Tobias Brückner / Monika Kolvenbach

ÜBERLEGUNGEN ZU EINEM VERFAHREN "WORTSEGMENTIERUNG DEUTSCHER TEXTE" (WORTSEG)

0. Vorbemerkungen

Die vorliegende Untersuchung steht in engem Zusammenhang mit den Überlegungen, die am Institut für deutsche Sprache bezüglich eines Programms zur Segmentierung beliebiger deutscher Texte angestellt wurden.

Es wird versucht, durch Beispiele und abwägende Überlegungen den Segment-Begriff deutlich herauszuarbeiten, um klarzustellen, daß mit Segment nicht nur das maschinell Faßbare gemeint ist, sondern daß dieser Begriff auch einen linguistisch motivierten Hintergrund aufweist, der die Ergebnisse einer solchen Segmentierung für weiterführende wissenschaftliche Untersuchungen nützlich und wünschenswert erscheinen läßt.

1. Zur Definition von Segment

Im folgenden sprechen wir von Segment als einer sinnvollen, durch Regeln erkennbaren Einheit. Was ist hierbei unter "sinnvoll" und was unter "erkennbar" zu verstehen?

Ausgangsmaterial für den Schritt der Segmenterkennung - vom Einzelsegment aus gesehen - bzw. der Segmentierung - von der Eingabeeinheit her gesehen - sind die Wörter, so wie sie in Texten vorgefunden werden, rein formal gesehen das, was zwischen zwei Leerzeichen steht. Diese Einschränkung ist - besonders im Deutschen - aus den folgenden Gründen notwendig: Es gibt Worteinheiten, die syntaktischen Regeln zufolge an bestimmten Wortgrenzen getrennt werden und dann an unterschiedlichen, stellungsmäßig definierbaren Plätzen innerhalb des Satzgefüges vorkommen. Z.B. werden in Hauptsätzen in einfachen Zeitformen von präfigierten

Verben die Präfixe abgetrennt und stehen erst am Satzende, wie in:

Hannes geht leider im Januar weg.

In diesem Beispielsatz, wenn er Eingabe für die Segmenterkennung wäre, wären *geht* und *weg* nach unserer Definition zwei verschiedene Eingabeeinheiten bzw. Wörter, die zu segmentieren wären. Im entsprechenden Nebensatz

..., weil Hannes leider im Januar weggeht.

wäre *weggeht* nur eine zu segmentierende Einheit bzw. ein Wort.

Von dieser Ausnahme abgesehen, kann man jedoch auch im Deutschen davon ausgehen, daß das, was in Texten und Sätzen von zwei Leerzeichen - ohne Berücksichtigung der Satzzeichen - eingeschlossen ist, ganz naiv gesagt "ein Wort" ist. Und dieses Wort ist die Einheit, auf der das Segmentierungsprogramm als Eingabe arbeiten soll und aus der nun "sinnvoll" kleinere Einheiten, eben die Segmente erkannt werden müssen.

Eine Vorbedingung, die auch schon vor der eigentlichen Segmentdefinition aufgestellt werden kann, ist die, daß die vorgenommene Segmentierung der Eingabedatei vollständig erfolgen muß; denn nur eine vollständige, "sinnvolle" Segmentierung läßt den Schluß zu, daß diese Segmentierung auch in sich schlüssig ist. Was wird nun unter "sinnvoll" auf diesem speziellen Gebiet verstanden? Eine Möglichkeit, eine solche Segmentierung vorzunehmen, wäre etwa die, daß man jedes Eingabewort in zweibuchstabige Segmente zerlegen würde, um so Aufschluß über das Vorkommen von Buchstabenkombinationen zu erhalten. Eine solche Segmentierung wäre für unsere Zwecke nicht "sinnvoll". Hier soll nur das als Segment abgetrennt werden, was auch in anderen Wörtern wieder "sinnvoll" auftaucht. Hier könnte man nun einwenden, daß auch zweibuchstabige Buchstabenfolgen wiederum in an-

deren Wörtern auftauchen, und dies kann unter anderem Aspekt dann wiederum "sinnvoll" sein. Für uns sind jedoch nur die Segmente "sinnvoll" erkannt, die Bausteine sein können bei der Bildung anderer Wörter und die dort in dieser "Baustein"-Funktion - ebenfalls nach den gleichen Regeln verwendet - wiederkehren.

