. (1995). Stochastische Wortartendisambiguierung des Deutschen. In *Lexikon & Text*, pages 241–254, Tübingen. Niemeyer.

Feldweg, H. und Hinrichs, E., editors (1995). *Lexikon & Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen. Vorträge des vom*

Seminar für Sprachwissenschaft der Universität Tübingen am 17. und 18. Februar auf Schloß Hohentübingen durchgeführten gleichnamigen Workshops, Tübingen. Niemeyer.

Gerdemann, D. und Hinrichs, E. (in Druck). Some Open Problems in Head-driven Generation. In *Linguistics and Computation*. CSLI Lecture Notes, CLSI, Stanford, California, USA.

Hauser, R. und Storrer, A. (1994). Dictionary entry parsing using the LexParse system. In *Lexicographica*, 9.

Hauser, R. und Storrer, A. (1995). Probleme und Lösungen beim Parsen von Wörterbüchern. In *Lexikon & Text*, pages 61–77, Tübingen. Niemeyer.

Hinrichs, E. und Nakazawa, T. (in Druck). Linearization AUXs in German Verbal Complexes. In J. Nerbonne, K. Netter, C. Pollard (eds.), *German Grammar in HPSG*. CSLI Lecture Notes, CSLI, Stanford, California.

Light, M., Reinhard, S., und Boyle-Hinrichs, M. (1995). INSYST: An automatic inserter system for hierarchical lexica. In *Lexikon & Text*, pages 111–119, Tübingen. Niemeyer.

Storrer, A., Feldweg, H., und Hinrichs, E. (1993). Korpusunterstützte Entwicklung lexikalischer Wissensbasen. *Sprache und Datenverarbeitung*, 17(2):59–72.

Thielen, C. (1994). Ein Tagset für die Wortartenklassifizierung des Deutschen. In Trost, H., editor, *KONVENS '94*. Österreichische Gesellschaft für Artificial Intelligence, WIEN.

Thielen, C. und Schiller, A. (1995). Ein kleines und erweitertes Tagset fürs Deutsche. In *Lexikon & Text*, pages 215–226. Niemeyer, Tübingen.

5.2 Arbeitsberichte

References

Breidt, E. (1993b). Extraktion von Verb–Nomen–Verbindungen aus dem Mannheimer Korpus I. SFS–REport 03–93, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Engelke, S. (1994). Grammatikentwicklung für Wörterbücher mit dem LexParse Formalismus. Arbeitsbericht, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen (in Vorbereitung).

Feldweg, H. (1993). Stochastische Wortartendisambiruierung für das Deutsche: Untersuchungen mit dem robusten System LIKELY. SFS–Report 08–93, Universität Tübingen, Seminar fr Sprachwissenschaft, Tübingen.

Feldweg, H. und Hipp, M. (1993). NGram: Ein Werkzeug zur statistischen Analyse von Wortfolgen. Unveröffentlichte Programmanleitung. Universität Tübingen, Seminar für Sprachwissenschaft.

Feldweg, H., Storrer, A., und Thielen, C. (1993). ELWIS–DB: Modellierung lexikalischer Einheiten des Deutschen in einem relationalen Datenbanksystem. Universität Tübingen, Seminar für Sprachwissenschaft. Unveröffentlichtes Arbeitspapier.

Hauser, R. (1993). LexParse — User's manual. SFS–Report 10–93, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Hauser, R. und Storrer, A. (1993). LexParse — Ein Parser zur maschinellen Analyse von Wörterbuchstrukturen. SFS–Report 09–93, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Hinrichs, E. (1991). Korpusunterstützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik. Forschungsantrag im Rahmen des Forschungsschwerpunktprogramms. Unveröffentlichter Projektantrag.

Hinrichs, E. und Nakazawa, T. (1993). Aspects of German VP Structure: An HPSG Account. SFS–Report–01–93. Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Hinrichs, E. und Nakazawa, T. (1994). Partial–VP and Split–NP Topicalization in German: An HPSG Analysis. In E. Hinrichs, D. Meurers, T. Nakazawa, *Partial–VP and Split–NP Topicalization in German: An HPSG Analysis and its Implementation.* SFB 340 Report, IBM Deutschland GmbH, Heidelberg.

