AUTOMATISCHE WORTFORMERKENNUNG IM DEUTSCHEN ANALYSEVERFAHREN UND SYSTEME
Ein Bericht über die auf dem 1. GLDV-Workshop zur automatischen Wortformerkennung in Erlangen präsentierten Entwicklungen

Uta Seewald

Universität Hannover

1 Der organisatorische Rahmen

Als Leiter des neu ins Leben gerufenen GLDV-Arbeitskreises "Parsing in Morphologie und Syntax" sowie als Initiator einer geplanten neuartigen Veranstaltungsreihe, MORPHOLYMPICS, hatte Roland Hausser, Leiter der Abteilung Computerlinguistik der Universität Erlangen- Nürnberg, für den 14. und 15. Oktober 1993 nach Erlangen eingeladen.. Die Initiative war aus der Beobachtung hervorgegangen, daß Programme zur automatischen Wortformerkennung sowohl in der Grundlagenforschung der Computerlinguistik als auch für praktische Anwendungen in der Textverarbeitung benötigt werden, eine Konsolidierung der in den vergangenen 20 Jahren entwickelten Teilergebnisse bisher jedoch nicht stattgefunden hat, nicht zuletzt auch aufgrund der schnellen Entwicklung auf dem Gebiet der Computerhard- und software sowie der linguistischen Theoriebildung.

Ziel der Veranstaltung war, die derzeit für das Deutsche in Entwicklung bzw. im Einsatz befindlichen Systeme zur automatischen Wortformerkennung vorzustellen und in einem aktuellen Überblick zu vergleichen. Ein besonderes Anliegen dieses Workshops lag auch darin, Bewertungskriterien für Systeme zur automatischen Wortformerkennung zu entwickeln. Zur Erarbeitung dieser Kriterien fand aus diesem Grunde eine gesonderte Diskussionsveranstaltung statt, deren Ergebnisse in einem Kriterienkatalog zur Bewertung automatischer Wortformerkennungssysteme gipfelte. Diese Kriterien sollen schließlich als Bewertungsmaßstab an jene Systeme angelegt werden, die sich dem Wettbewerb auf der 1. MORPHOLYMPICS stellen. Auf dieser für den Beginn des kommenden Jahres anberaumten Veranstaltung sollen verschiedene Systeme zur automatischen Wortformerkennung vorgestellt und *online* getestet werden.

2 Die vorgestellten Analyseverfahren und Systeme

Die zehn auf dem Workshop vorgestellten Systeme sind zum Teil im Hinblick auf bestimmte Anwendungsaspekte konzipiert worden, so daß ein unmittelbarer Vergleich der Systeme sowie der realisierten Analyseverfahren sich als schwierig oder zum Teil gar unmöglich erweist. Neben Systemen zur morphologischen Analyse und zum morphologischen Tagging finden sich denn auch Ansätze zur Inhaltserschließung von Lem-

mata sowie Hilfsmittel zur Verarbeitung von Nicht-ASCII-Zeichen.

Da die auf dem Workshop präsentierten Systeme alle *online* demonstriert wurden, war es den Teilnehmern dennoch möglich, bei Systemen mit ähnlicher Zielsetzung vergleichende Bewertungen anhand spontan zusammengetragener Testdaten vorzunehmen.

2.1 Word Manager (Mare Domenig, Universität Basel)

Der von Marc Domenig präsentierte Word Manager, der an der Universität Basel entwickelt wurde, ist ein System zur Erstellung und Nutzung morphologischer Wörterbücher. Das System beruht auf einer Netzwerkarchitektur, so daß verschiedene Benutzer bzw. natürlichsprachige Systeme auf ein Wörterbuch zugreifen können. Der Word Manager ist in der Lage, mehrere Sprachen zu bearbeiten und eine Vielzahl von Wörterbüchern zu verwalten. Die Idee, die diesem System zugrunde liegt, ist, - ein zentrales System bereitzustellen, daß das gesamte morphologische Wissen enthält, so daß verschiedene Anwendungen darüber verfügen können, ohne selbst mit einer morphologischen Komponente ausgestattet zu sein.

Ausgangspunkt der Konzeption des Word Manager ist das Wörterbuch. Diese Perspektive hat zur Entwicklung eines lexikalischen Datenbanksystems geführt, das im Vergleich zu verschiedenen regelbasierten Systemen den Vorzug hat, problemlos (von einem Linguisten) um große Datenmengen erweitert zu werden und für den menschlichen Benutzer doch transparent zu bleiben.

Nach der Vorstellung der Entwickler des Word Manager wird eine leere Datenbank zunächst von einem Linguisten bearbeitet, der Regeln über Flexion und Wortbildung einer bestimmten Sprache und entsprechende Beispiele formuliert sowie Ausnahmen von den aufgeführten Regeln angibt.

Die morphologische Regeldatenbank soll anschließend von einem Lexikographen, der lexikalische Einträge den formulierten Regeln zuordnet, weiter bearbeitet werden. Dabei kommt dem Lexikographen die Aufgabe zu, eine explizite Analyse einer Wortform in unmittelbare Konstituenten anzugeben und die bei jedem Schritt angewendete morphologische Regel zu nennen.

Die Sicht des Lexikographen auf das System wird durch ein Interface, das als CMKS (Conceptual Morphological Knowledge Specification environment) bezeichnet wird, geleistet. Bestandteil dieses Interface sind zahlreiche Fenster unterschiedlicher Funktionalität, auf die verschiedene Arten von Informationen verteilt sind. Zu diesen Fenstern gehören das Flexions- und das Wortbildungsfenster. Das Flexionsfenster zeigt Regeln und Formative an, die an der Flexion beteiligt sind, während das Wortbildungsfenster die entsprechenden an der Wortbildung beteiligten Regeln und Flexive visualisiert. Sowohl das Flexionsals auch das Wortbildungsfenster sind sogenannte *tree windows.* Bei der Implementierung von Wortbildungsmechanismen im Deutschen sind beispielsweise verzweigende Wortbildungsregeln erforderlich, die als Baumstruktur in einem Wortbildungsfenster angegeben werden.

