Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Bielefelder Lexikon und Morphologie in VERBMOBIL Phase II

Harald Lüngen
Karsten Ehlebracht
Dafydd Gibbon
Ana Paula Quirino Simões

Universität Bielefeld

Vm
Report 233
November 1998

November 1998

Harald Lüngen
Karsten Ehlebracht
Dafydd Gibbon
Ana Paula Quirino Simões

Universität Bielefeld (UBI)
Fakultät für Linguistik und Literaturwissenschaft
Universitätsstr. 25
Postfach 10 01 31
33501 Bielefeld

Tel.: (0521) 106 - 3510
Fax: (0521) 106 - 6008
e-mail: `luengen@Spectrum.Uni-Bielefeld.DE`

Gehört zum Antragsabschnitt: 1.3.1 Aufbau und Optimierung des Aussprachelexikons
6.4.1 Kommunikation, WWW-Dokumentation, Sprecherdatenbasis

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 B 2 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

ISSN 0949–6084

Inhaltsverzeichnis

1 Die Bielefelder Lexikographie in VERBMOBIL

Die Bielefelder Arbeitsgruppe befaßt sich im Rahmen von VERBMOBIL Phase II mit der automatisierten Erstellung von morphologisch erweiterten Aussprachelexika und mit der Koordination der Wortschatzabdeckung.

Da diese lexikographische Infrastrukturarbeit sich nicht als Modul im VERBMOBIL–System niederschlagen kann, sondern als Ressource von mehreren Modulen genutzt wird, haben wir uns entschlossen, die Arbeiten in der Form dieses Übersichtsberichts in überschaubarer Weise sichtbar zu machen.

Für die zweite Phase von VERBMOBIL sind Wortschätze im Bereich von 10000 Wörtern (Vollformen) vorgesehen. Die Wortformenvariation in deutschen Texten erhöht die Vollformenzahl gegenüber der Stammzahl aufgrund der Flexion um den Faktor 3-5, je nach Texttyp. Das Englische hat dem gegenüber einen 'Wortschatzvorteil' mit einer Wortformenvariation geringfügig ≥ 1. Auch in Bezug auf das häufige Vorkommen von Wortbildungen (Derivation und Komposition) sind die Erfahrungen mit dem Englischen nicht auf das Deutsche zu übertragen, hinzu kommt, daß im Deutschen Wortbildungen orthographisch konsequent als Einzelwörter geschrieben werden, während sie im Englischen zum großen Teil bereits durch Bindestrich oder Leerzeichen 'vorsegmentiert' sind, vgl. *twenty-one* vs. *einundzwanzig, travel agency* vs. *Reisebüro* oder *look up* vs. *nachschlagen.* Eine weitere Erkenntnis aus VERBMOBIL Phase I ist, daß ein erheblicher Anteil von Komposita, Derivationen und flektierten Formen bereits bekannter Wortstämme sich unter den unbekannten Wörtern (*Out Of Vocabulary (OOV) items*), d.h. solchen, die nicht im Vokabular des Spracherkenners sind, befindet, und sich mit wachsendem Vokabular weiter vergrößert.

Ziel der Arbeiten an der Universität Bielefeld ist daher der Aufbau eines Kernlexikons, dessen *extensionale Abdeckung* den projizierten Wortschatz der gesamten deutschen VERBMOBIL-Transliterationen umfaßt.

Die Arbeiten in VERBMOBIL Phase II setzen damit den Kern der lexikographischen Infrastrukturarbeiten fort, die von der Bielefelder Arbeitsgruppe in VERBMOBIL Phase I durchgeführt wurden. Insbesondere wurde auf folgende Kriterien geachtet:

1. lexikographische Konsistenz unter allen VERBMOBIL–Partnern durch Koordination der domänen– und systemspezifischen sowie multilingualen Abdeckungskriterien;
2. Datenbankkonsistenz durch Führung eines allgemein zugänglichen aktuellen Einzel–Tokens der DB im WWW;
3. Wiederverwendbarkeit der in früheren Arbeiten entwickelten Software zur morphologischen Paradigmengenerierung und zur HyprLex–Datenbankerstellung;
4. Evaluation der Ergebnisse durch Vergleich und Abgleich der Ausgaben unterschiedlicher Quellen (unterschiedliche morphologische Analyseverfahren, Ergebnisse der Partner).

Die wichtigsten theoretischen und methodischen Neuerungen gegenüber den Ergebnissen der Bielefelder Arbeiten in VERBMOBIL Phase I sind:

- der Einsatz einer speziell auf die Akquisition von Lexika für gesprochene Sprache entwickelten Erfassungssoftware mit eigenem Web–Server,
- die Optimierung der Klassifikationssoftware durch Approximation an finite Automaten (Finite State Transducer, FST),

- die Portierung der wissensbasierten Graphem–Phonem–Software von Prolog auf das XFST–System (Xerox) in Kooperation mit Lauri Karttunen, um die Vorteile des Xerox–FSA–Kalküls zu Optimierungszwecken nutzen zu können.

2 Lexikographische Akquisition des Ausspracheelexikons

2.1 Automatische morphologische Strukturierung des Wortschatzes

Der Import in diese Arbeitspakete sind die gefilterten Transliterationen der deutschen Dialoge aus der Datensammlung in VERBMOBIL, vgl. [bur96]. Die Vorverarbeitung (Konsistenzprüfung, Formatanpassung, Tokenisierung) erfolgt mit dem in Bielefeld in VERBMOBIL Phase I entwickelten, in München re–implementierten und modifizierten Transliterationsfilter ([GS95a], [GS95b], [Gib95]).

2.1.1 Anforderungen an die Lexikonakquisition

Die Akquisition des in VERBMOBIL verwendeten morphologisch erweiterten Ausspracheelexikons wird in sieben Stufen eingeteilt, die nicht notwendigerweise rein linear bearbeitet werden. An diese Stufen werden folgende Anforderungen gestellt:

Morphologische Segmentierung: Segmentierung der Graphem- und Phonemstrings in Strings, die Morpheme repräsentieren.

Lemmatisierung: Identifizierung des morphologischen Stamms (möglicherweise mehrere bei starken Verben) und der Zitierform.

Morphologische Klassifikation: Identifikation der Flexionsklasse von Substantiven und Verben.

Morphologisches Tagging: Tagging der Morphemsegmente mit morphologischen Kategorien.

Morphologisches Parsing: Zuweisung von Strukturbäumen zu segmentierten Graphemstrings.

Lexikalische Integration: Konsistenter und korrekter Einbau neu akquirierter Wörter und Morpheme in die Datenbank.

Paradigmengenerierung: Generierung aller möglichen Formen des Flexionsparadigmas in orthographischer und phonologischer Repräsentation für jedes Lemma.

Die Evaluation der Ergebnisse erfolgt durch folgende Methoden:

1. automatische formale Verifikation der Datentypen;
2. Abgleich der Ausgaben des Paradigmengenerators (overall verification) mit zwei Methoden:
   (a) automatische formale Verifikation des Datentyps;
   (b) automatischer Stichprobenabgleich mit manuell verifizierten Testsuiten.
3. operationale Tests durch die VERBMOBIL-Partner sowie koordinierende Absprachen.