Thielen, C. und Sailer, M. (1994). *Ein Tagset fürs Deutsche. Richtlinien für die manuelle Annotierung von Textkorpora.* Unveröffentlichtes Arbeitspapier. Universität Tübingen, Seminar für Sprachwissenschaft.

5.3 Vorträge und andere Aktivitäten

Helmut Feldweg

- 10/1992 Workshop Repräsentatives Korpus der Deutschen Gegenwartssprache", Universität Bonn: "Korpusunterstützte Entwicklung lexikalischer Wissensbasen".
- 10/1993 Workshop zur automatischen Wortformerkennung der Gesellschaft für linguistische Datenverarbeitung, Universität Erlangen: "Korpusunterstützte Entwicklung lexikalischer Wissensbasen – Erste Ergebnisse".
- 2/1994 Arbeitstreffen "Lexikon & Text", Tübingen: "Stochastische Wortartendisambiguierung des Deutschen"
- 5/1994 Kolloquium Lexikontheorie und Computerlexikographie, Universität Leipzig: "Korpusunterstützte Entwicklung lexikalischer Wissensbasen".

Helmut Feldweg & Angelika Storrer

- 2/1994 Treffen des AK Korpora der GLDV, Mannheim: "ELWIS: Korpusunterstützte Entwicklung lexikalischer Wissensbasen"
- 8/1994 6th EURALEX International Congress on Lexicography, Amsterdam: "Corpus-based development of lexical knowledge bases: the ELWIS project" (Poster)

Ralf Hauser

2/1993 Universität Heidelberg, Kolloquium zur Lexikographie: "Übersicht über das ELWIS-Projekt und die elektronische Aufbereitung von Wörterbüchern"

4/1993 EAMT-Workshop "Machine Translation and the Lexicon", Wissenschaftliches Zentrum der IBM Deutschland, Heidelberg: "LexParse – A Dictionary Entry Parser"

5/1994 Kolloquium zur Computerlexikographie, Universität Leipzig: "Das Parsen von Wörterbüchern"

12/1994 Wissenschaftliches Zentrum der IBM Deutschland, Heidelberg: "Das LexParse-System"

Ralf Hauser & Angelika Storrer

4/1994 Consensus ex Macchina? (Association for Literary and Linguistic Computing, Association for Computers in the Humanities), Paris: "Automatic Recognition of Textual Structures of Dictionaries"

Bop Carpenter & Erhard W. Hinrichs

8/1992 5th European Summer School in Logic, Language and Information, Lissabon, Portugal: "Implementations of Attribute-Value Logics for Grammar Formalisms" (Workshop)

Dale Gerdemann & Erhard W. Hinrichs

8/1992 4th European Summer School in Logic, Language and Information, Colchester, England: "Natural Language Generation" (Kurs)

Erhard W. Hinrichs

10/1993 Linguistisches Kolloquium des Wissenschaftlichen Zentrums der IBM Deutschland GmbH, Heidelberg: "NP Aufspaltung: eine HPSG Analyse"

1/1994 NTT Basic Research Laboratory, Tokyo, Japan: "Quantifier Scope Ambiguities"

1/1994 NTT Yokosuka Research and Development Center, Yokosuka, Japan: "Implementation of HPSG Grammars for German in the Troll System"

5/1994 Kyoto Workshop for German–Japanese Machine Translation, Kyoto, Japan: "HPSG for German"

Erhard W. Hinrichs & Paul J. King

10/1993 'Between Syntax, Semantics and Logic', Stuttgart: "What Feature Structures Do" (Workshop)

Erhard Hinrichs & Tsuneko Nakazawa

7/1993 1st International Workshop on HPSG, Columbus, Ohio, USA: "Topicalization in German: An HPSG Analysis"

8/1993 'Categorial Grammar' Workshop at the 5th European Summer School in Logic, Language and Information, Lissabon, Portugal: "Split NP Topicalization: A Challenge for Categorial Grammar"

