Morphologiesystem MORPHY[1]

Wolfgang Lezius
Universität Paderborn
Fachbereich 2 – Psychologie
Arbeitsgruppe Kognitionsforschung

1 Konzeptionelle Kriterien

1.1 Vorbemerkungen

Für jede Wortklasse wurde eine eigene Datenstruktur entworfen. Daher gibt es für jede Wortklasse ein eigenes Unterlexikon und einen eigenen Algorithmus für die Lemmatisierung und Generierung (siehe auch Abbildung 1). Das Verblexikon ist zudem in zwei Unterlexika für schwache und nicht-schwache Verben unterteilt. Die verwendeten Regeln sind direkt im Pascal-Programm implementiert und können vom Benutzer nicht geändert werden.

Substantive
Adjektive
schwache Verben
nicht- schwache Verben
Eigennamen
Sonstige

Abbildung 1: Die Struktur des Lexikons

1.2 Deklarative Spezifikation lexikalischer Einträge

Da die Datenstrukturen durch die Bildung von Klassen zum Teil sehr komplex sind, wird hier stellvertretend die Datenstruktur für Substantive erläutert.

Wortstamm: Klasse: ss/ß-Wechsel: Pluralbildung:

Abbildung 2 : Die Datenstruktur Substantiv

Jedem Stamm wird eine Klasse zugeordnet. Durch die Klasse sind alle Flexionsendungen, der Genus und die evtl. Umlautung festgelegt. Es folgt ein Ausschnitt aus der Endungstabelle der insgesamt 62 Klassen; die komplette Tabelle befindet sich in Anhang A.

Nr.	Singularendungen				Pluralendungen				Gen.	Umlaut.	Bsp.
	Nom	Gen	Dat	Akk	Nom	Gen	Dat	Akk			
4	-	s/es	e/-	-	e	e	en	e	m	ja	Kampf
8	-	s/es	e/-	-	er	er	ern	er	m	ja	Mann
11	-	en	en	en	en	en	en	en	m	nein	Mast
24	-	s/es	e/-	-	er	er	ern	er	n	nein	Kind
37	-	-	-	-	en	en	en	en	f	nein	Frau
48	um	ums	um	um	en	en	en	en	n	nein	Alb(um)

[1]Dieser Artikel ist erschienen in: Hausser, R. (Hg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994.* Niemeyer, Tübingen, 1996.

Für die eindeutige Festlegung der Flexionsformen wird noch die Information benötigt, ob ein ß-ss-Wechsel vorliegt (z.B. 'Kuß'–'Kusses'). Desweiteren kann es vorkommen, daß ein Substantiv im Plural nicht flektiert wird (z.B. 'Laub'). Dies wird durch die entsprechende Information angezeigt.

Wortstamm: Kuß Klasse: 4 ss/ß-Wechsel: JA Pluralbildung: JA

Abbildung 3 : Beispiel: Lexikoneintrag für das Substantiv 'Kuß'

1.3 Bezug zwischen lexikalischen Einträgen und Wortformen

Um eine Wortform zu lemmatisieren, wird in einem ersten Schritt versucht, auf alle möglichen Wortstämme zu schließen. Sofern ein Kandidat im Lexikon eingetragen ist, kann durch die in 1.1 erläuterte Datenstruktur für diesen Stamm im 2. Schritt das Flexionsparadigma generiert werden. Es bleibt schließlich der Vergleich der Wortform mit allen Flexionsformen. Anhand eines Beispiels wird diese Vorgehensweise deutlich.

Beispiel: Substantivform 'Tisches'

1. Durch Abschneiden aller möglichen Endungen werden die Kandidaten für den Wortstamm gebildet (für Wortformen wie 'Küsse' werden zusätzlich noch Umlautungsprozesse und ß-ss-Wechsel zurückgebildet):

   Tisches - / Tische - s / Tisch - es / Tisc - hes / Tis - ches / Ti - sches

   Für diese Kandidaten wird überprüft, ob ein Eintrag im Lexikon vorliegt. Damit bleibt nur noch 'Tisch' als Stamm übrig.