Wir haben nun nur noch zu klären, was wir unter "regelhaft" verstehen wollen, um dann die einzelnen Segmenttypen kurz zu beschreiben und an Beispielen das oben Gesagte zu verdeutlichen.

Jedem Segmenttyp und teilweise auch Einzelsegmenten kann in regelhafter Formulierung beigegeben werden, in welcher Umgebung dieser Baustein sinnvoll fungiert und welche Umgebungen nicht zulässig sind.

2. Zur Verwendung von Segment

Eine ausgezeichnete Gruppe von Segmenten sind die Basissegmente [1]. Zu der Gruppe der Basissegmente gehören die einsegmentigen Nomen, wie etwa *Haus, Schiff* oder *Kind*. Hierher gehören auch die Segmente, die entstehen, wenn man von unpräfigierten Verben das Rechtssegment (vgl. unten) *-en* abstreicht, wie etwa bei *spiel-en* oder *führ-en*. Hierher gehören ebenfalls Adjektiv-Simplicia wie etwa *klein, schön*.

Diese Basissegmente bilden das Kernstück des "Wortes", und an ihnen orientiert sich die Funktionendefinition der anderen Segmente.

Ausgehend von diesem Basissegment ist eine grobe Unterteilung nach Links- und Rechtssegmenten vorzunehmen, rein stellungsmäßig zum Basissegment bestimmt. Die Linkssegmente werden allgemein auch als Präfixe, die Rechtssegmente als Suffixe bezeichnet.

Diese Unterscheidung in Prä- und Suffixe erscheint uns jedoch nicht sehr glücklich. Wir würden neben der Unterscheidung in Links- und Rechtssegmente eine weitere Unterscheidung in Wortbildungssegmente und Flexive vornehmen. Unter Wortbildungssegmenten verstehen wir im folgenden Segmente, die die Fähigkeit besitzen, bestehende Wörter bezüglich entweder ihrer Wortklassenzugehörigkeit oder ihres Inhalts zu verändern. Hierzu einige Beispiele:

Änderung der Wortklassenzugehörigkeit:

zieh-en --→ *Zieh-ung* = Verb --→ Nomen

Sach-e --→ *säch-lich* = Nomen --→ Adjektiv

schön --→ *Schön-heit* = Adjektiv --→ Nomen

stark --→ *stärk-en* = Adjektiv --→ Verb

Änderung des Inhalts (welche semantischen Konsequenzen solche Änderungen haben und wie man sie semantisch beschreiben kann, soll hier völlig außer acht bleiben):

fass-en --→ *er-fass-en*

lauf-en --→ *aus-lauf-en*

klein --→ *klitze-klein*

Haus --→ *Toll-haus*

Alle diese Typen wollen wir weiterhin als Wortbildungssegmente bezeichnen, wobei wir uns völlig im klaren darüber sind, daß wir damit eine linguistisch eigentlich unvertretbare Vereinfachung vornehmen. Dies ist aber darin begründet, daß die Ergebnisse eines solchen Segmentierungsverfahrens dazu führen sollen, weitere Feinunterscheidungen und -gliederungen vorzunehmen, die jedoch durch eine Vorabdefinition nicht präjudiziert werden sollen.