1/1994 Annual Meeting of the Linguistic Society of America, Boston, Massachusetts, USA: "Split NP Topicalization in German: An HPSG Analysis"

Erhard W. Hinrichs & Heike Winhart

1/1995 Kolloquium 'Theorie der Semantik und Theorie der Lexikographie – Angewandte Semantik und Praxis der Lexikographie', IKP, Universität Bonn: "Praktische Semantik in der Computerlinguistik – ein Erfahrungsbericht"

Tsuneko Nakazawa & Erhard W. Hinrichs

8/1994 6th European Summer School in Logic, Language and Information, Kopenhagen, Dänemark: "Grammar Formalisms for NLP" (Workshop)

Sabine Reinhard, Marie Boyle-Hinrichs, Marc Light

4/1993 Sixth Conference of the European Chapter of the Association for Computational Linguistics, Universität Utrecht: "INSYST: An Automatic Inserter System for Hierarchical Lexica"

2/1994 Arbeitstreffen "Lexikon & Text", Tübingen: "INSYST: An Automatic Inserter System for Hierarchical Lexica"

Angelika Storrer

6/1994 Kolloquium zur Computerlexikographie, Universität Leipzig: "Mehrwotlexeme in der maschinellen Übersetzung"

Angelika Storrer & Ralf Hauser

2/1994 Arbeitstreffen "Lexikon & Text", Tübingen: "LexParse: Ein Parser zur maschinellen Analyse von Wörterbuchstrukturen"

Anglika Storrer & Ralf Hauser & Sabine Engelke

2/1993 Heidelberger Kolloquium zur Lexikographie und Lexikologie, Universität Heidelberg: "Artikelstrukturgrammatiken und Wörterbuchparsing"

Angelika Storrer & Ulrike Schwall

4/1993 EAMT-Workshop "Machine Translation and the Lexicon", Wissenschaftliches Zentrum der IBM Deutschland, Heidelberg: "Description and Acquisition of Multiword Lexemes"

Christine Thielen & Anne Schiller

2/1994 Arbeitstreffen "Lexikon & Text", Tübingen: "Ein kleines und erweitertes Tagset fürs Deutsche"

Christine Thielen

9/1994 KONVENS 94 "Verarbeitung natürlicher Sprache", Wien: "Ein Tagset für die Wortartenklassifizierung des Deutschen" (Poster)

6 Zitierte Literatur

References

Abney, S. (1991) Parsing by chunks. In: Berwick, R., Abney, S. und Tenny, C., Hrsg., *Principle–based parsing.* Kluwer, Dordrecht.

Alshavi, H. (1989). Analysing the dictionary definitions. In Boguraev, B. und Briscoe, T., Hrsg., *Computational Lexicography for Natural Language Processing*, pp. 153–169. London und New York.

Anick, P. und Pustejovski, J. (1990). Knowledge acquisition from corpora. In *Proceedings of COLING 13th, Helsinki.*

Bläsi, C. und Koch, H.-D. (1992). Dictionary entry parsing using standard methods. In *Proceedings of COMPLEX 92, Budapest.*

Breidt, E. (1993b). Extraktion von Verb–Nomen–Verbindungen aus dem Mannheimer Korpus I. SFS–Report 03–93, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Brustkern, J. (1988). Computerunterstützter Aufbau maschinenlesbarer Lexika. Eine Untersuchung zum Import lexikalischer Information aus traditionellen Lexika am Beispiel des dtv-Wörterbuchs der deutschen Sprache von Gerhard Wahrig. *Sprache und Datenverarbeitung*, 1:5–20.

Byrd, R. J., Chodorow, M. S. und et al. (1987). Tools and methods for computational lexikology. *Journal of Computational Linguistics*, 13(3–4):219–240.

Calzolari, N. und McNaught, J. (1994). EAGLES Interim Report. Editor's Introduction. (= EAG-EB-IR-2), Version of October, 1994.