Nachdem alle Regeln, Formative und Einträge einer Sprache angegeben sind, kann die betreffende Datenbank kompiliert werden. Stößt das Programm beim Kompilieren auf einen syntaktischen oder semantischen Fehler, wird eine Nachricht an das *message window* geschickt. Nach erfolgreicher Kompilierung stehen dem Benutzer zahlreiche *Browser* zur Verfügung. Zum Test bzw. zur Analyse einzelner Lexeme eignet sich vor allem der Lexembrowser, der ein Wort in seiner Zitierform zusammen mit kategorialen Angaben und möglichen davon abgeleiteten Wortformen ausgibt. Darüber hinaus enthält der Lexembrowser ein Teilfenster, das die am betreffenden Wort beteiligten Formative und deren syntaktische

Spezifikation auflistet. Spezifische morphologische Regeln, die sowohl bei der Wortbildung als auch bei der Flexion zum Tragen kommen, wie z.B. die Umlautung im Deutschen, werden ebenfalls in dafür vorgesehenen Fenstern dargestellt.

Durch die ausgebaute Fenster- und Browsertechnik sowie die menugesteuerten Funktionsaufrufe stellt der Word Manager dem Linguisten in einer offenen Client/Server-Architektur eine komfortable interaktive Entwicklungsumgebung zur Verfügung, die es erlaubt, große morphologische Wörterbücher effizient zu erstellen und zu verwalten.

2.2 Morphology Aid (Henriette Visser /Friederike Benjes, Universität Heidelberg)

Die von Henriette Visser und Friederike Benjes vorgestellte Morphology-Aid ist ein im Rahmen des Translater's Workbench-Projektes von Peter Hellwig entwickeltes Modul morphologischer Funktionen, das etwa von einem Texteditor aus aufgerufen werden kann, um morphologische Angaben einer Wortform abrufen zu können oder aber aufgrund der syntaktischen Angaben die korrekte Wortform eines Lemmas ermitteln zu können, um sie anschließend in den zu erstellenden Text zu übernehmen.

Der in der Morphology-Aid realisierte morphologische Ansatz geht auf das in den siebziger Jahren an der Universität Heidelberg entwickelte System PLAIN zurück. Das System greift auf zwei Lexika zu, wobei die morphologischen Phänomene einer Sprache in einem sogenannten morphosyntaktischen Lexikon beschrieben sind. Ein zweites Lexikon, als Valenzlexikon bezeichnet, enthält die Subkategorisierungsinformation, d.h. Angaben zur syntaktischen Kombinierbarkeit eines Lemmas. Das morphosyntaktische Lexikon ist als Übergangsnetzwerk aufgebaut und setzt sich aus mehreren Unternetzen zusammen. So existieren Unternetze für Stämme, solche für Derivationsmorpheme und andere für Flexionsendungen. Die morphosyntaktische Kombinierbarkeit der morphologischen Elemente wird durch die Übergänge im Netzwerk bzw. die Verbindungen zwischen den einzelnen Unternetzwerken beschrieben.

Die Eingabe eines Lexems in das morphosyntaktische Lexikon erfolgt anhand repräsentativer Wortformen des Lexems, in denen die möglichen Stämme des betreffenden Lexems enthalten sind. Im Fall des Verbs "sprechen" geschieht das anhand folgender Formen: *sprechen, sprichst, sprachst, sprächst, gesprochen,* das *gesprochene.* Anhand bestehender Muster erkennt das System die jeweils in den Wortformen enthaltenen Stämme und übernimmt sie in das Unternetzwerk, das Stämme enthält. Gleichzeitig werden die einzelnen Stämme mit Angaben über Netze bestimmter Flexionsformen verbunden, so daß entlang der autorisierten Netzwerkübergänge alle Formen der jeweils zu einem Lexem gehörenden Stämme erzeugt werden können. Aufgrund der in einem Wort enthaltenen Derivationsmorpheme und Flexionsendungen werden zunächst das Lexem sowie die morphosyntaktischen Angaben (Wortklasse, Kasus, Numerus, Genus, Person etc.) eines zu analysierenden Wortes ermittelt.

Die Anzeige der Flexionsparadigmen eines Lexems wird von sogenannten Tableaus gesteuert. Wird Information über eine bestimmte Wort form abgerufen, wird diese im Lexikon gesucht und die Stämme des zu dieser Wortform gehörenden Lexems ermittelt. Die möglichen Netzwerkübergänge und Verbindungen zu Unternetzen, die die Gesamtheit der Formen eines Lexems beschreiben, werden in einer sogenannten "section-table" gespeichert. Diese hat im Fall des Verbs "sprechen" folgende Form:

```
'sprech' Inf ,Prs1--t
     sprech    ' infen
     'sprech'    prs1-t
     'sprech'    prsK-est
     'sprech'    imp Sg-e
     'sprech'    imp Plt
```

```
     'sprech'      vBw-0
     'sprech' VN-en
          'Sprechen'       nSG-s
     'sprech' VN-er
          'Sprecher'       SMs-On
              'Sprecher'       mSg-s
              'Sprecher'       pI-On
     'sprech' VA-end
          'sprechend'      adjA
          'sprechend'      advO
       'sprich' Prs2Imp
     'sprich' prs2-st-t
     'sprich' impSg-O
'sprach' prtI-st
'spraech' prtK
'sproch' Ptz-ge-en
     'gesproch' ptzen
'sproch' VA-ge-en
     'gesproch' adjP-O
     'gesproch' adjA
```

Das System, mit dem menugesteuert verschiedene Optionen morphologischer Informationsgewinnung aufgerufen werden können, ist weniger für Zwecke der morphologischen Analyse großer Textmengen konzipiert, als für die interaktive Benutzung und den Zugriff von anderen Programmsystemen aus, wie es die Bezeichnung Morphology-Aid bereits nahelegt.

