BMBF

Verbmobil
Verbundvorhaben

Ein Synkretismusmodell für die deutsche Morphologie

Doris Bleiching, Guido Drexel, Dafydd Gibbon

Universität Bielefeld

Vm **Report 171**
September 1996

September 1996

Doris Bleiching, Guido Drexel, Dafydd Gibbon
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: {bleichin,drexel,gibbon}@Spectrum.Uni-Bielefeld.DE

Gehört zum Antragsabschnitt: 5.3 Morphologie

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 dem Autor.

Zusammenfassung

Morphologische Modelle in der Computerlinguistik reflektieren typischerweise sprachspezifische Eigenschaften, indem sie Datenstrukturen und Operationen verwenden, die der Typologie einzelner Sprachen entsprechen. Ausgehend von einer Diskussion der synkretistischen Eigenschaften der deutschen Flexionsmorphologie wird eine generische denotationelle Semantik für die bekannten sprachübergreifenden morphologischen Strukturen entwickelt. Diese Semantik liegt dem Konzept eines Compilers für ein generatives morphologisches Lexikon zugrunde, das 7000 Stämme aus einem Corpus spontan gesprochener deutscher Dialoge auf 30.000 Vollformen und 120.000 morphologische Kategorienabbildungen (nach Auflösung der Synkretismen) projiziert.

Morphology models in computational linguistics have tended to be language–specific, in that the data structures and operations used have reflected the typology of individual languages. Starting with a discussion of the syncretistic properties of German inflectional morphology, a generic denotational semantics for known language–independent inflectional structures is outlined. This semantics underlies the design of a generative morphological lexicon compiler for spoken German, which projects 7000 stems extracted from a corpus of spoken language dialogues to 30,000 fully inflected forms and 120,000 morphological category mappings (after resolution of syncretism).

1 Motivation

In diesem Beitrag wird eine generische denotationelle Semantik beschrieben, die die abstrakten Datenstrukturen für ein generatives morphologisches Lexikon mit expliziter Berücksichtigung der Synkretismen im Deutschen definiert.[1] Ein auf dieser Basis entwickelter Compiler wird zur Erzeugung großer Vollformenlexika für gesprochenes Deutsch eingesetzt; diese unterscheiden sich von bisherigen großen Aussprachelexika auf Graphem–Phonem–Übersetzungsbasis durch ihre formale Konsistenz.

Bisher entwickelte computerlinguistische Morphologiemodelle[2] sind von der Typologie der Sprachen, für die sie zuerst entwickelt wurden, stark geprägt worden. Beispiele für diese ‚Typologie–These' sind folgende:

- das schwach flektierende Englisch und klassische Vollformenmodelle (s. Ritchie et al., 1992);
- das agglutinative Finnisch und Zweiebenenmorphologie (s. Koskenniemi, 1983; Karttunen et al., 1987);
- das interkalative Arabisch und Mehrbandautomaten bzw. *template*–Strukturen (s. Kay, 1987; Gibbon, 1990; Reinhard und Gibbon, 1991);
- das fusionale Deutsch (wie auch Latein, Russisch) und Typen– oder Default–orientierte Objekthierarchien, Merkmalsstrukturen und verteilte Disjunktion (s. Bleiching, 1994; Cahill, 1993; Krieger und Nerbonne, 1993; Reinhard und Gibbon, 1991; Trost, 1993).

Die in diesem Papier beschriebenen Ergebnisse gehen von der Arbeitshypothese aus, daß formal adäquate Morphologietheorien und deren effiziente Verarbeitung auch in der Anwendungsentwicklung sprachtypologisch angemessen sein müssen (‚Typologie–These'). Für die Flexionsmorphologie des Deutschen ist also eine Paradigmentheorie angebracht, die auch die Synkretismusrelation morphologischer Neutralisierungen bzw. Mehrdeutigkeiten in der schriftlichen wie in der gesprochenen Modalität einbezieht.

Paradigmenstrukturen wurden in den letzten Jahren in Linguistik (z.B. von Stump, 1991) und Computerlinguistik (z.B. von Evans und Gazdar (in press))

[1] Dieser VM-Report wurde ursprünglich veröffentlicht als Beitrag in Dafydd Gibbon (ed.): *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference. Bielefeld, October 1996*, pp. 237–248. Berlin, etc.: Mouton de Gruyter.