Christ, O. (1994a). The IMS Corpus Workbench – The XKWIC User Manual. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Christ, O. (1994b). The IMS Corpus Workbench – Corpus Administrator's Manual. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Church, K., Gale, W., Hanks, P. und Hindle, D. (1991). Using statistics in lexical analysis. In Zernik, Hrsg., *Lexical Acquisition: Exploiting On–Line Resources to Build a Lexicon*, pp. 115–164. Lawrence Erlbaum, Hillsdale, NJ

Church, K. und Hanks, P. (1990). Word association norms, mutual information and lexicography. In *Proceedings of the 28th ACL*, pp. 76–83.

Child Language Analysis (CLAN) (1994). In: The CHILDES Database, September 1994, CD-ROM.

Cloeren, J. (1993). *Towards a cross–linguistic tagset.* TOSCA Research Group for Corpus Linguistics, Dept. of Language & Speech, University of Nijmegen. Unpublished document.

Collins German–English, English–German dictionary (1991) / by Peter Terell, Veronika Schnorr, Wendy V. A. Morris and Roland Breitsprecher. Second Edition, Reprint 1993, Glasgow: HarperCollins Publishers.

Association for Computational Linguistics (Hrsg.) (1994). European Corpus Initiative: Multilingual Corpus 1. CD-ROM, HCRC University of Edinbrugh and ISSCO University of Geneva.

Ejerhed, E., Källgren, G., Wennstedt, O. und Åström, M. (1992). *The Linguistic Annotation System of the Stockholm–Umeå Corpus Project.* Dept. of General Linguistics, University of Umeå, Report no. 33, (= DGL-UUM-R-33), Umeå.

Ejerhed, E. (1995). Design Principles for a Swedish corpus annotation system. In: Feldweg, H. und Hinrichs, E., editors. *Lexikon & Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen. Vorträge des vom Seminar für Sprachwissenschaft der Universität Tübingen am 17. und 18. Februar auf Schloß Hohentübingen durchgeführten gleichnamigen Workshops*, Tübingen. Niemeyer.

Feldweg, H. (1993). Transcription, Storage and Retrieval of Data. In: Adult language acquisition: cross-linguistic perspectives. Volume I: Field methods. Edited by Clive Perdue. Cambridge: Cambridge University Press, pp. 108-130.

Feldweg, H. (1993). Stochastische Wortartendisambiguierung für das Deutsche: Untersuchungen mit dem robusten System LIKELY. SFS–Report 08–93, Universität Tübingen, Seminar fr̈ Sprachwissenschaft, Tübingen.

Feldweg, H. und Hipp, M. (1993). Ngram: Ein Werkzeug zur statistischen Analyse von Wortfolgen. Anleitung zur Benutzung des gleichnamigen Programmes.

Feldweg, H. und Hinrichs, E., editors (1995). *Lexikon & Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen. Vorträge des vom Seminar für Sprachwissenschaft der Universität Tübingen am 17. und 18. Februar auf Schloß Hohentübingen durchgeführten gleichnamigen Workshops*, Tübingen. Niemeyer.

Feldweg, H., Kibiger, R. und Thielen, C. (1995) Zum Sprachgebrauch in deutschen Newsgruppen. In: Osnabrücker Beiträge zur Sprachtheorie (OBST): *Neue Medien und Gegenwartssprache* (erscheint demnächst).

Hauser, R. (1993). LexParse — User's manual. Software Manual. SFS–Report 10–93, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Hauser, R. und Storrer, A. (1993). LexParse — Ein Parser zur maschinellen Analyse von Wörterbuchstrukturen. SFS–Report 09–93, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen.

Heid, U. (1991). *Towards reusable lexical resources for natural language processing. Some proposals for linguistic knowledge representation.* Stuttgart.

Light, M., Reinhard, S., und Boyle-Hinrichs, M. (1995). INSYST: An automatic inserter system for hierarchical lexica. In *Lexikon & Text*, pages 111–119, Tübingen. Niemeyer.

MacWhinney, B. (1995). The CHILDES Project: Computational Tools for Analyzing Talk. Hillsdale: Lawrence Erlbaum.

Marcus, M. P., Santorini, B. und Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. In: Computational Linguistics, Volume 19, Number 2, pp. 313–330.