2. Anschließend wird das Paradigma aus dem Lexikoneintrag generiert:

   Wortstamm: Tisch Klasse: 3 ss/ß-Wechsel: NEIN Pluralbildung: JA

   Abbildung 4 : Lexikoneintrag für das Substantiv 'Tisch'

   Paradigma: Singular: der Tisch, des Tisches, dem Tisch(e), den Tisch,
   Plural: die Tische, der Tische, den Tischen, die Tische

3. Als Markierung bleibt somit der Genitiv Singular von 'Tisch' übrig.

1.4 Verständlichkeit und linguistische Motivation der Regeln

a) Flexion
   Ähnlich wie für Substantive wurden für schwache Verben, Adjektive und Eigennamen Klassifizierungen vorgenommen. Lediglich für nicht-schwache Verben werden 7 markante Formen gespeichert, die die Flexionsgenerierung in eindeutiger Weise festlegen. Das Vorgehen bei der Analyse ist dem bei Substantiven ähnlich, jedoch viel komplexer (z.B. durch eine Präfix-Suffix-Analyse bei Verbformen wie 'an-lächel-st').

b) Derivation
   Das Programm verzichtet auf die Behandlung von Derivationsprozessen.

c) Komposition
Die Komposition wird berücksichtigt durch die Analyse von Verbzusätzen und zusammengesetzten Substantiven. Da die Fugenbildung hier sehr unregelmäßig ist (z.B. 'Schwein-s-haxe', 'Schwein-e-braten', 'Schwein-kram'), möchte man gern auf die Entwicklung eines Algorithmus verzichten. Doch nehmen zusammengesetzte Substantive einen beachtlichen Anteil eines Textes ein. In den „vorläufigen Testdaten" lagen nach meiner Markierung immerhin bis zu 5% zusammengesetzte Substantive vor (siehe auch Testdatenergebnisse 2.2c). Der hier implementierte Algorithmus arbeitet nach dem „longest matching"–Verfahren. Es wird versucht, die längste von rechts abgeschnittene gültige Substantivform zu finden. Mit dem verbleibenden Rest wird dieser Vorgang solange wiederholt, bis man am Wortanfang angelangt ist. Eine Fuge wird damit als gültige Flexionsendung definiert. Die damit entstehenden Fehlinterpretationen ('Schwein-haxe' und 'Schwein-en-haxe' werden als gültige Formen definiert) werden in Kauf genommen. Einige Sonderfälle werden berücksichtigt: das Anhängen eines 's' bei femininen Substantiven (z.B. 'Heiterkeit-s-test'), das Auslassen eines Konsonanten bei 3-fachem Auftreten (z.B. 'Schif/f/fahrt') bzw. das Nicht-Auslassen (z.B. 'Sauerstoff-flasche').

Beispiel: 'Hausdächern'
Hausdächer-n
Hausdäche-rn
Hausdäch-ern
Hausdäc-hern
Hausdä-chern
Hausd-ächern
Haus-dächern
Hau-sdächern
Ha-usdächern
H-ausdächern
Haus
Hau-s
Ha-us
H-aus

Da am Wortanfang angelangt, wurde eine Form von 'Haus-Dach' gefunden.

1.5 Morpho-syntaktische Analyse / Kategorisierung

a) morpho-syntaktische Analyse
Da die Analyse kontextfrei ist, werden keine Zusammenhänge zwischen den Wörtern hergestellt. Eine Erweiterung des Programms in dieser Richtung ist jedoch möglich. Beispielsweise könnte im Satz „Er kauft es mir ab." der Zusammenhang von 'kaufen' und 'ab' hergestellt werden.

b) Kategorisierung
Das Merkmalssystem ist detailliert genug, um als Front-End für einen Syntax-Parser eingesetzt werden zu können. Es schlüsselt nach den traditionellen Kriterien wie Genus, Numerus, Kasus, Modus, Tempus, Person auf. Hier als Beispiel die Aufschlüsselung für Substantive; das vollständige System befindet sich in Anhang B.

```
SUB (Stamm) NOM MAS SIN
            GEN FEM PLU
            DAT NEU
            AKK
```

Die Kategorisierung hat sich im Einsatz bereits bewährt (siehe 2.9); durch eine entsprechende Reduktion bzw. Adaption kann sie für den gewünschten Einsatzzweck modifiziert werden — für die Textindexierung reicht ja beispielsweise der Wortstamm.