2.3 LA-Morph (Gerald Schüller/Oliver Lorenz, Universität Erlangen)

Das von Gerd Schüller und Oliver Lorenz vorgestellte System zur Wortformerkennung ist die Implementierung eines als LA-Morph bezeichneten Ansatzes zur morphologischen Analyse. LA-Morph basiert auf dem Algorithmus der Linksassoziativen Grammatik, der von R. Hausser (1992) mathematisch ausgewertet und (1989) im Hinblick auf morphologische Anwendungen beschrieben wurde.

Ein LA-Morph-System einer beliebigen Sprache setzt sich aus drei Komponenten zusammen: (1) einem Grundformenlexikon, (2) einer Regelmenge zur Ableitung von Allomorphen *(allo-rules)* aus den Einträgen des Grundformenlexikons und (3) einer Regelmenge, die die Kombination von Allomorphen *(combi-rules)* beschreibt. Eine der wesentlichen Maxime, die bei der Entwicklung von LA-Morph berücksichtigt wurde, ist die Speicherplatzeffizienz und die Verarbeitungsgeschwindigkeit des Systems. Um die Verarbeitungsgeschwindigkeit gering zu halten, werden beispielsweise die Regeln zur Erzeugung der Allomorphe vor der Laufzeit des Analysesystems angewendet, so daß das Lexikon zur Laufzeit bereits alle Allomorphe enthält. Enthält das Grundformenlexikon eines englischen Lexikons beispielsweise den Eintrag ("happy" (ADJ_GRAD) happy), dessen erste Position die Oberflächenform des Eintrags, dessen zweite Position die Kategorie und dessen dritte Position den Stamm bzw. die Semantik des Eintrags angibt, so sorgt eine entsprechende Allomorphieregel für die Erzeugung der zwei von dieser Grundform abgeleiteten Allomorphe ("happy" (ADJ) happy) und ("happi" (SR ADLGRAD) happy).

Allomorphieregeln bestehen aus einer Eingabebedingung und einer oder mehreren Allomorphdefinitionen. Die auf das englische Adjektiv *happy* angewendete Regel überprüft beispielsweise, ob die Eingabekette mit einem 'y' endet und ob das Lemma als (ADJ_GRAD) kategorisiert ist. Trifft das zu, so werden die beiden oben genannten Allomorphe erzeugt. Da bei einer zutreffenden Allomorphieregel die entsprechenden Allomorphe generiert und nachfolgende Allomorphieregeln ignoriert werden, ist die Abfolge der Regeln für die Erzeugung der Allomorphe von Bedeutung. Insbesondere ist wichtig, daß als letzte Regel eine Default-Regel zur Anwendung kommen kann, die solche Lemmata bearbeitet, deren einziges Allomorph mit dem Stamm identisch ist. Neben den aus dem Grundformenlexikon erzeugten Allomorphen enthält das Lexikon auch alle Affixe. Affixallomorphe werden - im Unterschied zu lexikalischen Morphemen - jedoch nicht über Regeln abgeleitet, sondern direkt spezifiziert, da es sich bei diesen um eine geschlossene Menge von Elementen handelt. Die Erzeugung der Allomorphe erfolgt in Abhängig-

keit vom Regularitätsgrad der jeweiligen Grundform. Für das Deutsche werden reguläre *(sagen -> sag),* semi-reguläre *(laecheln -> laechel) laech0,* semi-irreguläre *(hAus -> haus, haeus)* sowie irreguläre Lemmata *(gut -> gut, bess,* be) unterschieden.

Die morphologische Analyse wird zur Laufzeit von den Analyse- bzw. Kombinationsregeln gesteuert, die in editierbarer und in kompilierter Form vorliegen. Die Kombinationsregeln sind als links assoziative Regeln formuliert und enthalten im rechten Regelteil das Ergebnis aus der Konkatenation zweier im linken Regelteil enthaltenen Elemente sowie die Angabe eines Regelpaketes, das jene Regeln auflistet, die auf die nach Anwendung der Regel erzeugte Wortform angewendet werden können. Die erste Regel einer links assoziativen Ableitung beginnt mit einer Startregel, bei der als Startelement die leere Zeichenkette mit einem ersten Allomorph verbunden wird. Die morphosyntaktische Wohlgeformtheit von Wortformen wird mittels kategoriebasierter Restriktionen überprüft. Hierzu gehört beispielsweise die Angabe von Kategorien, die lexikalische Einträge charakterisieren, die unvollständig sind und die Anwendung einer Kombinationsregel erfordern, oder die Angabe von Elementen, die in einer Wortform niemals initial auftreten können.

Das Lexikon liegt bei der Analyse als *trie* vor, ebenfalls eine Maßnahme zur Effizienzsteigerung des Systems. Das deutsche Grundformenwörterbuch umfaßt ca. 13.000 Einträge, das Wörterbuch des englischen Systems ca. 8.000. Um eine Flexibilität des Systems zu gewährleisten, wurde LA-Morph auf verschiedene Plattformen portiert. Da sowohl das Lexikon als auch die Regeln deklarativ formuliert und als ASCII-Dateien gespeichert sind, ist das System praktisch sprachunabhängig.