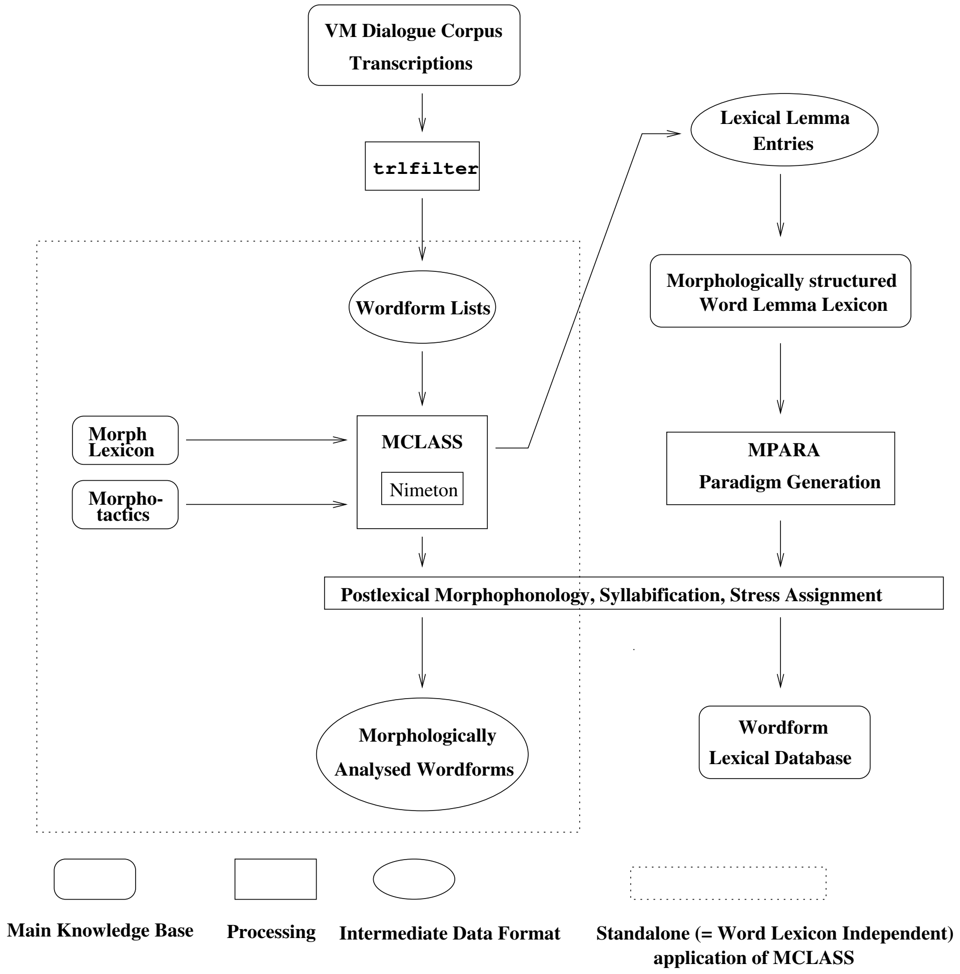

Abbildung 1: Korpusbearbeitung in VM-II AP 1.3.1.

2.1.2 Architektur der lexikographischen Korpusbearbeitung

Die Hauptmodule zur automatischen Bewältigung der Lexikonakquisition sind der morphologische Parser MCLASS und der Paradigmengenerator MPARA. Eine Vorstufe zur Entwicklung von MCLASS war das in VERBMOBIL Phase I entwickelte MSEG (vgl. [Ste95]). MPARA wurde in VERBMOBIL Phase I entwickelt, und ist ausführlich in [BDG96] beschrieben. Aufgabe von MPARA ist die konsistente morphologisch fundierte Erweiterung des durch das Corpus belegten Lexikons. Unter anderem wird hierdurch ein notwendiger infrastruktureller Beitrag zur Lösung des *Out Of Vocabulary (OOV) item*–Problems geleistet. Das Modul MPARA wird in VERBMOBIL Phase II zur Konstruktion des morphologisch erweiterten Lexikons intensiv eingesetzt. Der Generatorprototyp hat sich als stabil und als nahezu empirisch vollständig erwiesen; nur relativ kleinere Erweiterungen mußten in VERBMOBIL Phase II aufgrund der Evaluationszyklen durchgeführt werden.

Der Kern der flexionsmorphologischen Wissensbank wurde in VERBMOBIL Phase I entwickelt (vgl.

[Ble94], [Ble95]) und wird weiter verwendet.

2.2 Computerlinguistische Module

2.2.1 Mclass: Morphotaktik als Unifikationsgrammatik

Die Analyse der Morphotaktik (Wortstrukur) erfolgt mit einer HPSG–basierten Wortgrammatik, die im Format einer DCG implementiert wird. Warum wird hier eine unifikationsbasierte, kontextfreie Grammatik verwendet, und nicht etwa ein endlicher Automat? Linguistische Verallgemeinerungen, auch über Morphologie, lassen sich besser in einer kontextfreien Grammatik ausdrücken, außerdem bietet dieses Format eine explizite Schnittstelle zur lexikalischen Semantik. Lineare Strukturen, wie durch endliche Automaten beschrieben, lassen sich nicht in geeigneter Weise semantisch interpretieren.