1.6 Generierung

Wie bereits in 1.1 und 1.2 erläutert, werden bei der Lemmatisierung Regeln für die Generierung benutzt. Das Programm kann daher durch die entsprechenden Algorithmen das Paradigma aus der jeweiligen Grundform eines Wortes generieren. Im Programm steht dazu die Funktion „Anschauungssystem“ zur Verfügung.

1.7 Übertragbarkeit auf andere Sprachen

Da das Programm speziell für das Deutsche entworfen wurde, entfällt dieser Punkt.

2 Technische Konzeption und Einsatzfähigkeit

2.1 Zielsetzung der Konzeption

Das Programm ist aus einer Semesterarbeit im Seminar „Einführung in die automatische Sprachanalyse“ von Prof. Wettler im WS 1991 entstanden. Es sollte dort als Front-End für einen Syntaxparser eingesetzt werden. Da als Hardware nur ein PC zur Verfügung stand, ist das Programm speziell für den PC entwickelt worden. Aus diesem Grunde stand die Minimierung des Speicherplatzbedarfs, insbesondere die Komprimierung des Lexikons, im Vordergrund.
Diese Anwendungsaspekte haben die Entwicklung des Programms stark beeinflußt. Es belegt als ausführbarer Code 500 KB Hauptspeicher; die 16500 im Lexikon eingetragenen Wortstämme benötigen 800 KB Festplattenspeicher, was nur etwa 50 byte/Wort bedeutet. Diese starke Komprimierung und die detaillierten Analyseergebnisse führen zu einer relativ niedrigen Geschwindigkeit von etwa 10 Wörtern/Sekunde.

2.2 Portabilität der Software und Daten

Als reines PC-Programm wurde die Software unter Borland Pascal entwickelt. Als Betriebssystem wird MS-DOS 5.0 oder höher benötigt. Eine Adaption für andere Betriebssysteme ist nicht geplant.

2.3 Schnittstellen zur Syntax und Semantik

In Abschnitt 1.3 wurde die Syntax-Schnittstelle bereits vorgestellt. Sie wird in Projekten unserer Arbeitsgruppe eingesetzt (siehe 2.9). Eine Semantik-Schnittstelle ist nicht implementiert.

2.4 Hilfestellung bei Benutzerfehlern

Fehler kann der Benutzer nur bei der Lexikonpflege machen. Dort kann er bei einer Fehleingabe den Eintrag eines Stammes jederzeit abbrechen. Auch läßt sich ein fehlerhaft eingetragener Stamm problemlos wieder löschen.

2.5 Größenbeschränkung des Systems

Die Lexikongröße wird nur durch den vorhandenen Festplattenspeicher begrenzt. Da das Lexikon stark komprimiert ist, kann man von einem beliebig großen Lexikon sprechen. Die Länge eines Wortstammes ist durch 20 Buchstaben begrenzt, die Länge einer Wortform somit durch Länge des Stamms (max. 20 Buchstaben) plus Länge der Endung (max. 4 Buchstaben). Zusammengesetzte Substantive dürfen bis zu 80 Buchstaben lang sein.

2.6 Schnittstelle zu Nicht-ASCII-Zeichen

Nicht-ASCII-Zeichen werden nicht unterstützt, dafür Nicht–ASCII–Texte. Die sehr verbreiteten Formate WordPerfect (DOS und Windows ab Version 5.0) und Word (DOS ab Version 4.0) werden in das ASCII-Format konvertiert und direkt lemmatisiert.