2.4 Weiterentwicklungen von SADA W (Heinz Dieter Maas, IAI Saarbrücken)

Das von Heinz Dieter Maas vorgestellte Programmpaket *mpro,* das neben umfangreichen Lexikonfunktionen über die Möglichkeit verfügt, Wortformen zu generieren oder ein Texttagging durchzuführen, geht in seinen Anfängen auf das im Rahmen des SFB 100 entwickelte System SADAW zurück.

Das Programmpaket *mpro* analysiert Wörter des Deutschen morphologisch und liefert, sofern das entsprechende Ausgabeformat ('dima') gewählt wird, für Wörter bzw. Morpheme, die im Lexikon semantisch spezifiziert sind, Erklärungen, die den Inhalt der jeweiligen Wortbildung in Form einer Paraphrase darstellen. Neben dieser Option kann als Ausgabeformat der Analyse auch das CAT2-Format gewählt werden! Dabei werden sogenannte slex-Einträge erzeugt, die die Morphosyntax des zu analysierenden Wortes angeben, sowie sogenannte slex-Einträge, die die syntaktisch-semantische Information des betreffenden Wortes enthalten. Bei der Eingabe des Kompositums *Computerfreak* z.B., das in seine zwei nominalen Konstituenten zerlegt wird, enthält *mlex* als Wert des Attributs lex den Eintrag selbst, also 'Computerfreak' , der sich aus den lexikalischen Einheiten (Zu) 'Freak' und 'Computer' zusammensetzt. Die Angaben der syntaktisch-semantischen Information beziehen sich schließlich gesondert auf die beiden Konstituenten des Kompositums, wie das nachfolgende Analysebeispiel zeigt.

Eingabe: *Computerfreak*

```
mlex= {lex='Computerfreak', lu=freak,
       graphiks=eap, known=no, clu=
       {lu='eomputer', head={eat=n, ehead= {agr=
       {gen=masc}}, semf= {abstraet=eoner,
       anim=nil, temp=nil, bound=count,
       gran=nil}}},
```

l CAT2 ist ein maschinelles Übersetzungssytem, das als Seitenlinie von EUROTRA-D am IAI in Saarbrücken entwickelt wurde.

```
        head= {cat=n, deriv=nil, pref=_, ehead=
        {case=nom; dat; acc), agr= {num=sing,
        gen=masc}}}}.[].

slex= {lu='freak', head= {cat=n, ehead= {agr=
        {gen=masc}}, semf= {abstract=concr,
        temp=nil, gran=nil, anim= {'T'=hum,
        hum=male}}}}.[].

slex= {lu='computer', head= {cat=n, ehead= {agr =
        {gen=masc}}, semf= {abstract=concr,
        temp=nil, gran=nil, anim= nil,
        bound=count}}}.[].
```

Das Ausgabeformat mit den Angaben zur morphosemantischen Information des Kompositums enthält weitergehende Angaben, so z.B. daß es sich bei 'Computerfreak' um ein zählbares Substantiv handelt, das entweder im Nominativ, im Dativ oder im Akkusativ vorliegt und als menschliches Agens spezifiziert ist. Als Interpretation des obigen Wortes liefert *mpro* die Angabe "Ein Computerfreak ist jemand, der Computer gerne mag". In den Fällen; in denen das System eine Interpretation des zu analysierenden Wortes geben kann - bei dem Derivat 'Bäcker' lautet sie z.B. "Ein Bäcker ist jemand, der backt" - beruht diese auf einer semantischen Klassifikation der beteiligten Lexeme und Morpheme. So sind die im Lexikon enthaltenen Substantive in verschiedene Klassen untergliedert, wobei jede Klasse gleichsam als Hyperonym der unter sie fallenden Elemente aufgefaßt werden kann. Neben Klassen wie " abstract " , " agent" , "act" und "animal" sind z.B. Klassen für Gefäße ("box"), Krankheiten ("disease"), Instrumente ("instr"), Körperteile ("koerper") oder Materialien („material") vorgesehen. Das Substantiv *Computer* ist z.B. Element der Klasse ("instr"). Im Gegensatz zum Substantiv *Hammer,* das eben falls dieser Klasse angehört, ist *Computer* jedoch als komplexes Instrument markiert. Verben sind danach klassifiziert, welcher semantischen Klasse sie zuzuorden sind bzw. welche semantische Relation zwischen dem abgeleiteten Verb und seiner Basis vorliegt. Als Verb klassen finden sich beispielsweise "mitteilen" mit Verben wie *befehlen, diktieren, erzählen, melden, sagen,* "manipulate" mit Verben wie *ackern, boxen, drücken, reißen, stechen* oder "move_displace" mit Verben wie *fahren, fliegen, rollen* und *ziehen.* Eine Vielzahl von Verbklassen bezieht sich auf den jeweils vorliegenden Ableitungstyp. So gehören Ableitungen von Nomina des Typs *spionieren* (i.e. das tun, was ein Spion tut) zur mit "act_like" bezeichneten Klasse, während Verben, die ihr transitives Objekt mit über anschließen (Bsp.: *beherrschen* [über etw. herrschen], *bejammern* [über etw. jammern], *besiegen* [über etw. siegen], *bestaunen* [über etw. staunen]), als Elemente der Klasse "tra(ueber)" klassifiziert sind.