Die DCG-Morphotaktik, die der morphologischen Analyse in Mclass zugrundeliegt, enthält derzeit 43 Regeln, wie z.B die folgende Regel zur Definition der linearen Präzedenzrelation zwischen Präfixen in komplexen abgeleiteten Stämmen:

```
% Rule2:      stemPrefixed --> part  stemPre1fixed
% Example:    aufbewahr    --> auf   bewahr

% Linear precedence and feature vectors:

stemPrefixed(X0_MID,X0_MCAT,X0_COMB_SUF,X0_LBND,X0_LBND_ARG,X0_RBND,X0_RBND_ARG,X0_BCAT,
             X0_SCAT,X0_NAT,X0_SEP,X0_STR,X0_ADJ,X0_LINK,X0_LINK_MID,X0_UMGL,X0_UMLB,X0_UMLD,X0_ABGL,
             X0_ABLB,X0_ABLD,X0_ABL_SUF_MID,X0_PRET,X0_REG,X0_INTERF,X0_INTERF_ARG,X0_GEN,X0_HEAD_MID,
             X0_ARG_MID,X0_VROOT,X0_STEMCLASS,X0_SUFFIXCLASS,X0_FLEX,X0_PARTSTEM):-

   part(X1_MID,X1_MCAT,X1_COMB_SUF,X1_LBND,X1_LBND_ARG,X1_RBND,X1_RBND_ARG,X1_BCAT,
        X1_SCAT,X1_NAT,X1_SEP,X1_STR,X1_ADJ,X1_LINK,X1_LINK_MID,X1_UMGL,X1_UMLB,X1_UMLD,X1_ABGL,
        X1_ABLB,X1_ABLD,X1_ABL_SUF_MID,X1_PRET,X1_REG,X1_INTERF,X1_INTERF_ARG,X1_GEN,X1_HEAD_MID,
        X1_ARG_MID,X1_VROOT,X1_STEMCLASS,X1_SUFFIXCLASS,X1_FLEX,X1_PARTSTEM),

   stemPre1fixed(X2_MID,X2_MCAT,X2_COMB_SUF,X2_LBND,X2_LBND_ARG,X2_RBND,X2_RBND_ARG,
        X2_BCAT,X2_SCAT,X2_NAT,X2_SEP,X2_STR,X2_ADJ,X2_LINK,X2_LINK_MID,X2_UMGL,X2_UMLB,X2_UMLD,
        X2_ABGL,X2_ABLB,X2_ABLD,X2_ABL_SUF_MID,X2_PRET,X2_REG,X2_INTERF,X2_INTERF_ARG,X2_GEN,
        X2_HEAD_MID,X2_ARG_MID,
        X2_VROOT,X2_STEMCLASS,X2_SUFFIXCLASS,X2_FLEX,X2_PARTSTEM),

% Category-specific constraints:
  X0_MCAT=stemPrefixed,
  X0_BCAT=X1_SCAT,
  X2_SCAT=v,
  X1_SEP='+',
  X1_NAT='+',
  X1_STR='+',
  X1_ADJ='-',
  X1_BCAT=X2_SCAT,
  not(X2_LBND==lbndFree),
  X1_SCAT=X0_SCAT,
  X1_LBND=X0_LBND,

% Projection (head-feature) constraints:
  X2_RBND=X0_RBND,
  X2_LINK=X0_LINK,
  X2_LINK_MID=X0_LINK_MID,
```

```
X2_NAT=X0_NAT,
X2_SUFFIXCLASS=X0_SUFFIXCLASS,
X2_UMGL=X0_UMGL,
X2_UMLB=X0_UMLB,
X2_PRET=X0_PRET,
X2_FLEX=X0_FLEX,
X2_VROOT=X0_VROOT,
X2_REG=X0_REG,
X2_STEMCLASS=X0_STEMCLASS,
X2_INTERF=X0_INTERF,
X2_PARTSTEM=X0_PARTSTEM.
```

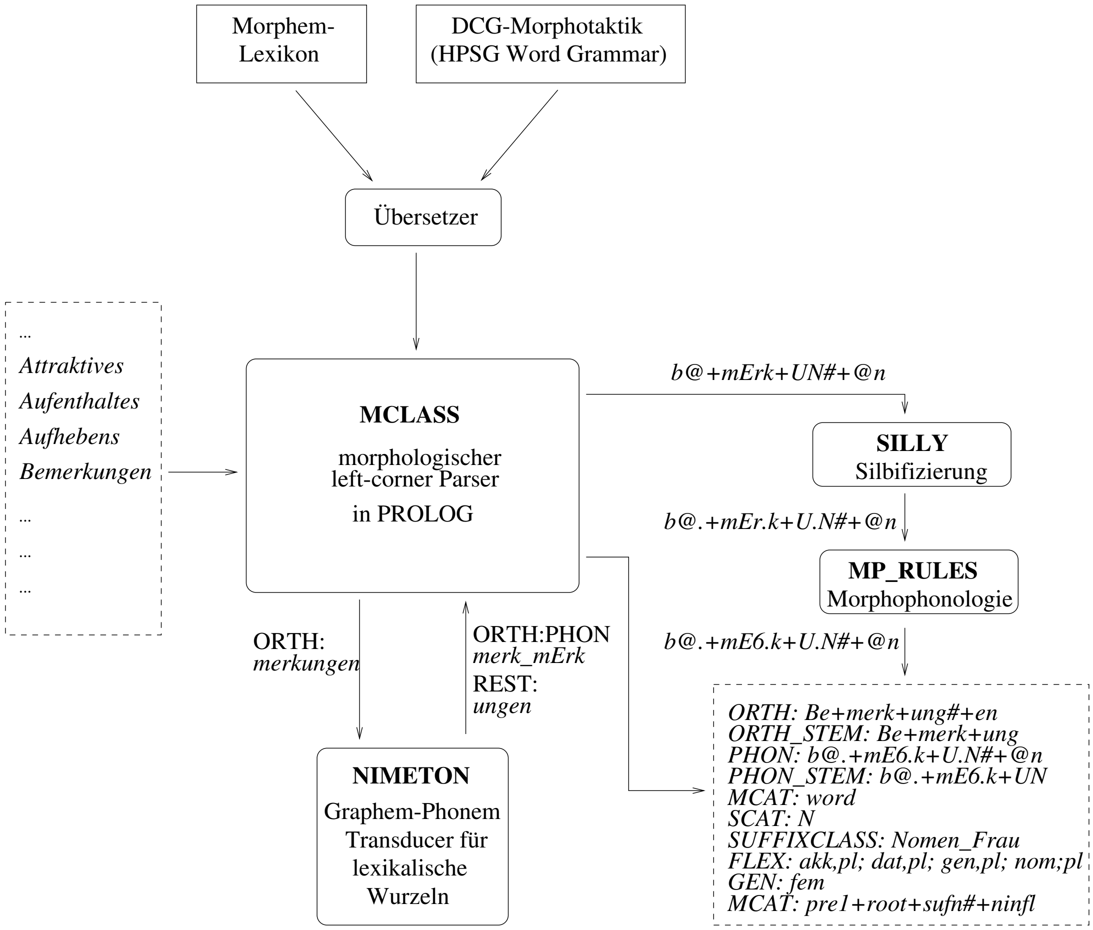

Abbildung 2: Architektur der morphologischen Analyse mit MCLASS.

Aus der DCG-Wortgrammatik und der Morph–Datenbank wird ein Prolog–left–corner–Parser MCLASS automatisch erzeugt, der für die Korpusanalyse eingesetzt wird (vgl. Abb. 2).

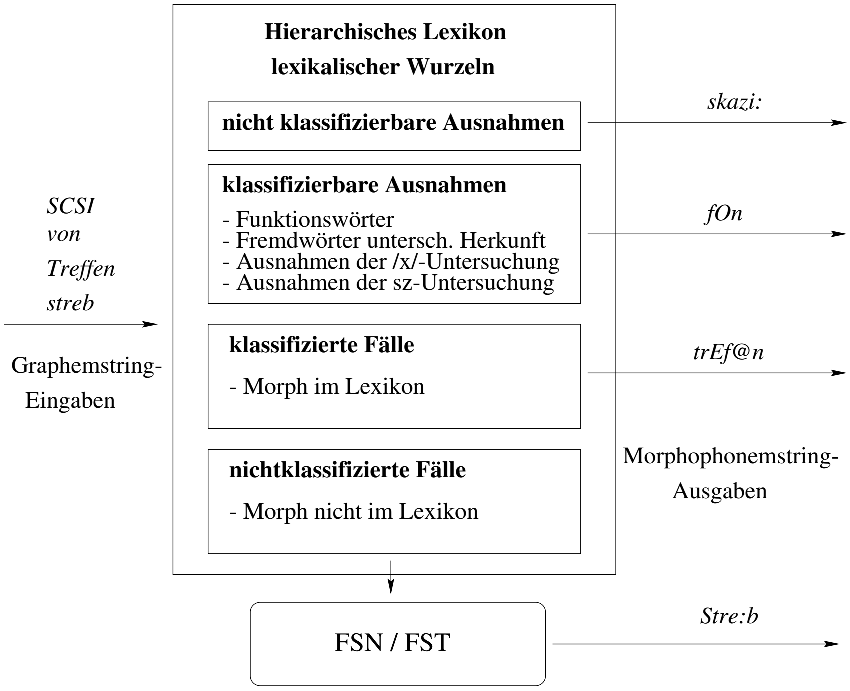

Abbildung 3: Architektur des FST für lexikalische Wurzeln NIMETON.