2.7 Benutzerfreundlichkeit des Turn-Around

a) linguistisch-empirische Modifikationen
Da die morphologischen Regeln nicht vom Benutzer geändert werden können, muß das Programm in einem solchen Falle neu compiliert werden. Abgesehen von der Änderung im Quellcode benötigt die Compilierung einige Sekunden.

b) unbekannte Wortformen
Auf unbekannte Wortformen reagiert das System robust. Aufgrund einer Endungsanalyse wird eine Wortklassenprognose gegeben, die in etwa 80% aller Fälle zutrifft.

c) Lexikonpflege
Bei der Eingabe eines neuen Wortes werden dem Benutzer umfangreiche Hilfen gegeben. Lediglich den Stamm muß der Benutzer selbst eintippen. Ein Expertensystem schlägt nun für markante Formen/Fälle eine Reihe von Wortformen vor, aus denen die richtige gewählt werden muß. Nach 2-5 Anfragen dieser Art (abhängig von der Wortklasse und dem Stamm) ist der Stamm klassifiziert.

Dialog beim Eintrag des schwachen Verbs 'segeln':

1. Geben Sie den Stamm ein: segeln
2. Wird das Verb schwach konjugiert ?
   1: Ja
   2: Nein
3. Wie lautet die 2. Person Singular Präsens ?
   1: du telefonierst
   2: du telefonierest
   3: du telefoniert

4. Wie lautet das Partizip des Verbs ?
   1: telefoniert
   2: getelefoniert
   Verb klassifiziert !

Dialog beim Eintrag des Substantivs 'Mann' :

1. Geben Sie den Stamm ein: Mann
2. Welchen Genus hat das Substantiv ?
   1: maskulinum
   2: femininum
   3: neutrum
3. Muß bei der Pluralbildung umgelautet werden ?
   1: ja
   2: nein
   3: Plural ex. nicht
4. Wie lautet der Genitiv Singular ?
   1: Manns
   2: Manns/Mannes
   3: Mannens
5. Wie lautet der Dativ Plural ?
   1: Männen
   2: Männern
   Substantiv klassifiziert !

2.8 Transparenz und Vollständigkeit der Dokumentation

Für den Benutzer steht ein Bedienungshandbuch zur Verfügung, das im Januar 94 zuletzt überarbeitet wurde. Dort werden alle Programmfunktionen ausführlich beschrieben. Auf der Grundlage des Handbuchs ist ein Testbericht entstanden, der im Rahmen eines Artikels über Shareware-Übersetzer in der Computer-Fachzeitschrift „DOS International“ in der Ausgabe 10/93 veröffentlicht wurde.

2.9 Verfügbarkeit und Wartung

Universitäten und wissenschaftlichen Instituten wird das Programm zur rein wissenschaftlichen Nutzung kostenlos zur Verfügung gestellt. Es wird zur Zeit von 12 Universitäten benutzt. Im Falle von bugs können sich die Anwender an den Programmautor wenden.

Im Rahmen der Arbeitsgruppe „Kognitionsforschung“ wird das Programm als eine Komponente für einen Tagging-Algorithmus für das Deutsche eingesetzt. Zudem wird es als Grundlage für einen Textindexierer benutzt.