Bei der Segmentierung mit *mpro* werden alle in einer Wortbildung enthaltenen Stämme identifiziert, wobei auch Allomorphe berücksichtigt werden. Zu allen Teilketten der Zerlegung werden die Lexikoninformationen ermittelt, auf die dann je nach ausgewähltem Ausgabeformat zugegriffen wird. An zwei Beispielen seien die Angaben zum Inhalt der zu analysierenden Wörter abschließend noch einmal dargestellt:

kroatisch
Struktur des Wortes: [derived, a, n, isch, na,
[kroatien, [hyper=loc, loc=country]]]
Interpretation des Wortes:
kroatisch: bezieht sich auf Kroatien
Baecker Struktur des Wortes: [derived, n, v, er, er,
[backen, _31459]]] Interpretation des *Wortes:*
Ein Baecker ist jemand, der backt.

2.5 Korpusunterstützte Entwicklung lexikalischer Wissensbasen (Helmut Feldweg, Universität Tübingen)

Das von Helmut Feldweg beschriebene System zur Entwicklung lexikalischer Wissensbasen ist ein System ~ur automatischen Wortartenzuordnung für deutsche Texte (Taggingsystem), das nicht auf der Grundlage eines morphologischen Analysealgorithmus operiert, sondern auf einem sto-

chastischen Verfahren beruht. Das System LIKELY, um das es sich hier handelt, wurde im Rahmen des Tübinger Projektes ELWIS entwickelt, um umfangreiche annotierte Textkorpora des Deutschen zu erstellen. Das System baut auf einem Algorithmus zur stochastischen Wortartendisambiguierung auf, der von Marshall (1983, 1987) und .de Rose (1988) im Rahmen der Entwicklung von Taggingverfahren für das Englische beschrieben wurde.

Das System LIKELY sucht die einzelnen im Text auftretenden Wörter in einem Vollformenwörterbuch auf, das sowohl Angaben zu möglichen Wortarten des jeweiligen Wortes enthält als auch Angaben zu deren lexikalischen Häufigkeiten. Ist ein Wort in bezug auf seine Wortartenangaben ambig, so wird ein Netzwerk der verschiedenen Lesarten aufgebaut, das links und rechts durch ein eindeutig klassifiziertes Wort begrenzt wird. Die Übergangswahrscheinlichkeiten der aufeinander folgenden Wortarten werden einer Tabelle entnommen, und schließlich wird der durch- das Netzwerk führende optimale Pfad (mittels des Viterbi-Algorithmus) berechnet.

Das System unterscheidet 40 Wortarten, eine Teilmenge jener Wortarten, die im Saarbrücker Lemmatisierungsprogramm SALEM verwendet wurden, auf dessen Erfahrungen bei der Konzeption von LIKELY aufgebaut wurde. Die Wahrscheinlichkeitswerte, auf die bei der Analyse zugegriffen wird, werden einer sogenannten Übergangsmatrix entnommen, die im Falle von LIKELY zweidimensional ist. D.h. die in ihr angegebenen Werte basieren auf den Werten, die aus den absoluten Häufigkeiten der im Referenzkorpus auftretenden Zweierwortgruppen, i.e. Wortbigramme, errechnet wurden. Die Matrix der Übergangswahrscheinlichkeiten sowie ein Vollformenwörterbuch mit der Angabe der Wortarten und der jeweiligen relativen lexikalischen Wahrscheinlichkeiten wurden aus einem Trainingskorpus, das sich aus den ersten zwei Dritteln eines 239889 Wörter umfassenden Referenzkorpus zusammensetzte, ermittelt. Das Referenzkorpus seinerseits basiert auf drei in analysierter Form vorliegenden Segmenten des Mannheimer Korpus I, die mit dem Saarbrücker System SALEM bearbeitet worden waren, und hier einer maschinellen und intellektuellen Überarbeitung unterzogen wurden.

Auf einer Sun Sparcstation 10/20 annotiert das System ca. 6000 Wortformen pro Sekunde. Die ermittelte Fehlerquote von 7,38 % bei unvollständigem Wörterbuch, d.h. einem Wörterbuch, das nur die in einem vorgegebenen Trainingskorpus auftretenden Wörter, deren Wortarten und lexikalische Häufigkeiten berücksichtigt, ergibt sich u. a. aus den verschiedenen zur Analyse herangezogenen Textsorten. Das Ergebnis der Annotierung könnte sowohl durch ein umfangreicheres Wörterbuch als auch eine anschließende morphologische Analyse verbessert werden.

2.6 X2MORF - erweiterte 2-Ebenen-Morphologie (Harald Trost, Universität Wien)

Bei dem von Harald Trost vorgestellten System X2MORF handelt es sich um ein System, das ein erweitertes Zwei-Ebenen-Modell (Two-Level Morphology)2 mit einem merkmalbasierten Lexikon kombiniert. In der ursprünglichen Formulierung des Two-Level-Modells wurden Wortbildungsregeln als reguläre Grammatik.. in Form von Fortsetzungsklassen ausgedrückt. In X2MORF wurden diese Fortsetzungsklassen durch eine merkmalbasierte Wortbildungsgrammatik ersetzt. Darüber hinaus wurden die Zwei-Ebenen-Regeln mit einem Filter versehen, der die Anwendung der Regeln auf bestimmte morphologische Klassen beschränkt. Da der Filter als morphologischer Kontext in Form von Merkmalstrukturen ausgedrückt wird, können die Merkmalstrukturen der Two-Level-Regeln mit

2 Koskenniemi (1983)

den Merkmalstrukturen des jeweils konkret vorliegenden Morphs unifiziert werden.