2.2.2 Nimeton: Erkennung unbekannter lexikalischer Wurzeln durch endliche Automaten

Es sind natürlich nicht alle neu im Korpus vorkommenden lexikalischen Wurzeln dem System bekannt. Daher gibt es eine Schnittstelle zu NIMETON, dem Graphem-Phonem-Transducer für lexikalische Wurzeln (vgl. [Mat96]). Hier wird graphophonotaktisch analysiert, welcher Teil eines unbekannten orthographischen Substrings im Input potentiell eine lexikalische Wurzel darstellt. NIMETON ist ein reiner FST, in Prolog implementiert, der ein in Prolog repräsentiertes Übergangsnetzwerk verwendet. Als Übergangsnetzwerke stehen ein graphophonotaktisches Netzwerk für native deutsche Wurzeln und eines für deutsche Namensbestandteile zur Verfügung, sowie eine Sammlung von Ausnahmepfaden.

Zur Optimierung von Repräsentation und Performanz wird zur Zeit ein Übersetzer entwickelt, der das NIMETON–Prolog–Format in das XFST–Format konvertiert, das mit der XFST–Software von Xerox Research Grenoble unter Lizenz verwendet wird.

Die kleinere Silbifizierungskomponente SILLY (vgl. [Mat98]), die morphophonologische Komponente MP_RULES und eine sich in der Entwicklung befindliche Komponente zur Zuordnung der lexikalischen Grundbetonung TONY sind ebenfalls als FST realisert. Eine Forschungsrichtung ist daher die Integration der einzelnen Verarbeitungsgrammatiken zu einem einzigen (riesigen) endlichen Automaten, vgl. Abschnitt 4.1.

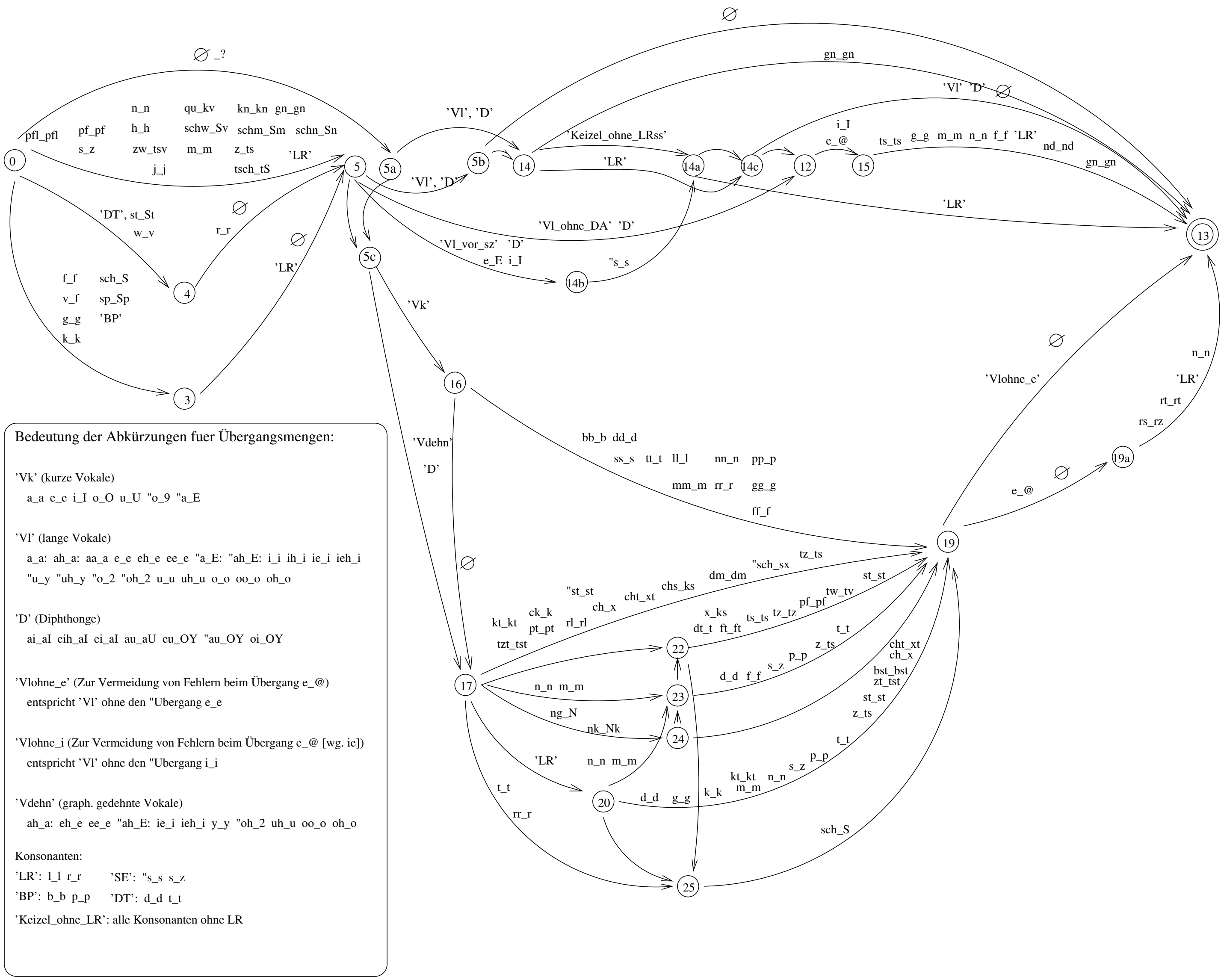

Abbildung 4: NIMETON–FSN für native lexikalische Wurzeln.

3 Lexikographische Koordination und WWW-Kommunikation

3.1 Lexikondatenbank: Typen lexikalischer Information und WWW-Distribution

Das wichtigste *deliverable* der Universität Bielefeld ist die regelmäßig aktualisierte Bielefelder Lexikondatenbank (vgl.[GE95]). Sie enthält Tags zum projizierten deutschen VERBMOBIL-Wortschatz, das sind derzeit 55891 Vollformen, bei 8272 im Korpus vorkommenden Wortformen. Die Datenbank hat also derzeit 55891 Datensätze. Die morphologische Information aus AP 1.3.1 wird hierdurch bereitgestellt, darüber hinaus auch statistische (Frequenz-) Information sowie weitere Tags aus Stuttgart und München. Insgesamt gibt es derzeit 31 Tags.