Anhang A: Klassensystem für Substantive

Nr.	Singularendungen				Pluralendungen				Gen.	Umlaut.
	Nom	Gen	Dat	Akk	Nom	Gen	Dat	Akk		
1	-	s	-	-	-	-	n	-	m	-
2	-	s	-	-	-	-	n	-	m	u
3	-	s/es	e/-	-	e	e	en	e	m	-
4	-	s/es	e/-	-	e	e	en	e	m	u
5	-	s	-	-	-	-	-	-	m	-
6	-	s	-	-	-	-	-	-	m	u
7	-	s/es	e/-	-	er	er	ern	er	m	-
8	-	s/es	e/-	-	er	er	ern	er	m	u
9	-	s/es	e/-	-	en	en	en	en	m	-
10	-	s	-	-	n	n	n	n	m	-
11	-	en	en	en	en	en	en	en	m	-
12	-	n	n	n	n	n	n	n	m	-
13	e	en	en	en	en	en	en	en	m	-
14	e	ens	en	en	en	en	en	en	m	-
15	en	ens	en	en	en	en	en	en	m	-
16	en	ens	en	en	en	en	en	en	m	u
17	-	s	-	-	s	s	s	s	m	-
18	-	ses	-	-	se	se	sen	se	m	-
19	-	s	-	-	-	-	n	-	n	-
20	-	s	-	-	-	-	n	-	n	u
21	-	s/es	e/-	-	e	e	en	e	n	-
22	-	s/es	e/-	-	e	e	en	e	n	u
23	-	s	-	-	-	-	-	-	n	-
24	-	s/es	e/-	-	er	er	ern	er	n	-
25	-	s/es	e/-	-	er	er	ern	er	n	u
26	-	s/es	e/-	-	en	en	en	en	n	-
27	en	ens	en	en	en	en	en	en	n	-
28	-	s	-	-	s	s	s	s	n	-
29	e	es	e	e	e	e	en	e	n	-
30	e	es	e	e	en	en	en	en	n	-
31	n	ns	n	n	n	n	n	n	n	-
32	-	ens	en	-	en	en	en	en	n	-
33	-	ses	-/se	-	se	se	sen	se	n	-
34	-	-	-	-	e	e	en	e	f	-
35	-	-	-	-	e	e	en	e	f	u
36	-	-	-	-	-	-	n	-	f	u
37	-	-	-	-	en	en	en	en	f	-
38	-	-	-	-	n	n	n	n	f	-
39	e	e	e	e	en	en	en	en	f	-
40	-	-	-	-	s	s	s	s	f	-
41	-	-	-	-	se	se	sen	se	f	-
42	-	-	-	-	nen	nen	nen	nen	f	-
43	-	-	-	-	-	-	-	-	f	-
44	-	-	-	-	-	-	-	-	m	-
45	-	s	-	-	ien	ien	ien	ien	n	-
46	a	a	a	a	en	en	en	en	f	-
47	a	as	a	a	en	en	en	en	n	-
48	um	ums	um	um	en	en	en	en	n	-
49	um	ums	um	um	a	a	a	a	n	-
50	is	is	is	is	en	en	en	en	f	-
51	on	ons	on	on	a	a	a	a	n	-
52	us	us	us	us	ora	ora	ora	ora	m	-
53	us	us	us	us	een	een	een	een	m	-
54	o	os	o	o	en	en	en	en	n	-
55	us	us	us	us	en	en	en	en	m	-
56	os	os	os	os	en	en	en	en	n	-
57	x	x	x	x	zen	zen	zen	zen	f	-
58	s	s	s	s	nten	nten	nten	nten	m	-
59	-	s	-	-	nen	nen	nen	nen	n	-
60	s	s	s	s	en	en	en	en	m	-
61	-	-	-	-	-	-	-	-	n	-
62	-	s	-	-	ten	ten	ten	ten	m	-

Anhang B: Merkmalssystem

Substantive:

```
SUB (Grundform) [INF] NOM MAS SIN
                      GEN FEM PLU
                      DAT NEU
                      AKK NOG
```

Eigennamen:

```
EIG VOR SUB (Grundform) NOM MAS ART
    NAC                 GEN FEM NOA
    PER                 DAT NEU
    STD                 AKK
    COU
    WAT
    GEB
    MOU
    GEO
```

Nominalkomposita:

```
KMP [Grundform] (Komponenten) SUB (Grundform des rechtesten Lemmas)  ...
                              - weiter s.o. -
                              EIG VOR SUB (Grundform des rechtesten Lemmas) ...
                              - weiter s.o. -
```

Verben:

```
VER AUX Infinitiv SFT INF         [NEB]
    MOD           NON PA1
                      PA2
                      EIZ
                      IMP SIN
                          PLU
                      1   SIN PRAE
                      2   PLU PRT
                      3       KOJ
                              KJ2
```