Bei der Analyse greift das System auf ein Morphlexikon zu, das zu jedem Morph eine Merkmalstruktur enthält. Es handelt sich aufgrund der jedem einzelnen Morph zugeordneten Information also um ein morph- bzw. morphembasiertes System, im Gegensatz etwa zu einem paradigmenorientierten System. Dies bildet die Voraussetzung dafür, daß das System neben Flexion auch Derivation und Komposition behandeln kann. Dem Morphlexikon gewissermaßen übergeordnet ist ein Lexemlexikon, aus dem die syntaktisch-semantische Information eines Lexems entnommen wird. Unterhalb des Morphlexikons sind die Zwei-Ebenen-Regeln angesiedelt, die phonologische Phänomene behandeln. Durch die den Morphen zugeordneten Merkmale werden die Zwei-Ebenen-Regeln um einen morphologischen Kontext erweitert. Auf diese Art und Weise können mit dem System auch morphologische Phänomene wie nichtkonkatenative Morphologie, Schwa-Epenthese oder Ablaut und Umlaut, die in der ursprünglichen Form des Zwei-Ebenen-Modells problematisch waren, behandelt werden.

Die Kanten des dem Analysesystem zugrunde liegenden Übergangsnetzwerkes enthalten einen Filter. Anhand dieses Filters besteht die Möglichkeit, Tests durchzuführen, um bestimmte Morphkonkatenationen gegebenenfalls zu blockieren. Die Regel "A: ä {=}-; [MORPH: [HEAD: [UMLAUT: [VALUE: +]]]]" enthält bspw. in dem dem Semikolon folgenden Klammerausdruck den Kontext für die Umlautregel, Dieser Kontext wird hergestellt durch Suffixe, die Umlaut erfordern, d.h. deren entsprechender Wert mit '+' angegeben ist, so daß bei der Morphemkonkatenation in der Merkmalstruktur des Stamms das Attribut Umlaut schließlich in Abhängigkeit von den jeweiligen Suffixen bzw. deren Umlautwerten gesetzt wird. Auf diese Weise ist es möglich, kränklich und handlich zuzulassen, nicht aber händlich. Diese Bildung wird durch Unifikation ausgefiltert.

Insgesamt werden unnötige Unifikationen bei der Analyse vermieden, um die Effizienz des Systems zu steigern.

2.7 Deutsche Flexions- und Kompositionsmorphologie auf 2-Ebenen-Basis (Anne Schiller, Universität Stuttgart)

Auch das von Anne Schiller vorgestellte morphologische Analysesystem basiert auf dem Zwei-Ebenen-Modell, wobei es gegenüber dem ursprünglichen Modell von Koskenniemi nur geringfügige Veränderungen aufweist, was auch durch die Verwendung des PC-Kimmo-Systems als Grundlage der Implementierung deutlich wird.
Ziel dieser Arbeit ist die Erstellung eines. großen Lexikons mittels der hier beschriebenen morphologischen Analyse.

Entsprechend der in der Zwei-Ebenen-Morphologie eingeführten Konvention werden Morphemgrenzen oder auf lexikalischer Ebene unterspezifizierte Merkmale durch diakritische Zeichen dargestellt. Je nachdem ob ein morphologischer Stamm umlautet oder nicht, wird die Morphemgrenze zum nachfolgenden Suffixmorphem unterschiedlich dargestellt.

Bsp: jung+er →jünger
jung(+)er →junger

Die Einträge des aus bestehenden Wörtern zu kreierenden Lexikonteils werden hergeleitet aus Wortlisten, die kategoriale Angaben zu den einzelnen Wörtern enthalten. Da das System in erster Linie für die Analyse verwendet wird, enthält es auch Regeln, die zu Übergenerierung führen.

Diskontinuierliche Affixe, wie sie bei der Flexion von Verben im Deutschen beispielsweise im Fall des Partizips Perfekt auftreten können, werden im Lexikon ebenfalls mittels diakritischer Zeichen angegeben. Der

Stamm des Infinitivs *bauen* wird dementsprechend mit dem Merkmal [-Part] versehen, so daß aufgrund der lexikalischen Form „[-Part]bau" bei der Bildung des Partizips die Form *eingebaut* erkannt, *eingebaut* jedoch blockiert wird. Regeln für die Umlautung werden ebenfalls durch Diakritika an den Morphemgrenzen angegeben (Bsp: *Haus $er* oder: *groß $er).* Über solche Regeln hinaus verfügt das System über Kontrollregeln, die beispielsweise Fälle separierbarer präfixaler Elemente kontrollieren. So sorgt die Regel " [-Imp] :=::} [-Sep]_*" dafür, daß aufgrund des. Lexikoneintrags von *steh,* der als "steh[-Imp]" angegeben ist, und desjenigen von auf, der mit "auf[Sep]" angegeben ist, zwar *steht,* aber nicht *aufsteht* akzeptiert wird. Komposita des Typs N N (Substantiv+Substantiv), N Adj (Substantiv+Adjektiv) oder Adj N (Adjektiv+Substantiv) können im vorliegenden System auch unter Zuhilfenahme von Kontrollregeln bearbeitet werden, erfordern jedoch die Einführung spezieller zusätzlicher Kontrollregeln.

Das System in seiner derzeitigen Form verfügt über 15 Regeln. Die übrigen morphosyntaktischen Zusammenhänge werden unter Verwendung von Fortsetzungsklassen ausgedrückt. Für Nomina bestehen insgesamt 60 Fortsetzungsklassen, für Adjektive 20 und für Verben 40 solcher Klassen.

2.8 MORPH - ein modulares und robustes Morphologieprogramm für das Deutsche (Gerhard Hanrieder, FORWISS Erlangen)

Das von Gerhard Hanrieder entwickelte System MORPH ist ein morphologisches Programm, das sich aus drei Analysemodulen zusammensetzt. Ein Modul ist für die Flexionsanalyse konzipiert, ein weiteres Modul bearbeitet Wortbildungen, und das dritte Modul bearbeitet Formen mit unbekannten morphologischen Kernen und erzeugt auf der Grundlage der jeweils vorliegenden Endungen Hypothesen über die morphosyntaktischen Kategorien. Sofern eine Wortform also nicht mittels der ersten beiden Module analysiert werden kann, werden im dritten Modul Hypothesen über seine morphosyntaktische Kategorie angegeben. Dieses dritte Modul verleiht dem System den Aspekt der Robustheit.