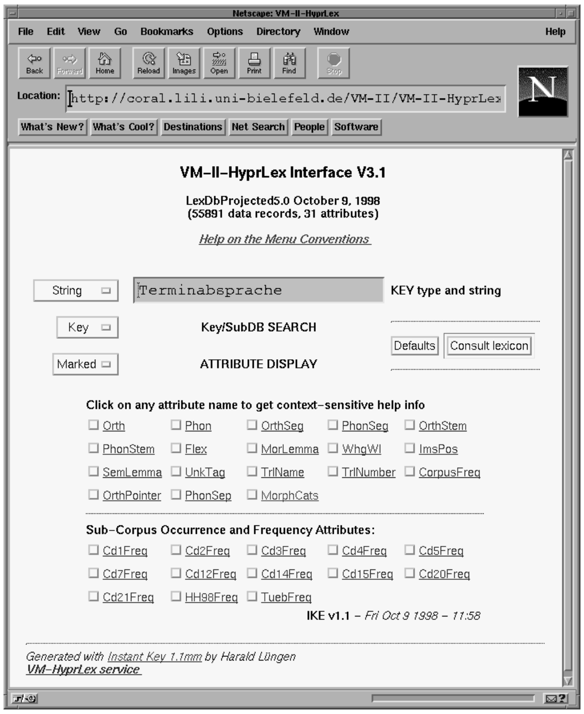

Abbildung 5: Lexikon–Datenbank: WWW-Schnittstelle.

Die Wortformen–Datenbank wird als einzelne Relation mit Unterrelationen, die in einzelnen Feldern direkt kodiert sind, formatiert:

```
Orth:          Terminabsprache
Phon:          tE6mi:n?apSpra:x@
OrthSeg:       Termin#ab#sprach#+e
PhonSeg:       tE6.m'i:n#?''ap#Spr''a:.x#+@
OrthStem:      Termin#ab#sprach
PhonStem:      tE6.m'i:n#?''ap#Spr''a:x
Flex:          N,akk,sg,fem;N,dat,sg,fem;N,gen,sg,fem;N,nom,sg,fem
MorLemma:      Terminabsprache
WhgWl:         1
ImsPos:        NN
SemLemma:      Terminabsprache
UnkTag:        *nil*
TrlName:       0
TrlNumber:     0
CorpusFreq:    15
OrthPointer:   *nil*
PhonSep:       t-E6-m-i:-n-?-a-p-S-p-r-a:-x-@
MorphCats:     stem_nn#part#stem_n#+infl_n
Cd1Freq:       2
Cd2Freq:       0
Cd3Freq:       2
Cd4Freq:       3
Cd5Freq:       1
Cd7Freq:       0
Cd12Freq:      7
Cd14Freq:      0
Cd15Freq:      0
Cd20Freq:      0
Cd21Freq:      0
HH98Freq:      0
TuebFreq:      0
```

Die Datenbank wird den Partnern in zwei Formen zur Verfügung gestellt:

1. ASCII–kodierte UNIX–Datenbank;
2. interaktive WWW–Datenbank mit online–Kontexthilfe und Archiv, über CGI–Server (Abb. 5).

Beide Versionen werden mit den Bielefelder lexikographischen Werkzeugen automatisch generiert.

3.2 Wortlisten in VERBMOBIL

Wortlisten definieren in VERBMOBIL den Inhalt der Modullexika, und damit den überhaupt verarbeitbaren Wortschatz, anhand der Schnittstelle zwischen Spracherkennung und Syntax (*Worthypothesengraph*).

Die Universität Bielefeld ist in der Gruppe der Modulverantwortlichen vertreten, um die Erstellung der auf dem Worthypothesengraph basierenden deutschen Wortliste (WHG–Wortliste) mit den Projektpartnern zu koordinieren und um Datenkonsistenz bezüglich der in VERBMOBIL Phase II aktuell gültigen, von verschiedenen Partnern beigesteuerten Wortlisten zu gewährleisten. Im multilingualen Kontext von VERBMOBIL gehört hierzu auch die Koordination mit den Partnern, die die englischen und japanischen Wortlisten erstellen (s. Abb. 6.

Das Konzept der *WHG-Wortliste* (vgl. [Lü98]) wurde zur Definition des deutschen Wortschatzes für das VERBMOBIL-System entwickelt.

Abbildung 6: Wortlistenverwaltung.

3.3 WWW-Schnittstellen für Komponenten der morphologischen Korpusbearbeitung

Die Erfahrungen mit der Nutzung der interaktiven lexikographischen WWW–Dienste, insbesondere für die Arbeit mit MCLASS und MPARA, machte es notwendig, einen Dämon für die Überwachung des Client–Server–Betriebs zu entwickeln.

Wird Software direkt mittels CGI-Script im WWW aufgerufen, so hat einerseits der Benutzer nach Absenden einer Anfrage keine Kontrolle über den Prozess, den er auf dem Server startet, andererseits hat der Prozess keine Information darüber, ob der Benutzer die Ausgabe erwartet. So würde ein Benutzer durch wiederholtes Absenden einer Anfrage (durch Absenden, Unterbrechen der Datenübertragung im Webbrowser und danach erneutes Absenden) mehrere Prozesse auf dem Server starten, während er nur die Ausgabe des zuletzt gestarteten Prozesses erhalten kann. Insbesondere bei Software mit langen Verarbeitungszeiten, wie MCLASS und MPARA, würde der Server unverhältnismäßig stark belastet.

Der Dämon MDAEMON kontrolliert diese Interaktion. Er läuft permanent und erhält alle Anfragen, die über einen bestimmten Port eingehen. Die generierte Eingangsseite (CGI-Form) beinhaltet eine Session-ID, so daß der MDAEMON jede Anfrage anhand der Session-ID einem Benutzer zuordnen kann. Existiert bereits ein Prozess für einen Benutzer, der einen neuen Prozess starten möchte, so löscht der MDAEMON nun zunächst den laufenden Prozess und startet dann den neuen. Die Zahl der Prozesse, auf deren Ausgabe niemand wartet, wird auf ein Minimum reduziert.

Abbildung 8 zeigt eine WWW-Anfrageseite, deren oberer Frame das Eingangsformular zum Aufruf von

MCLASS enthält. Der Benutzer kann die angezeigten Parameter setzen und mit dem START-Knopf die Verarbeitung auf dem Server in Gang setzen. Der untere Frame zeigt die aktiven Prozesse nach Absenden der Anfrage durch den Benutzer. Der MDAEMON hat die Prozess-ID 1802, darunter finden sich die HTML-Schnittstelle (PID 27235) und MCLASS (PIDs 27241 und 27242).