[2] Eine Übersicht über einige Ansätze wird von Sproat (1992) gegeben; eine ‚Typologie–These' analog der hier vorgestellten entwickelt er jedoch nicht.

mehrfach thematisiert, wobei Stump den herkömmlichen Paradigmenbegriff[3] mit dem Konzept der *paradigm function*, die Stämme auf ihre deklinierten und konjugierten Formen abbildet, explizierte, während Evans, Gazdar und andere die hierarchische Modellierung von Regularitäten und Teilregularitäten in Paradigmen einführten. Eine Komponente der *paradigm function* sollte auch der für Sprachen wie Deutsch charakteristische Synkretismus sein, der von Spencer (1991) als ‚neutralized [...] morphosyntactic category' bzw. als ‚systematic inflectional homonymy' (S. 219) charakterisiert wird. Der Synkretismusbegriff wurde in informelleren linguistischen Arbeiten (s. die Diskussion durch Carstairs-McCarthy, 1992; Corbett, 1991; Matthews, 1991; Spencer, 1991; Stump, 1993; Zwicky, 1985) detailliert beschrieben und von Evans und Gazdar (in press) stichwortartig erwähnt; Mechanismen wie (z.B. verteilte) Disjunktionen, ‚Archikategorien' oder Default–Vererbung werden in der computerlinguistischen Literatur dafür ansatzweise verwendet. Die Synkretismusrelation als eigene Komponente einer *paradigm function* wurde aber bisher in computerlinguistischen Arbeiten noch nicht explizit modelliert.

In den folgenden Abschnitten wird ein explizites Synkretismusmodell vorgestellt, das mit einer Unterscheidung zwischen Affix– bzw. Stammoperationszuordnungen und Synkretismusrelationen ähnlich der zwischen *rules of exponence* und *rules of referral* (Zwicky, 1985) eine Forschungslinie von Zwicky über Stump weiterführt. Neu ist die Synthese folgender Aspekte: die generische denotationelle Semantik für computerlinguistische Morphologie–Anwendungen; die Erfassung der vollständigen Flexion einer Sprache (Deutsch) für beide Modalitäten der Orthographie und Phonologie, einschließlich Wortprosodie (Stammoperationen wie Apophonie, Konsonantenmodifikation, Betonungsverschiebung, (Re–)Silbifizierung, Betonung); die klare Trennung von Affix– und Stammsynkretismen, sowie die Hierarchiebildung sowohl mit Affix– als auch mit Stammsynkretismen.

Die folgende Darstellung orientiert sich nicht an gängigen Formalismen sondern an linguistischen Modellierungskonventionen. Nach einer Charakterisierung des Synkretismus im Kontext der Typologie in Abschnitt 2 wird in Abschnitt 3 zunächst der Synkretismus im Deutschen besprochen; anschließend wird in Abschnitt 4 das Synkretismusmodell explizit formuliert, in Abschnitt 5 eine Anwendung für die Generierung großer Aussprachelexika beschrieben und schließlich

[3]Der traditionelle Terminus ‚Paradigma' bedeutet die Menge der Abbildungen von Flexionskategorien auf die flektierten Formen eines Stamms; ‚Paradigmenklasse' bedeutet eine Menge von Stämmen mit ähnlich gebildeten Paradigmen (Deklinationen bzw. Konjugationen). Ein ‚Lexem' wird hier als die Menge der flektierten Formen eines Stamms definiert; ein (morphologisches) ‚Lemma' ist eine unterspezifizierte intensionale Repräsentation eines Paradigmas (also u.a. ohne Berechnung der Formen).

werden in Abschnitt 6 die Ergebnisse im Kontext der vorgesehenen Ziele diskutiert.

2 Synkretismus und Typologie der deutschen Flexion

Grundlage der Beschreibung eines Flexionssystems ist die Frage, „what is the extent and nature of the constraints, if any, on the deviation from the one–to–one pattern of inflectional realisation?“ (Carstairs-McCarthy, 1992, S. 193f.). Zur Beantwortung dieser Frage werden *constraints* über folgende morphologische Relationen benötigt:

1. morphosyntaktische Kategorien (z.B. als Merkmalsstrukturen mit Merkmalen wie ‚Person‘, ‚Numerus‘, ‚Genus‘);

2. Zeitrelationen in und zwischen morphophonologischen Basisformen im Lexikon oder frei gebildet (z.B. durch Affixverkettung, Stammvariation);

3. Abbildungen von morphologischen Kategorien auf morphophonologische Formen.

Die rein phonologisch bedingten temporalen Oberflächenrelationen in morphologisch komplexen Formen oder im Satzkontext (‚postlexikalische‘ oder Sandhi–*constraints*, einschließlich Auslautverhärtung, /r/–Vokalisierung) sind mit endlichen Maschinen sehr einfach modellierbar und können in diesem Rahmen außer Acht gelassen werden.