Bei der Analyse eines Wortes werden alle möglichen Lesarten ausgegeben. Die Lexikoneinträge enthalten keine Angaben über Valenzen. Alle Einträge enthalten ausschließlich Angaben zu morphosyntaktischen Merkmalen wie Wortklasse, Numerus oder Kasus. Das Lexikon ist ein gemischtes Voll- und Stammformenlexikon, das als Liste von Listen angelegt ist. Kürzel zur Bezeichnung von Morphemklassen stellen die Verbindung einzelner Morpheme zu Klassen von Morphemen her. Mehrdeutigkeiten sind im Lexikon als verschiedene Lesarten kodiert, Allomorphe sind als gesonderte Einträge verzeichnet, wie das Beispiel (HAUS (MSING (ALLO HAEUS))) zeigt.

Die Lexikonlisten sind aus Effizienzgründen als Buchstabenbäume kompiliert. Die Suche eines Eintrags im Lexikon kann somit auf der Grundlage einer Baumtraversierungsfunktion erfolgen. Die Wortbildungsanalyse des zweiten Moduls erfolgt als *left-to-right-Analyse.* Die Kombination von Morphemen ist jeweils als Übergangsnetzwerk implementiert. Die möglichen Nachfolger eines Substantivs (N) sind z.B. angegeben mit (N), (Adj), (Präf) und (Suff).

Bei der " Endungshypothesenanalyse" wird nach dem Verfahren des *longest matching* die Endung eines Wortes mit unbekannter Basis isoliert. Die einer Endung zugeordneten Merkmale erlauben schließlich, Hypothesen über die vorliegende Wortform anzugeben.

2.9 Informationsgewinnung aus der Struktur lexikalischer Lemmata (Nico Weber, Universität Bonn)

Die von Nico Weber vorgestellte Informationsgewinnung aus der Struktur lexika

lischer Lemmata basiert auf der Untersuchung von Definitionstexten eines maschinenlesbaren Wörterbuchausschnitts des Deutschen3, in denen die morphologische Struktur der Lemmata thematisiert bzw. durch geeignete Lexeme variiert wird. Besonders ergiebig für die Informationsgewinnung aus der Struktur lexikalischer Lemmata haben sich dabei 1- bis 3-WortDefinitionen erwiesen, ein Definitionstyp, der - so die Schätzungen des Autors ca. 90 % der Definitionen im betrachteten Wörterbuchausschnitt umfaßt. Definitionen dieses Typs enthalten als Definiens des in Frage stehenden Lemmas meistens ein Wort, das entweder das Lemma, die Bedeutung von Ableitungsaffixen am Lemma oder die Wortbildungsbedeutung eines abgeleiteten bzw. zusammengesetzten Lemmas variiert.

Außer den Satzbanddaten des DUW zur Sichtung und anschließenden Klassifikation von Definitionstexten nach phrasalen Strukturen wurden zu Testanalysen 300.000 Einträge der Bonner. Wortdatenbank (Bonniex), 230.000 Einträge des Wortanalytischen Wörterbuchs [WAW]4, die sowohl in segmentierter als auch in unsegmentierter Form vorliegen, sowie eine Morphemliste mit 7677 Elementen herangezogen:

Bei den verwendeten Tools zur Extraktion der Daten aus den Definitionstexten handelt es sich um einen Satzbandparser mit SGML-konformer Ausgabe sowie. um ein Selektionsprogramm für Lemmata, Lesartenangaben und die eigentlichen Definitionstexte.

Der morphologische Parser, der die vorbereitende Segmentierung der Lemmata und Einwortdefinitionen vornimmt, ist lexikonorientiert und wurde auf der Grundlage der Daten des Wortanalytischen Wörterbuchs (WAW) erstellt. Die Segmentierung erfolgt von links nach rechts nach dem Verfahren des *longest matching.* Das Segmentierungsergebnis wird zunächst daraufhin überprüft, ob es in der vom Segmentierungsprogramm gelieferten Form im W A W, das "Einträge in segmentierter Form verzeichnet, vorhanden. ist. Wird im WAW keine entsprechende Segmentierung des in Frage stehenden Wortes gefunden, wird eine Kompositaanalyse durchgeführt. Die dabei ermittelten Lexeme werden mit den Wörterbucheinträgen verglichen. Ist das betreffende Lexem im WAW enthalten, wird es auf die in den Definitionstexten auftretenden Variationen seiner Wortbildungsbedeutung hin überprüf~. Verschiedene Typen von Variationen treten hier auf:5

(1) Kombinatorische Varianten: Lexeme variieren]

Bahn	verbindung
Zug	verbindung
Abend	mahl
Abend	essen
Ausguß	becken
Ausguß	0

(2) Diasystematische Ersetzung: [Lexemsubstitution]

achtern
hinten

(3) Affixexplikation mit Stammwiederholung:

[Präfixe variieren]

Ab	-	lauf
Ver	-	lauf

[Suffixe variieren]

abänder - lich
abänder - bar

Die in den Lexemdefinitionen auftretenden Alternanzen lassen sich semantisch in-

3 Es handelt sich bei dem betrachteten Wörterbuchausschnitt um die Einträge <A> bis <Band> des in Form maschinenlesbarer Satzbanddaten vorliegenden Buden - Deutsches Universalwörterbuch [DUW]..