Monacchini, M. und Calzolari, N., Coords. (1994). *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages.* (Draft Version: October, 1994) Istituto di Linguistica Computazionale, ILC, Pisa.

Montemagni, S. und Vanderwende, L. (1992). Structural patterns vs. string patterns for extracting semantic information from dictionaries. In *Proceedings COLING 92, Nantes*, pp. 553–560.

Neff, M. und Boguraev, B. (1990). From machine-readable dictionaries to lexical data bases. Report, IBM T.J. Watson Research Center, New York, New York.

Neff, M., Byrd, R. und Rizk, O. (1988). Creating and querying hierarchical lexical databases. In *Proceedings of the Second Conference on Applied Natuaral Language Processing*, pp. 84–93.

Neumann, G. (1992). diac – a suite of lex programs programs for converting Ascii German into diacritical German. Version 1.3. Wirtschaftsuniversitaet Wien.

Prinz, J. (1993). cgrep, Version 1.4. 1993/03/10. Institut für instrumentelle Mathematik und Informatik Wesfälische-Wilhelms-Universität.

Schulze, B. M. und Christ, O. (1994). The IMS Corpus Workbench – The CQP User's Manual. Stuttgart: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Smadja, F. (1991). Macrocoding the lexicon with co–occurrence knowledge. In Zernik, U., Hrsg., *Lexical acquisition: exploiting on–line resources to build a lexicon*, pp. 165–189. Lawrence Erlbaum Associates, Hillsdale, NJ.

Sonderforschungsbereich 100 "Elektronische Sprachforschung", P.A., Hrsg. (1980). *SALEM. Ein Verfahren zur automatischen Lemmatisierung deutscher Texte.* Max Niemeyer Verlag, Tübingen.

Sperberg-McQueen, C. M. und Burnard, L., Hrsg. (1990, 1992, 1993, 1994). *Guidelines for the Encoding and Interchange (TEI P3).* Association for Computational Linguisitcs, Associaton for Computers in the Humanities, Association for Literary and Linguistic Computing, Chicago und Oxford.

TEI AI1W2 (1991). List of Common Morphological Features For Inclusion in TEI Starter Set Of Grammatical-Annotation Tags. (Draft Version: June 11, 1991).

Thielen, C. und Sailer, M. (1994). *Ein Tagset fürs Deutsche. Richtlinien für die manuelle Annotierung von Textkorpora.* Unveröffentlichtes Arbeitspapier. Universität Tübingen, Seminar für Sprachwissenschaft.

Tuthill, W. (1981). Hum – A Concordance and Text Analysis Package. Berkeley, California: Comparative Literature Development, University of California.

Veronis, J. und Ide, N. (1991). An assessement of semantic information automatically extracted from machine readable dictionaries. In *Proceedings of the Fifth Conference of the European Chapter of the ACL, 9-11 April 1991*, pp. 227–232, Berlin Alexanderplatz.

Wiegand, H. (1989a). Component parts and structures of general monolingual dictionaries: A survey. In Hausmann und et al., Hrsg., *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, Band 1, pp. 328–360. Mouton de Gruyter, Berlin und New York.

Wiegand, H. (1989b). Der Begriff der Mikrostruktur im allgemeinen einsprachigen Wörterbuch: Geschichte, Probleme, Perspektiven. In Hausmann und et al., Hrsg., *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, Band 1, pp. 409–461. Mouton de Gruyter, Berlin und New York.

Wiegand, H. (1989c). Formen von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In Hausmann und et al., Hrsg., *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, Band 1, pp. 462–501. Mouton de Gruyter, Berlin und New York.

Wiegand, H. (1990). Printed dictionaries and their parts as text. *Lexicographica*, 6:1–126.

Wothke, K., Weck–Ulm, I., Heinecke, J., Mertineit, O. und Pachunke, T. (1993). Statistically Based Automatic Tagging of German Text Corpora with Parts–of–Speech — Some Experiments. Heidelberg, Wissenschaftliches Zentrum der IBM Heidelberg.

Zimmermann, M. (1987). transportable version of the browser program – 871008. u.a. verfuegbar via: ftp.switch.ch:/mirror/vms/DECUS/decuslib/browser/.