2.1 Strukturelle Klassifikation

Als Referenz für die weitere Diskussion wird von der ‚ideal einfachen‘ Flexionsmorphologie mit umkehrbar eindeutiger Abbildung zwischen morphosyntaktischen Kategorien und morphophonologischen Basisformen ausgegangen. Agglutinative Sprachen (z.B. Finnisch) kommen diesem ‚Ideal‘ recht nahe. Das ‚Ideal‘ muß allerdings nicht nur auf Affixverkettung sondern auch auf Operationen für Stammvariation, ‚prosodische‘ Überlappungsoperationen für Ton– oder Wortakzentzuordnung bezogen werden. Bei den Abweichungen von der umkehrbar eindeutigen Abbildung für Deutsch kommen alle Abweichungstypen von Carstairs (1987), S. 14ff., vor und stellen (wie in vielen indoeuropäischen Sprachen) ein komplexes Paradigmensystem dar:

1. *one–to–many syntagmatic*: Partizip Perfekt auf *ge ... (e)t/en*;

2. *one–to–many paradigmatic*: Plural auf *–e, –en, –er, –s* usw., oder auch freie oder stilistisch bedingte Variation, z.B. *Hunds – Hundes*;

3. *many–to–one syntagmatic*: Nominativ zusammen mit Plural auf *–e* usw.;

4. *many–to–one paradigmatic*: Alle Kategorien außer Nominativ Singular auf *–en*, z.B. *Matrose – Matrosen*.

Die *paradigm function* für Deutsch muß also alle vier Relationen erfassen. Von speziellem Interesse an dieser Stelle ist allerdings der vierte Abweichungstyp, der die Synkretismusrelation der systematischen flexionsmorphologischen Neutralisierung bzw. Mehrdeutigkeit charakterisiert, bei der „a single inflected form may correspond to more than one morphosyntactic description" (Spencer, 1991, S. 45).

2.2 Funktionale Klassifikation

Der Synkretismus in einzelnen Sprachen und für einzelne Wortarten in einer Sprache, wird auch nach morphosyntaktischen Kategorien klassifiziert, so etwa ‚Genus–Synkretismus', ‚Genus–Numerus–Synkretismus' (s. Corbett, 1991, S. 194f.). Corbett (S. 154ff., 190ff.) unterscheidet weiter zwischen verschiedenen Typen der ‚many–to–one'–Relation, indem er die verschiedenen Neutralisierungsrelationen in morphosyntaktischen Kategorien klassifiziert. Danach ist Deutsch als *convergent system* zu bezeichnen (Corbett, 1991, S. 155, S. 190), in dem z.B. bei Substantiven (sowie Adjektiven und Artikeln) die Kategorie Genus und bei Verben die Kategorie Person im Plural konsistent neutralisiert wird, aber nicht umgekehrt.

2.3 Lexikalische Klassifikation

Es soll aber auch auf folgende weitere Komplexitäten hingewiesen werden, die ein Synkretismusmodell unterscheiden bzw. erfassen muß:

1. Lexemsynkretismus (Vollformensynkretismus, z.B. bei suppletiven Formen: *sind* als Person–Synkretismus; einige ältere Ansätze kennen sogar nur den Lexemsynkretismus – also auch für nichtsuppletive Fälle – und beziehen den Begriff nicht auf Mehrdeutigkeit von Affixen oder Stämmen);

2. Affixsynkretismus (Mehrdeutigkeit von Affixformen, z.B. *–en*);

3. Stammsynkretismus (Mehrdeutigkeit von Stammformen, z.B. bei konstanter Stammform über mehrere morphosyntaktische Kategorien, etwa *brauch–e, brauch–te* oder *bräuch–te, bräuch–ten*);

4. Derivationssynkretismus (Mehrdeutigkeit bei Derivationsformen bzw. zwischen Derivationsformen und Flexionsformen, z.B. *entschieden* als Partizip und Adjektiv oder Adverb).

2.4 Modellierungskonventionen

Strenggenommen kann die Synkretismusrelation nicht bereits für einzelne Paradigmen oder Paradigmenklassen formuliert werden, da die jeweils dazugehörigen morphosyntaktischen Kategorien mit ihren verschieden spezifizierten Merkmalsstrukturen auch unabhängig voneinander formuliert oder auch schlicht als atomare Kategorien aufgefaßt werden können. Erst der Vergleich aller Paradigmenklassen im Hinblick auf gemeinsame morphosyntaktische Eigenschaften erlaubt die Definition von maximal spezifizierten morphosyntaktischen Kategorien, die wiederum die Formulierung der Synkretismusrelation als Neutralisierung bzw. Homonymie bezogen auf diese maximal mögliche Spezifikation erlauben. Hierdurch entsteht eine generalisierte Definition des traditionellen Begriffs ‚Paradigmenklasse' als Menge von Paradigmen, die dieselbe Synkretismuseigenschaft besitzen. Die *paradigm function* erscheint nicht als nicht weiter analysierte Funktion, sondern als eine differenzierte, durch Synkretismen verschiedener Art zerlegbare Funktion.