4 Wortanalytisches Wörterbuch. Deutscher Wortschatz nach Sinn-Elementen (192ff.). Kandler, Günther/Winter Stefan. München: Wilhelm Fink Vlg.

5 Die Unterscheidung Webers nimmt die Klassifikation morphosemantischer Definitionen von J. Rey-Debove (1971) auf. Vlg. Weber (1992:10).

terpretieren auf der Grundlage der klassischen Relationen wie Synonymie, Hyponymie oder Antonymie. Der Parser, der die Analyse der Lexemdefinitionen vornimmt, sucht nun mit regulären Ausdrücken nach formalen Variationsmustern. Die regulären Ausdrücke berücksichtigen die Forderung, daß in zwei variierenden Wörtern mindestens ein Element an derselben Position materiell identisch auftreten muß. Der Abgleich mit dem Referenzmuster Ax Lx ist z.B. möglich in Fällen wie

Ax Lx:	*Ab-zug*	*Ab-guß*	*Ab-tausch*
"-Lx	*Abdruck*		
Ax-"		*Aus-guß*	
0-"			0- *Tausch*

Dabei variiert jeweils die im Suchpattern explizit aufgeführte Form (Lx = Lexem; Ax = Affix; 0 = Nullmorphem), während das mit dem Referenzmuster identische Element jeweils durch <<“>> angegeben wird. Eine Interpretation kann sowohl bei Komposita als auch bei Derivaten erfolgen. Die Derivationsbedeutung wird auf der Grundlage der Derivationsaffixe und den jeweils vorliegenden Definitionstexten ermittelt. So werden beispielsweise Ableitungen mit dem Suffix *ung* wie *Programmierung* in der Regel durch eine entsprechende substantivierte Infinitivform *(das Programmieren)* wiedergegeben, wodurch das Suffix eine prozessuale Interpretation erhält, und aufgrund des Definitionstextes des Eintrags *andünsten* ("kurz dünsten") kann dem Präfix *an* z.B. die Bedeutung "kurz" zugeordnet werden.

2.10 Schnittstellen zu Nicht-ASCII-Zeichen (S.Y. Cho, Universität Saarbrücken)

Den Abschluß der Präsentationen des Erlanger Workshops bildete die Vorstellung der Schnittstellen zu Nicht-ASCII-Zeichen von S.- Y. Cho, der die Möglichkeiten der Darstellung koreanischer, japanischer oder chinesischer Schriftzeichen in einem koreanischen Textsystem auf der einen bzw. einem englischen Textsystem auf der anderen Seite demonstrierte. Im Vordergrund dieses Beitrags stand kein morphologisches Analyseverfahren, sondern die von Cho und Lew 1992 entwickelte Lösung des Problems, in Textverarbeitungs- oder natürlichsprachlichen Systemen auf Schriftzeichen, die nicht auf dem lateinischen Alphabet beruhen bzw. nicht im Umfang der ASCII-Codierung enthalten sind, zugreifen zu können.

Die vorgelegte Lösung basiert auf der internen Darstellung eines Hangul-Zeichens durch zwei ASCII-Codes. Im koreanischen Textverarbeitungsprogramm lassen sich die Zeichen so in ihrer koreanischen Form anzeigen. In einer englischen Programmumgebung werden die jeweiligen Zeichen den zwei ASCII-Codes entsprechend durch zwei miteinander verbunden ASCIIZeichen (etwa: ea für ASCII: 136 97) wiedergegeben.

3 Ausblick

Die verschiedenen auf dem Workshop präsentierten Ansätze und Systeme sowie die sich an die Systemvorführungen anschließende engagierte Diskussion der Workshopteilnehmer über Bewertungskriterien für automatische Wortformerkennungssysteme deuten sicher schon jetzt auf einen spannenden Verlauf der 1. MORPHOLYMPICS, auf der sich im März nächsten Jahres konkurrierende Systeme zur Wortformerkennung des Deutschen einer Jury stellen sowie um den ersten Platz und ein Preisgeld <computerlinguistisch kämpfen> werden.

Literatur

De Rose, S.J. (1988): Grammatical category disambiguation by statistical optimization. In: Computational Linguistics, 14/1, S. 31-39.

Domenig, Marc/Ten Hacken, Pius (1992): Word Manager: A System for Morphological Dictionaries. Hildesheim: Olms.

Feldweg, Helmut (1993):
Stochastische Wortartendisambiguierung für das Deutsche. Untersuchungen mit dem robusten System LIKELY. SfS-Report-08-93, Universität Tübingen.

Hausser, Roland (1989): Principles of Computational Morphology. Laboratory of Computational Linguistics, Carnegie Mellon University.

Hausser, Roland/Schüller, Gerald/Zierl , Marco (1993): MAGIC. A Tutorial in Computational Morphology. Univ. Erlangen-Nürnberg.

Hellwig, Peter (1992): The Morphology-Aid Function. Workpackage 2.4., T2, Universität Heidelberg.

Koskenniemi, Kimmo (1983): Two-level morphology. A general computational theory for word-form recognition and production. Department of General Linguistics, University of Helsinki, Publications no. 11.

Marshall. I. (1983): Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus. In: Computers in the Humanities, 17, S. 139-150.

Marshall. I. (1987): Tag selection using probabilistic methods. In: Garside, R./Leech, G./Sampson, G. (Hrsg.): The computational analysis of English. London/New York: Longman, S. 4256.

Trost, Harald (1993): Coping with Derivation in a Morphological Component. Erscheint
in: Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL '93), Utrecht.

Weber, Nico (1992):
Morphosemantische Wörterbuchdefinitionen. In: Sprache und Datenverarbeitung 16/2, S. 45"::63.

Uta Seewald, Hannover