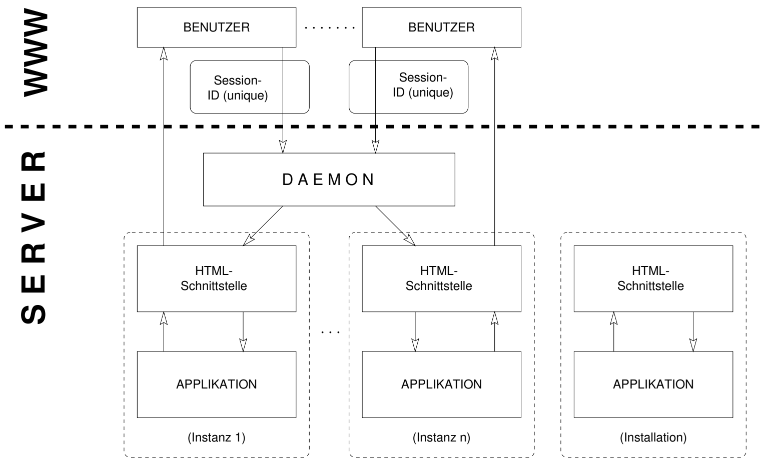

Abbildung 7: Kontrolle der Interaktion durch MDAEMON.

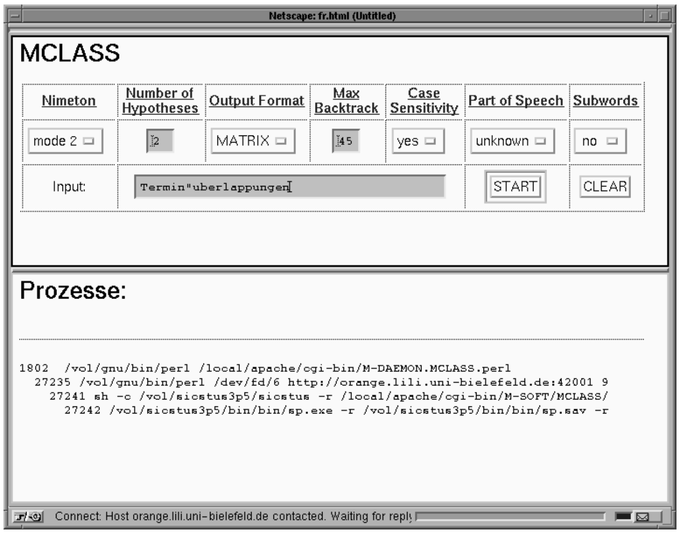

Abbildung 8: Experimentelle WWW-Schnittstelle für MCLASS

.

4 Beiträge zur lexikographischen Forschung

4.1 Morphologische Modellierung in HPSG mit Finite-State-Verarbeitung

Die Bielefelder Arbeitsgruppe hat sich seit vielen Jahren auf *Computational Linguistics and Spoken Language* spezialisiert, mit besonderer Fokussierung auf die Lexikographie gesprochener Sprache. Das Forschungs– und Entwicklungsinteresse der Arbeitsgruppe im Projekt VERBMOBIL ist die qualitativ hochwertige computerlinguistische Modellierung der Morphologie gesprochener Sprache mit dem Ziel einer effizienten, konsistenten und viele Informationstypen umfassenden Verarbeitung (vgl. [BDF$^+$95]). Ein Bielefelder Modul zur morphologischen online–Analyse spontan gesprochener Sprache war Bestandteil der inkrementellen Architektur von TP 15, VERBMOBIL Phase I, und ein Modul zur Reparatur unterbrochener Komposita war Bestandteil des Forschungsprototyps von VERBMOBIL Phase I, das für die Korrekturverarbeitung durch andere Partner in VERBMOBIL Phase II Pate stand. In VERBMOBIL Phase II liegt der Schwerpunkt auf der lexikographischen, insbesondere morphologischen Korpusbearbeitung zum Zweck der Lexikonakquisition.

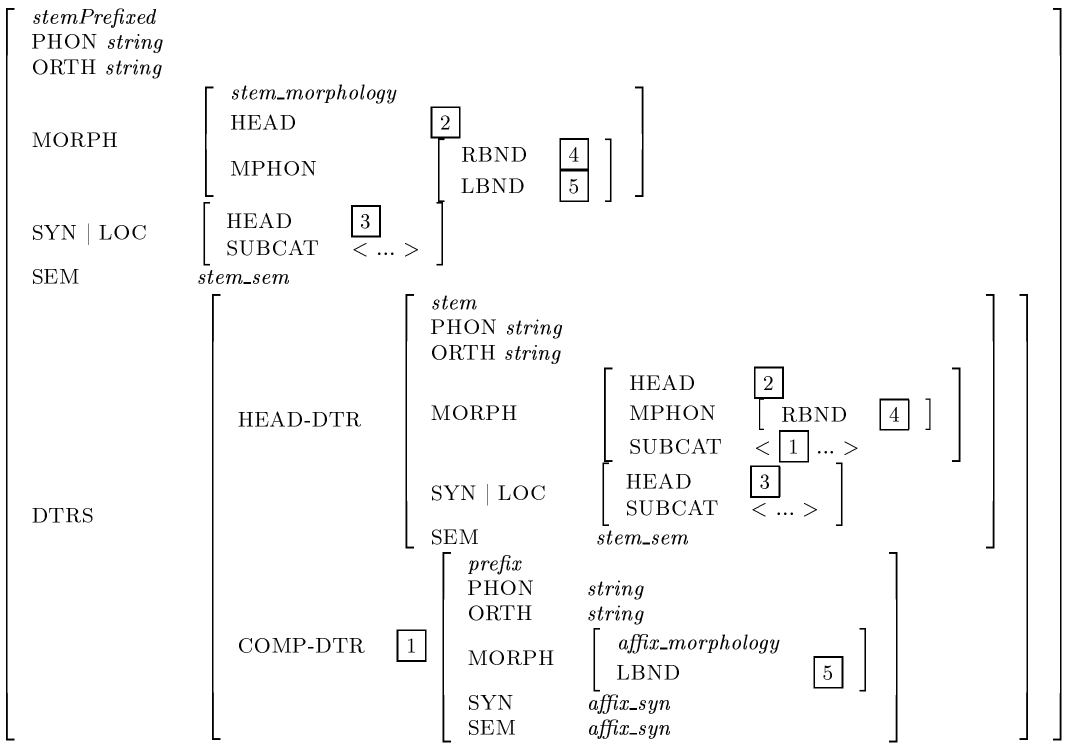

Abbildung 9: HPSG–sign *stemPrefixed* als AVM–Matrix

.

Die theoretische Fundierung wird im HPSG–Paradigma durchgeführt. Alle morphologischen Kategorien und Constraints werden in diesem Rahmen unter Berücksichtigung der Ansätze in [Kri93], [Rie], [Rie94] modelliert, sowie des Bielefelder ILEX–Ansatzes zur Modellierung von Kompositionalität im Lexikon (cf. [Gib91], [Gib93], [Gib]). Diese linguistische Theorie bietet eine explizite Schnittstelle zur lexikalischen

Semantik, die von anderen Partnern bearbeitet wird. Ein Beispiel eines HPSG-signs, das einerseits einen lexikalischen Wortbildungstyp im hierarchischen Lexikon repräsentiert, andererseits eine Wortbildungsregel wie in Abschnitt 2.2.1, findet sich in Abbildung 9.