Diese Definition erlaubt es, Ähnlichkeiten zwischen Paradigmen festzustellen, Klassen zu bilden und dadurch Hierarchien von verwandten Spezialfällen der Synkretismusrelation zu definieren. Diese Hierarchie kann als Typenhierarchie, aber auch (unter zusätzlichen Annahmen über ‚Normalabbildungen') als Defaulthierarchie dargestellt werden, und kann nun aufgrund einer Charakterisierung der Synkretismusrelationen differenzierter formuliert werden als in bisherigen hierarchischen Ansätzen zur Beschreibung der deutschen Flexionsmorphologie. Insbesondere kann für die flektierenden Wortarten zwischen folgenden Paradigmenhierarchien im Deutschen unterschieden werden:

- Hierarchie der Stammsynkretismen (einschl. prosodische Operationen),
- Hierarchie der Affixsynkretismen (vor allem für Suffixe),
- Hierarchie der Lexemsynkretismen (für suppletive Alternationen).

3 Synkretismus im Deutschen

Eine Darstellung aller Spezialfälle der Synkretismusrelationen für die deutsche Flexion wäre sehr umfangreich und würde den gegebenen Rahmen sprengen. Deshalb wird hier exemplarisch nur der Affixsynkretismus der Substantive an einem

Tabelle 1: Paradigma der Vollform ‚Kind'.

	Numerus:	*Singular*	*Plural*
Kasus:	*Nominativ*	Kind	Kind#+er
		kˈɪnt	kˈɪnd#+ɐ
	Akkusativ	Kind	Kind#+er
		kˈɪnt	kˈɪnd#+ɐ
	Genitiv	Kind#+es	Kind#+er
		kˈɪnd#+əs	kˈɪnd#+ɐ
	Dativ	Kind#+e	Kind#+er+n
		kˈɪnd#+ə	kˈɪnd#+ɐ+n

Fallbeispiel erläutert. In der in Abschnitt 5 beschriebenen Anwendung sind die flexionsmorphologischen Synkretismen des Deutschen aber vollständig erfaßt.

Die Beschreibung des Synkretismus wird anhand der charakteristischen lexikalischen Eigenschaften, die Paradigmenklassen zugewiesen werden, vorgenommen. Beispielsweise wird dem Substantiv *Kind*, das auch als Prototyp für die Bezeichnung der Klasse verwandter Paradigmen genutzt wird, folgendes Paradigma[4] (s. Tabelle 1) zugewiesen.

Aus dem Paradigma können die orthographischen und phonologischen Affixmengen[5] extrahiert werden:

OrthSet: {∅, *es, e, er, er+n*}
PhonSet: {∅, əs, ə, ɐ, ɐ+n}

Flexionskategorien, die auf einzelne Endungen abgebildet werden, werden traditionell als Disjunktionen und Negationen dargestellt und als Regeln, *constraints* oder ‚Archikategorien' formuliert, dargestellt im nachfolgenden Beispiel durch Boole'sche Ausdrücke:

1. NomAkkSing $= (sing \wedge \neg(dat \vee gen))$
2. GenSing $= (sing \wedge gen)$
3. DatSing $= (sing \wedge dat)$

[4]Die nicht–traditionelle Reihenfolge der Kasus verdeutlicht einige der Synkretismen; die phonologischen Transkriptionen folgen aus praktischen Gründen einer morphologischen Erweiterung der IPA–Konventionen.

[5]Auf die Generalisierung über die agglutinativen Formen *‚er+n"* bzw. */ɐ+n/* im Dativ Plural wird nicht weiter eingegangen; die Wahl des ‚Schwa–Genitivs' bzw. des ‚Schwa–Dativs' als Basisform ist hier ebenfalls nicht ausschlaggebend, beruht aber auf einer formal begründeten Präferenz für Elision gegenüber Epenthese.

Tabelle 2: *Exponence* und *Referral* bei der Klasse ‚Affix_KIND'.

AffixKlasse:	Affix_KIND	
OrthVektor:	NomAkkSing:	[1] ∅
	GenSing:	[2] es
	DatSing:	[3] e
	NonDatPlur:	[4] er
	DatPlur:	[5] er+n
PhonVektor:	NomAkkSing:	[6] ∅
	GenSing:	[7] əs
	DatSing:	[8] ə
	NonDatPlur:	[9] ɐ
	DatPlur:	[10] ɐ+n
SyncMap:	NomSing:	Orth: [1], Phon: [6]
	AkkSing:	Orth: [1], Phon: [6]
	GenSing:	Orth: [2], Phon: [7]
	DatSing:	Orth: [3], Phon: [8]
	NomPlur:	Orth: [4], Phon: [9]
	AkkPlur:	Orth: [4], Phon: [9]
	GenPlur:	Orth: [4], Phon: [9]
	DatPlur:	Orth: [5], Phon: [10]

4. NonDatPlur = $(plur \wedge \neg dat)$

5. DatPlur = $(plur \wedge dat)$

Paradigmen und Paradigmenklassen werden durch charakteristische Synkretismen (‚SyncMap') und Abbildungen (‚OrthVektor', ‚PhonVektor') auf Affixe, Stammoperationen oder ganze Wörter definiert. Diese Komponenten sind in einem typologisch orientierten Modell klar zu trennen. Die Abbildungen der Klasse ‚Affix_KIND'[6] werden beispielhaft in Tabelle 2 als Attribut–Wert–Struktur mit *structure–sharing* formuliert.