Adjektive:

```
ADJ(Grundform) [PA0 Partizip-Grundform] PRD GRU
                                            KOM
                                            SUP
                                        ATT SOL GRU MAS NOM SIN
                                            IND KOM FEM GEN PLU
                                            DEF SUP NEU DAT
                                                    AKK
```

Artikel:

```
ART DEF NOM SIN MAS
    IND GEN PLU FEM
        DAT     NEU
        AKK
```

Partikel:

```
PAR ADV LOK NOM
    PRP TMP GEN
        MOD DAT
        CAU AKK
        INR     (nur bei PRP, dort jedoch nicht verpflichtend)
        PRO
PAR KON
    NEG
```

Interjektion:

```
INJ
```

Pronomen:

```
PRO RIN NOM MAS
        GEN FEM
        DAT NEU
        AKK
    IND STV SIN 1
        BEG PLU 2A
        B/S     2B
                3
    DEM NOM SIN MAS
    REL GEN PLU FEM
    POS DAT     NEU
        AKK
    PER NOM SIN MAS 1
    REF GEN PLU FEM 2A
        DAT     NEU 2B
        AKK          3
    NEG NOM SIN MAS
        GEN PLU FEM
                NEU
```

ohne Angaben : undeklinierbares Pronomen

Abkürzungen:

```
ABK (Abkuerzungsgrundform) (Wortkl., also SUB, VER, ADJ, ART, PAR, PRO, INJ)
```

Hinweis: In eckigen Klammern stehen jeweils Optionen, d.h. der dort zu machende Eintrag ist nicht verpflichtend.

Bedeutung der 3-Buchstaben-Abkürzungen:

ABK - Abkürzung
ADJ - Adjektiv
ADV - Adverb
AKK - Akkusativ
ART - Artikel bzw. mit Artikel
ATT - attributiver Gebrauch
AUX - Hilfsverb
CAU - kausal
COU - Land
DAT - Dativ
DEF - bestimmt
DEM - demonstrativ
EIG - Eigenname
EIZ - erweiterter Infinitiv mit zu
FEM - feminin
GEB - geogr. Gebiet
GEN - Genitiv
GEO - geogr. Eigenname (Sonstiges)
GRU - Grundform
IMP - Imperativ
IND - unbestimmt
INF - Infinitiv
INJ - Interjektion
INR - interrogativ
INT - intransitiv
LOK - lokal
KOM - komparativ
KON - Konjunktion
KMP - Kompositum
MAS - maskulin
MDD - modal
MOD - Modalverb
MOU - Gebirge
NAC - Nachname
NEB - Gebrauch nur im Nebensatz
NEG - Negation
NEU - neutrum
NIL - keine Form gefunden
NOA - ohne Artikel
NOG - ohne Genus
NOM - Nominativ
NON - nicht-schwach
PAR - Partikel
PA0 - Partizip
PA1 - Partizip 1
PA2 - Partizip 2
PER - personal bzw. Personenname
PLU - Plural

POS - possessiv
PRAE - Präsens
PRD - prädikativ
PRO - Pronomen bzw. pronominal
PRP - Präposition
PRT - Präteritum bzw. Imperfekt
REL - relativ
RIN - relativ oder interrogativ
SE - Satzendezeichen
SFT - schwach
SIN - singular
SOL - ohne Artikel
SOR - Source (morphologische Herkunft)
STD - Stadt oder Ort
SUB - Substantiv
SUP - Superlativ
SZ - Satzzeichen
TMP - temporal
TRN - transitiv
VER - Verb
VOR - Vorname
WAT - Gewässer
ZAL - Zahlwort
ZAN - Zahl (=Ziffernfolge)
1,2,3 (bei Verben) - Person bei Numerus
1,2A,2B,3 (bei Pronomen) - nähere Bestimmung eines Pronomen :

- 1 - von sich selbst sprechende Person
- 2 - angesprochene Person
- 2A - angesprochene Person, höflich
- 2B - angesprochene Person, vertraut
- 3 - über eine Person/Sache gesprochen