Andererseits ist aus der Computermorphologie bekannt, daß morphologische Verarbeitung mit endlichen Automaten erfolgen kann, dem einfachsten und effizientesten Typ von abstrakter Maschine. Es wird daher zur Zeit ein Übersetzer implementiert, der, zunächst nach den Algorithmen aus [PW91] und [Roo96], aus der vorliegenden Wortgrammatik automatisch ein äquivalentes oder zumindest approximativ äquivalentes Finit State Transition Network (FSN) generiert, das einen FST darstellt. Damit würden alle Module der morphologischen Korpusbearbeitung auf endlichen Automaten beruhen, und können mithilfe des FSA–Kalküls von Kaplan und Kay (vgl. [KK94]) optimiert werden. Hierzu wird unter Lizenz von Xerox Research (Grenoble) die Implementierung XFST verwendet. Mit XFST soll das morphologische FSN mit den Morphem– und Wortlexika zu einem einzigen FST komponiert werden. Erste Evaluationen haben gezeigt, daß der resultierende FST zwar extrem groß, jedoch extrem effizient ist.

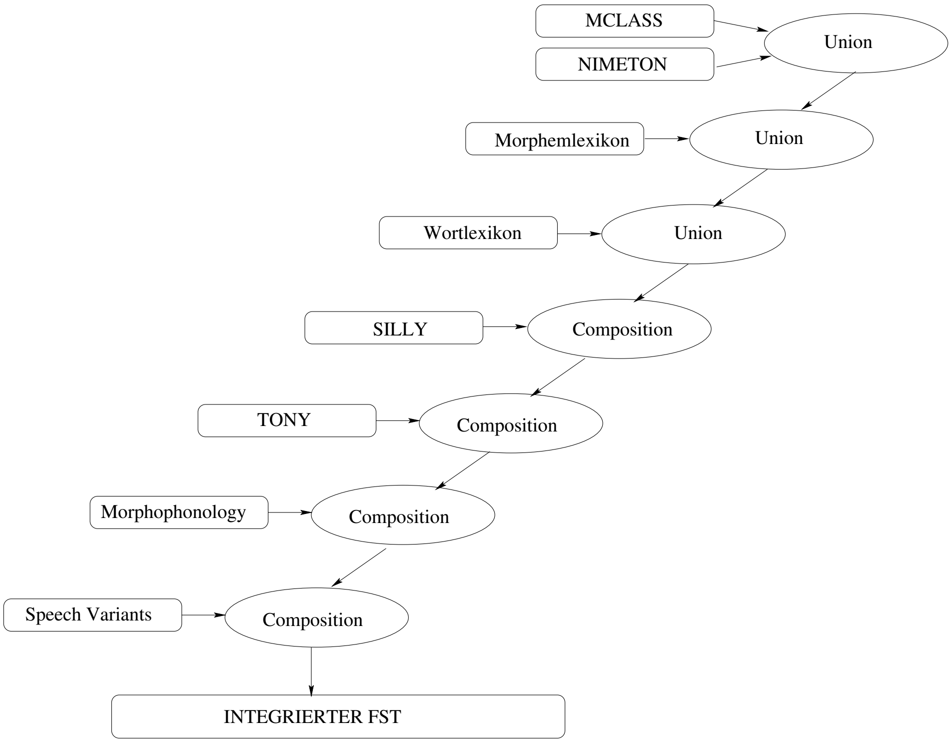

Abbildung 10: Integration endlicher Automaten.

Schließlich liefert dieses Vorgehen auch eine ideale Schnittstelle für Anwendungen in der stochastischen Spracherkennung hochflektierender Sprachen, z.B. mit dem in Bielefeld in VERBMOBIL Phase I in Kooperation zwischen TP5 und TP15 entwickelte Morphologie–Modul MEWES (vgl. [Alt97], [LPD+96], [AS96], [ADL+96]), das im Forschungsprototyp von VERBMOBIL Phase I eingesetzt wurde.

4.2 Interaktive Akquisition morphologischen Wissens (IAMW)

Als Datenbasis für MCLASS wird eine Morphem-Datenbank verwendet, die sämtliche Affixe des Deutschen und derzeit ca. 800 lexikalische Wurzeln nach insgesamt 40 Merkmalen klassifiziert, enthält. Eine Forschungsrichtung ist die Automatisierung der Entscheidungsbaum–gesteuerten interaktiven Akquisition neuer Merkmalsvektoren für diese Morphem–Datenbank. Die Entscheidungsbaum–Constraints über den sehr großen Vektorraum werden durch manuelle Steuerung definiert.

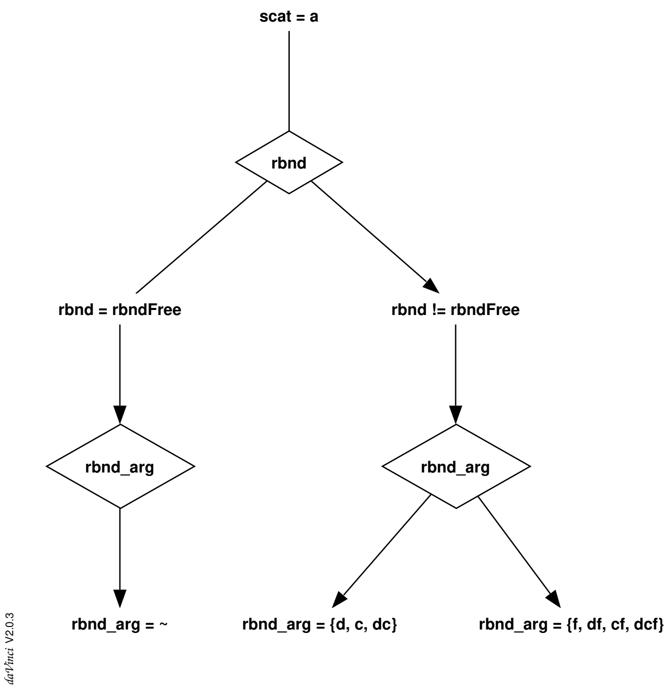

Abbildung 11: Teil eines Entscheidungsbaums zur Morphemklassifizierung.

Das Wissen über Zuordnung und Wertebereich aller Felder eines Datensatzes (entspricht einer expliziten Typisierung der Datensätze) wird in einem Entscheidungsbaum dargestellt. Die in diesem Entscheidungsbaum vorgegebenen Attribute werden von dem Programm IAMW interpretiert und steuern die weiteren Interaktionen mit dem Benutzer.

Nur wenn der Wert eines Attributes noch nicht aus dem vorhandenen Wissen ermittelt werden kann, wird der Benutzer zur Eingabe eines Wertes aufgefordert. Dazu wird der Wertebereich gezeigt, aus dem der Benutzer einen Wert auswählen kann.