Für Verkettungs– und Sandhi–Operationen (z.B. Auslautverhärtung, wie /kˈɪnd#+ɐ/ – /kˈɪnt/) ist eine automatenbasierte Funktion definiert, die aber hier nicht weiter relevant ist. Eine Erweiterung analog zur *form–based* (im Gegensatz zur *matrix–based*) Variante des Ansatzes von Krieger und Nerbonne (1993) erscheint möglich; dort wurden jedoch weder eine vergleichbare Lemma–Klassen–Unterscheidung, noch ein explizites Modell der Synkretismusrelation, noch eine vollständige Erfassung der Flexion des gesprochenen Deutsch anvisiert.

[6]Paradigmenklassen werden mit Typ und Stamm gekennzeichnet: ‚Affix_KIND', ‚StammOp_GOTT'.

Nach Herausfaktorisierung der morphographischen und morphophonologischen Sandhi–Bedingungen sind orthographische und phonologische Abbildungen oft (nicht notwendigerweise) gleich strukturiert. Ein weitergehendes Generalisierungskonzept (z.B. eine Vererbungshierarchie) kann also Vorteile bringen (vgl. Langer und Gibbon, 1992). Eine alternative Erweiterung im Sinne der DATR–Theorie von Bleiching (1994) erscheint also auch möglich. Die einschlägige Default–Hierarchie der regulärsten Paradigmenklassen (stark irreguläre Fälle und Fremdplurale sind hier nicht eingeschlossen) wird in Abbildung 1 wiedergegeben.

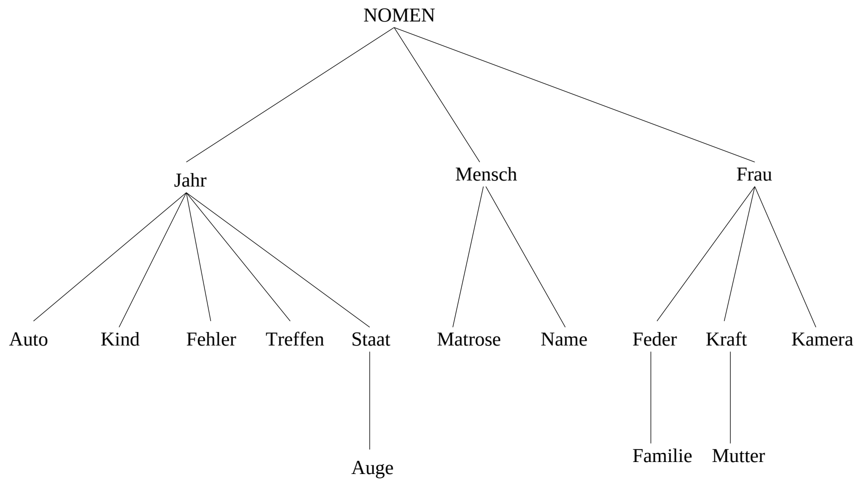

Abbildung 1: Default-Hierarchie für Affixsynkretismen im Deutschen.

4 Ein Modell zur Interpretation von Synkretismustheorien

Als Referenz für die Interpretation von Synkretismustheorien wird als Ergebnis ein explizites Modell über $< L, M, X, P, \Pi >$ (L = Lemmamenge, M = morphosyntaktische Kategorienmenge, $X \subseteq M$ = synkretismusspezifische morphologische Kategorienmenge, P = Inventar von Formen und Operationen über Formen, Π = Menge von Funktionen) im folgenden beschrieben.

1. L ist die Menge der Lemmata (mindestens die Simplexlemmata und Ableitungssuffixe; Komplexlemmata ‚erben' die Paradigmen des lexikalischen Kopfes);

2. $X \subseteq M$ sind Mengen von morphosyntaktischen Kategorien: M ist die Menge der Flexionskategorien, (z.B. $[NomAkkSing : Orth : x, \ldots, DatPlur :$

$Phon : y]$), X ist die für einen spezifischen Synkretismus charakteristische Menge von ,Archikategorien', die auf die in einem Lexem vorkommenden Affix–, Stamm– oder Lexemvarianten, umkehrbar eindeutig abgebildet werden (s. ,OrthVektor' und ,PhonVektor' in Tabelle 2);

3. $P = \{F_{orth}, F_{phon}, A_{orth}, A_{phon}, S_{orth}, S_{phon}\}$ ist eine Partition, wobei für $U = F\ oder\ A\ oder\ S : U_{orth}, U_{phon}$ (F = modalitätsspezifische Vollformen, A = Affixe, S = Stämme) und die Relation $U_{orth} \times U_{phon}$ (Paare aus komplexen morphologischen Repräsentationen in beiden Modalitäten) durch partiell kompositionelle Korrespondenzregeln definiert wird;

4. Es gilt die Konvention $f(U) =_{def} \{f(x)|x \in U\}$, f ist eine Funktion und U eine Teilmenge des Definitionsbereichs. Die Subskripte _p, _s (*Präfix*, *Suffix*) werden zur Bezeichnung von Partitionen der subskriptlos bezeichneten Affixmengen in Präfix– und Suffixmengen verwendet:
   $\Pi = \{$
   $\pi : L \times M \to F_{orth} \times F_{phon}$, (Suppletion)
   $\alpha_{syn\_p} : L \times X \to A_{orth} \times A_{phon}$, (Präfixalternation)
   $\alpha_{syn\_s} : L \times X \to A_{orth} \times A_{phon}$, (Suffixalternation)
   $\omega_{syn} : L \times X \to S_{orth} \times S_{phon}$, (Stammoperationen)
   $\lambda_{syn\_p} : L \times X \to \alpha_{syn\_p}(L, X)^\frown\omega_{syn}(L, X)$, (Präfixsynkretismus)
   $\lambda_{syn\_s} : L \times X \to \omega_{syn}(L, X)^\frown\alpha_{syn\_s}(L, X)$, (Suffixsynkretismus)
   $\alpha_{synpar\_p} : L \times M \times X \to A_{orth} \times A_{phon}$, (Präfixklassen)
   $\alpha_{synpar\_s} : L \times M \times X \to A_{orth} \times A_{phon}$, (Suffixklassen)
   $\omega_{synpar} : L \times M \times X \to S_{orth} \times S_{phon}$, (Stammoperationen)
   $\lambda_{synpar\_p} : L \times M \times X \to \alpha_{synpar\_p}(L, M, X)^\frown\omega_{synpar}(L, M, X)$, (Präfigierung)
   $\lambda_{synpar\_s} : L \times M \times X \to \omega_{synpar}(L, M, X)^\frown\alpha_{synpar\_s}(L, M, X)$ (Suffigierung)
   $\}$

5. $F_{orth} \times F_{phon} \subseteq ((A_{orth} \times A_{phon})^\frown(S_{orth} \times S_{phon})) \cup ((S_{orth} \times S_{phon})^\frown(A_{orth} \times A_{phon}))$
   Die folgende Verkettungskonvention wird verwendet:
   $\alpha^\frown\beta$, α und β sind Mengen, ist die Menge aller $x^\frown y$ für $x \in \alpha, y \in \beta$.

Die Funktionen werden folgendermaßen verstanden:

1. Lexemsynkretismus: π, d.h. die holistische Paradigmenfunktion (*paradigm function*) für suppletive Vollformen und als ,Morphologie–Ersatz' in älteren Vollformentheorien;

2. Synkretismusfunktionen (s. ,OrthVektor' und ,PhonVektor' in Tabelle 2): α_{syn_p} (Präfixsynkretismusfunktion); α_{syn_s} (Suffixsynkretismusfunktion);

ω_{syn} (Stammsynkretismusfunktion); λ_{syn_p} (Synkretismusfunktion für Präfigierung); λ_{syn_s} (Synkretismusfunktion für Suffigierung);

3. Synkretistische Paradigmenfunktionen (s. ‚SyncMap' in Tabelle 2): α_{synpar_p} (für Präfixe); α_{synpar_s} (für Suffixe); ω_{synpar} (für Stämme); λ_{synpar_p} (für Präfigierung); λ_{synpar_s} (für Suffigierung).

Auf der Basis dieser Definitionen werden synkretismusbasierte Paradigmenhierarchien als Relationen (partielle Ordnungen: Typ– bzw. Defaultverbände) über die Mengen $\alpha_{synpar}(L, M, X)$, $\omega_{synpar}(L, M, X)$, $\lambda_{synpar}(L, M, X)$ definiert.

Eine Typologie der Synkretismusrelationen kann durch *constraints* über die Mengen X, A, S im Modell definiert werden; auf diese Spezifizierung und deren Anwendung auf mehrere typologisch unterschiedliche Sprachen kann aber hier nicht weiter eingegangen werden. Es ist auch klar, daß für eine spezifische deskriptive Theorie weitere *constraints*, z.B. über die Relation zwischen orthographischen und phonologischen Einheiten, definiert werden müssen. Das Ziel ist aber an dieser Stelle, in erster Linie die Entwicklung eines Modells für synkretismusbasierte Theorien der deutschen Flexionsmorphologie und in zweiter Linie generische Datenstrukturen für Morphologiewerkzeuge zu definieren.