IAMW ist ein generisches Werkzeug, das einen Entscheidungsbaum in einem vorgegebenen externen Format erwartet; somit können beliebige Datenbanken mit einer Struktur wie der beschriebenen interaktiv akquiriert werden. Die Entscheidungsbäume können z.B. von Linguisten spezifiziert werden oder maschinell aus Trainingsdaten gelernt werden. IAMW ist in Prolog implementiert.

5 Fazit: Lexikographie gesprochener Sprache in Forschung und Anwendung

Beim Anlauf des VERBMOBIL–Projekts Anfang der 90er Jahre war die Rolle der Lexikographie gesprochener Sprache, und speziell der Morphologie gesprochener Sprache, für die Systementwicklung noch nicht in allen Aspekten bekannt.

Heute kann festgehalten werden, daß einerseits eine umfassende lexikographische Koordination, andererseits eine theoretisch wohlfundierte und empirisch validierte kompositionelle Morphologie unverzichtbare Elemente für die Entwicklung komplexer sprachverarbeitender Systeme sind:

- Durch die kontinuierliche lexikographische Koordination anhand neuer Disseminationstechniken, z.B. das WWW, und unter Nutzung wiederverwendbarer Daten und morphologischer Analysewerkzeuge kann die lexikographische Datenkonsistenz schneller erreicht und effektiver gehalten werden.
- Ein *online*–Einsatz morphologischer Verarbeitung in der Spracherkennung (im Gegensatz etwa zur Sprachsynthese) liegt noch in der Zukunft, aber durch die Verwendung morphologischer *offline*–Verarbeitung können insbesondere für hochflektierende Sprachen Probleme wie

  - die Konsistenzerhaltung großer Lexika,
  - das *Out Of Vocabulary item problem*,
  - die Generierung großer Inventare von Aussprachevarianten

  systematisch und effektiv behandelt werden.

Literatur

[ADL+96] Frederek Althoff, Guido Drexel, Harald Lüngen, Martina Pampel, and Christoph Schillo. The treatment of compounds in a morphological component for speech recognition. In Dafydd Gibbon, editor, *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference*. Mouton de Gruyter, 1996.

[Alt97] Frederek Althoff. MEWES: Ein Modul für den Einsatz morphologischen Wissens bei der Erkennung gesprochener Sprache. Master's thesis, Universität Bielefeld, 1997.

[AS96] Frederek Althoff and Christoph Schillo. Morphology Component: Morphy Manual Version 1.1. VERBMOBIL Technisches Dokument 39. Universität Bielefeld, 1996.

[BDF+95] Doris Bleiching, Guido Drexel, Kerstin Fischer, Harald Lüngen, and Martina Pampel. Morphologie im Forschungsprototypen. VERBMOBIL Memo 83. Universtiät Bielefeld, 1995.

[BDG96] Doris Bleiching, Guido Drexel, and Dafydd Gibbon. Ein Synkretismusmodell für die deutsche Morphologie. In Dafydd Gibbon, editor, *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference*, Berlin, 1996. Mouton de Gruyter.

[Ble94] Doris Bleiching. Integration von Morphophonologie und Prosodie in ein hierarchisches Lexikon. In *Proceedings of KONVENS 94*, Vienna, 1994.

[Ble95] Doris Bleiching. Morphologie in der Lexikondatenbank. VERBMOBIL Technisches Dokument 21. Universtiät Bielefeld, 1995.

[bur96] Christof Burgard, Reinhard Karger, Wolfgang Wahlster (Hrsg.). Wissenschaftliche Ziele und Netzpläne für Verbmobil Phase 2. VERBMOBIL Technisches Dokument 44.DFKI GmbH, 1996.

[GE95] Dafydd Gibbon and Ute Ehrlich. Spezifikation für ein VERBMOBIL-Lexikondatenbankkonzept. VERBMOBIL Memo 69. Universtiät Bielefeld, Daimler Benz AG, 1995.

[Gib] Dafydd Gibbon. DATR theories or DATR programmes? On current computational models of morphology. Lecture Notes DGfS Jahrestagung 1995.

[Gib91] Dafydd Gibbon. ILEX: A linguistic approach to computational lexica. *Computatio Linguae. Zeitschrift für Dialektologie und Linguistik*, Beiheft 73, 1991.

[Gib93] Dafydd Gibbon. Generalised DATR for flexible lexical access: Prolog specification. VERBMOBIL Report 2. Universität Bielefeld, 1993.

[Gib95] Dafydd Gibbon. Verbmobil lexicon: Conventions for spelling and pronunciation. VERBMOBIL Technisches Dokument 31. Universität Bielefeld, 1995.

[GS95a] Dafydd Gibbon and Daniela Steinbrecher. VERBMOBIL-Standardfilter für Transliterationen Version 1.0. VERBMOBIL Technisches Dokument 32. Universität Bielefeld, 1995.

[GS95b] Dafydd Gibbon and Daniela Steinbrecher. VERBMOBIL-Standardfilter für Transliterationen Version 2.2. VERBMOBIL Technisches Dokument 38. Universtiät Bielefeld, 1995.

[KK94] Ronald M. Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, 1994. (First appeared as a paper presented to the Winter Meeting of the Linguistic Society of America, New York, 1981.).

[Kri93] Hans-Ulrich Krieger. Derivation without lexical rules. DFKI Research Report RR-93-27, 1993. DFKI Saarbrücken.

[LPD+96] Harald Lüngen, Martina Pampel, Guido Drexel, Dafydd Gibbon, Frederek Althoff, and Christoph Schillo. Morphology and speech technology. In *Proceedings of the 2nd ACL-SIGPHON Conference*, University of California, Santa Cruz, 1996. Association for Computational Linguistics.

[Lü98] Harald Lüngen. Extracting consistent vocabularies from corpora: The german whg-wordlist vmii-whg.wl.2.0. VERBMOBIL Technisches Dokument 64. Universtiät Bielefeld, 1998.

[Mat96] Martin Matthiesen. Nimeton - Ein Graphem-Phonem-Übersetzer für das Deutsche. VERBMOBIl Memo 109. Universität Bielefeld, 1996.

[Mat98] Martin Matthiesen. SILLY - silbifizierung mittels morphologischer Informationen. VERBMOBIL Memo 137. Universität Bielefeld, 1998.

[PW91] F. Pereira and Rebecca Wright. Finite-state approximation of phrase-structure grammars. In *Proceedings of the 29th ACL Conference*, 1991.

[Rie] Susanne Riehemann. Word formation in lexical type hierarchies. a case study of *bar*-adjectives in german. Magisterarbeit Universität Tübingen, 1993 und http://doors.stanford.edu/ sr/.

[Rie94] Susanne Riehemann. Morphology and the hierarchical lexicon. Unpublished Paper, http://doors.stanford.edu/ sr/, 1994.

[Roo96] C.M. Rood. Efficient finite-state approximation of context-free grammars. *Proceedings of the ECAI Workshop on Extended Finite State Models of Language*, 22, 1996. Budapest 1996.

[Ste95] Daniela Steinbrecher. MSEG: Morphologische Segmentierung deutscher Wortformen. VERBMOBIL Memo 99. Unversität Bielefeld, 1995.