Es liegt nahe, eine Synkretismustheorie, die durch diese denotationelle Semantik interpretiert werden kann, als ein lemmabasiertes Lexikon mit Merkmalsstrukturen (s. Tabelle 2) im Rahmen einer Typ– oder Default–Vererbungshierarchie (s. z.B. Krieger und Nerbonne, 1993; Gibbon, 1992) zu formulieren.

Zwei Theorien wurden in unterschiedlichen Formalismen auf der Grundlage dieses Modells erstellt, die mit verschiedenen prozeduralen Eigenschaften ausgestattet sind und über das Synkretismusmodell hinaus mit einer morphographischen und morphophonologischen Sandhi–Komponente versehen sind:

1. *Theorie 1:* Die morphosyntaktischen Kategorien wurden mit Attribut–Wert–Strukturen dargestellt und die Lemma– und Klassendefinition erfolgte mit Default–Vererbung. Die Theorie wurde mit einer funktionalen Abfrage–Antwort–Semantik im Sinne des Synkretismus für Deutsch in DATR implementiert.

2. *Theorie 2:* Die morphosyntaktischen Eigenschaften wurden als Terme von Prädikaten über Lemmata und Klassen definiert und in einem Implikationssystem in Prolog implementiert. Von dieser Implementierung wurden zwei Versionen erstellt:

   (a) *Theorie 2a:* eine relationale Version, die flexible Anfrage–Antwort–Kombinationen erlaubt,

(b) *Theorie 2b:* eine funktionale Version im Sinne des Synkretismusmodells für Deutsch, die große durch die Theorie definierte ,virtuelle Wortschätze' effizient auskompiliert.

Weitere Theorien oder weitere prozedurale Semantiken für die Theorien können definiert werden (z.B. bei *Theorie 2a*: inverse Anfragemodi mit möglicherweise mehrdeutigen Antworten). Im Modell entsprechen solche Anfragemodi den folgenden inversen Funktionen von Formen auf Potenzmengen $\mathcal{P}$ von Lemma–Kategorie–Paaren[7], z.B. für Vollformen:

1. $\pi^{-1}(F_{orth} \times F_{phon}) = \pi^{-1}((S_{orth} \times A_{orth})^\frown(S_{phon} \times A_{phon})) = \mathcal{P}(L \times M)$
2. $\lambda^{-1}_{synpar}(F_{orth} \times F_{phon}) = \lambda^{-1}_{synpar}((S_{orth} \times A_{orth})^\frown(S_{phon} \times A_{phon})) = \mathcal{P}(L \times M \times X)$

Diese inversen Funktionen entsprechen lexikalischen Zuordnungsaufgaben bzw. morphologischen Analyseaufgaben. Für die vorgesehene Anwendung wurden diese Modi aber nicht benötigt.

5 Anwendung: Generierung von Aussprachelexika

Eine zentrale Aufgabe bei der Verarbeitung gesprochener Sprache ist die Erstellung eines Aussprachelexikons, in der Regel als Tabelle flektierter orthographischer und phonemischer Wortformen, die in den Trainings– und Testcorpora für Spracherkenner vorkommen. Zur Behandlung nicht belegter Flexionsformen müssen die Stämme der belegten Formen auf alle Flexionsformen projiziert werden (im vorliegenden Corpus eine Größenrelation von ca. vier bis fünf).

Zur Lösung dieser Aufgabe wurde die oben beschriebene *Theorie 2b* gewählt und zur Erstellung einer Wissensbasis über die *paradigm function* des Deutschen verwendet. Ziel war es, die Flexionsmorphologie zu komprimieren und Orthographie–Phonologiepaare von flektierten Formen mit ihren Kategorien automatisch und konsistent zu generieren, damit die Wissensbasis lediglich linear mit Anzahl und Umfang der Stämme wächst, nicht mit Anzahl und Umfang der dazugehörigen Vollformen. Die Synkretismusklassen wurden in einer Vererbungshierarchie geordnet.

Zur Zeit umfaßt die Wissensbasis 6141 Stämme aus Corpora gesprochener Sprache (ca. 14000 Dialogbeiträge bzw. *turns* zu Terminabsprachen). Zur Operationalisierung der *paradigm function*, wie sie in der denotationellen Semantik

[7]Wegen der möglichen Mehrdeutigkeiten wird auf Potenzmengen abgebildet.

expliziert wird, wurde ein Generator entworfen und in Quintus–Prolog implementiert. Die Wissensbasis definiert damit ein ‚virtuelles Vollformenlexikon', das mit dem Generator *on demand* oder auskompiliert verwendet werden kann. Bei den 6141 definierten Stämmen werden vom Generator 27690 Vollformen erzeugt, die nach Auflösung der Synkretismen 120598 morphosyntaktische Abbildungen auf Vollformen ergeben. Zur Zeit wird dieses Lexikon in mehreren Forschungslabors zur Entwicklung von Spracherkennungssystemen für Deutsch eingesetzt.

6 Zusammenfassung und Ausblick

Zur Behandlung des Synkretismusproblems als Komponente einer *paradigm function* für gesprochenes Deutsch wurde zunächst eine Synthese bisheriger typologischer Ansätze zu einzelnen Aspekten des Synkretismus erarbeitet, erweitert und als Grundlage für ein Synkretismusmodell und eine operationalisierte Synkretismustheorie genutzt. Aus empirischen Gründen lag der Schwerpunkt auf dem Modell als Grundlage für Theoriebildung und Implementierungen. Die Anwendbarkeit des Modells auf typologisch andersartige *paradigm functions* und Synkretismen als die des Deutschen muß noch untersucht werden.

Die am Anfang aufgestellte Arbeitshypothese, daß formal adäquate Morphologietheorien und deren effiziente Verarbeitung auch in Anwendungskontexten auf sprachtypologisch angemessene Modelle bezogen werden müssen, wurde, wenn nicht ‚bewiesen', so doch gestützt vom Ergebnis der Anwendung zur Erzeugung großer Aussprachelexika. Neu an dieser Anwendung ist auch die Tatsache, daß sie die beiden Modalitäten des geschriebenen und des gesprochenen Deutsch erfaßt.

Darüber hinaus stellt die Anwendung eine ungewöhnliche, aber fruchtbare Kooperation wissensbasierter und stochastischer Methoden in der Spracherkennung dar: *offline* werden neue wissensbasierte Methoden eingesetzt, um einen Teil der Infrastruktur für die Entwicklung effizienter stochastischer Systeme bereitzustellen.

Literatur

D. Bleiching (1994). Integration von Morphophonologie und Prosodie in ein hierarchisches Lexikon. In: H. Trost, Hg., *KONVENS 94, Wien*, S. 32–41, Berlin. Springer.

L. Cahill (1993). Morphonology in the lexicon. In: *Sixth Conf. of the European Chapter of the Assoc. for Computational Linguistics*, S. 87–96, Utrecht.

A. Carstairs (1987). *Allomorphy in Inflexion.* Croom Helm, London.

A. Carstairs-McCarthy (1992). *Current Morphology.* Routledge, London.

G. Corbett (1991). *Gender.* Cambridge University Press, Cambridge.

R. Evans und G. Gazdar (in press). DATR: A language for lexical knowledge representation.

D. Gibbon (1990). Prosodic association by template inheritance. In: W. Daelemans und G. Gazdar, Hg., *Proc. of the Workshop on Inheritance in Natural Language Processing*, S. 65–81, Tilburg. Institute for Language Technology.

D. Gibbon (1992). ILEX: A linguistic approach to computational lexica. In: U. Klenk, Hg., *Computatio Linguae*, Nummer 73 in Beiheft, S. 32–53, Stuttgart. Zeitschrift für Dialektologie und Linguistik, Franz Steiner.

L. Karttunen, K. Koskenniemi und R. Kaplan (1987). TWOL: A compiler for two-level phonological rules. In: R. Kaplan und L. Karttunen, Hg., *Computational Morphology.* Palo Alto Research Center, Stanford University.

M. Kay (1987). Nonconcatenative finite-state morphology. In: *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, Denmark. U Copenhagen.

K. Koskenniemi (1983). *Two-Level Morphology: A General Computational Model for Word Form Recognition and Production.* Dissertation, Universität Helsinki.

H.-U. Krieger und J. Nerbonne (1993). Feature-based inheritance networks for computational lexicons. In: T. Briscoe, V. de Paiva und A. Copestake, Hg., *Inheritance, Defaults, and the Lexicon*, S. 90–136, Cambridge. Cambridge University Press.

H. Langer und D. Gibbon (1992). DATR as a graph representation language for ILEX speech oriented lexica. Verbundprojekt ASL-TR-43-92/UBI, Universität Bielefeld.

P. Matthews (1991). *Morphology.* Cambridge University Press, Cambridge. 2nd edition.

S. Reinhard und D. Gibbon (1991). Prosodic inheritance and morphological generalisations. In: *Fifth Conference of the European Chapter of the Association for Computational Linguistics*, S. 131–136, Berlin.

G. Ritchie, G. Russell, A. Black und S. Pulman (1992). *Computational Morphology.* MIT Press, Cambridge, MA.

A. Spencer (1991). *Morphological Theory.* Blackwell, Oxford.

R. Sproat (1992). *Morphology and Computation.* MIT Press, Cambridge, MA.

G. Stump (1991). A paradigm-based theory of morphosemantic mismatches. *Language* 67: 675–725.

G. Stump (1993). On rules of referral. *Language* 69: 449–479.

H. Trost (1993). Coping with derivation in a morphological component. In: *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, S. 368–376, Utrecht.

A. Zwicky (1985). How to describe inflection. *Berkeley Linguistics Society* 13: 714–733.