---

1) Vgl. Krallmann, D.: Zur Untersuchung funktionaler Merkmale in Wortstrukturen, ersch. in: C.H. Heidrich, Morphologie II, Hamburg, 1981

Als Flexive bezeichnen wir im folgenden die Segmente, die eine Differenzierung des Wortes bezüglich Numerus, Kasus, Tempus oder Modus leisten und zwar nur dann, wenn zu dieser Unterscheidung ein eigenes Segment herangezogen wird. Die Unterscheidung bezüglich Genus klammern wir bei Nomen absichtlich aus, weil *-in* in *Spiel-er-in* bei uns als Wortbildungssegment aufgefaßt wird und sich die Genus-Unterscheidung zwischen *Spiel-er* und *Spiel-er-in* schon in dem unflektierten Wort, das als Eintrag in einem Grundformenlexikon erscheint, niederschlägt. Daß man im Fall der Verben von der Grundform (dem Infinitiv) jeweils ein Flexiv abtrennen muß, um das Basissegment zu erhalten, mag von einigen als inkonsequent erachtet werden, soll jedoch aus systematischen Erwägungen so festgelegt werden.

Flexive können sowohl als Links- als auch als Rechtssegmente auftreten. Sie stehen meist als Rechtssegmente, es gibt jedoch die folgenden Ausnahmen:

arbeit-en : *en* = Flexiv/ *ge-arbeit-et*: *ge* und *et* = Flexive

aus-arbeit-en : *aus* = Wortbildungssegment, *en* = Flexiv

aus-zu-arbeit-en : *aus* = s.o., *zu* und *en* = Flexive

Außer den Rechts- und Linkssegmenten unterscheiden wir noch die Mittelsegmente, die sowohl stellungsmäßig als auch funktional bestimmt sind. Diese von uns Mittelsegmente genannten Einheiten werden gemeinhin auch als Infixe bezeichnet. Sie stehen nur in Wörtern, die mehr als ein Basissegment enthalten, wobei sich um diese Basissegmente wiederum Wortbildungs- und Flexionssegmente gruppieren können. Sie stehen dann als Trenner an der Stelle, wo die Rechtssegmente des einen Basissegments aufhören und die Linkssegmente des nächsten Basissegments anfangen (es braucht sich nicht notwendig um nur zwei Basissegmente zu handeln).

Beispiele: 1. Beim Aufeinandertreffen von zwei Basissegmenten:
Kind-s-vater, Esel-s-brücke

2. Beim Aufeinandertreffen von Rechts- und Basissegment:
   Be-nutz-ung-s-recht, Keusch-heit-s-gür-tel
3. Beim Aufeinandertreffen von Basis- und Linkssegment:
   Arbeit-s-ver-trag
4. Beim Aufeinandertreffen von Rechts- und Linkssegment:
   Auf-lass-ung-s-be-scheid

Damit ergibt sich die folgende schematische Darstellung für Wörter mit einem Basissegment:

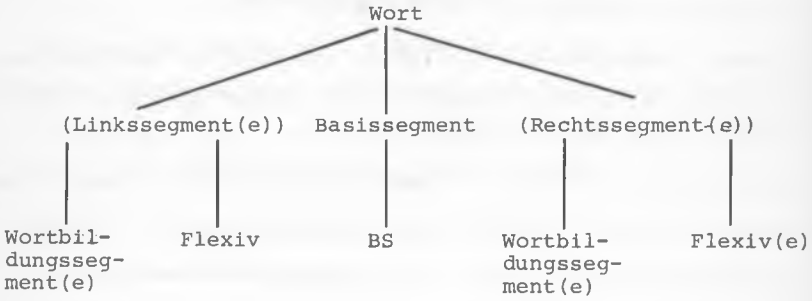


Abb. 1

Die nächste Abbildung zeigt die schematische Darstellung eines Wortes mit mehr als einem Basissegment:

Wort
BS1
BSn
LSn
LWn
LWn,m
LWn,m-1
LWn,1
LFn
LFn,1
BSn
BSn
BSn
RSn
RWn
RWn,1
RWn,j-1
RWn,j
RFn
RFn,1

Abb. 2/I

Abb. 2/II

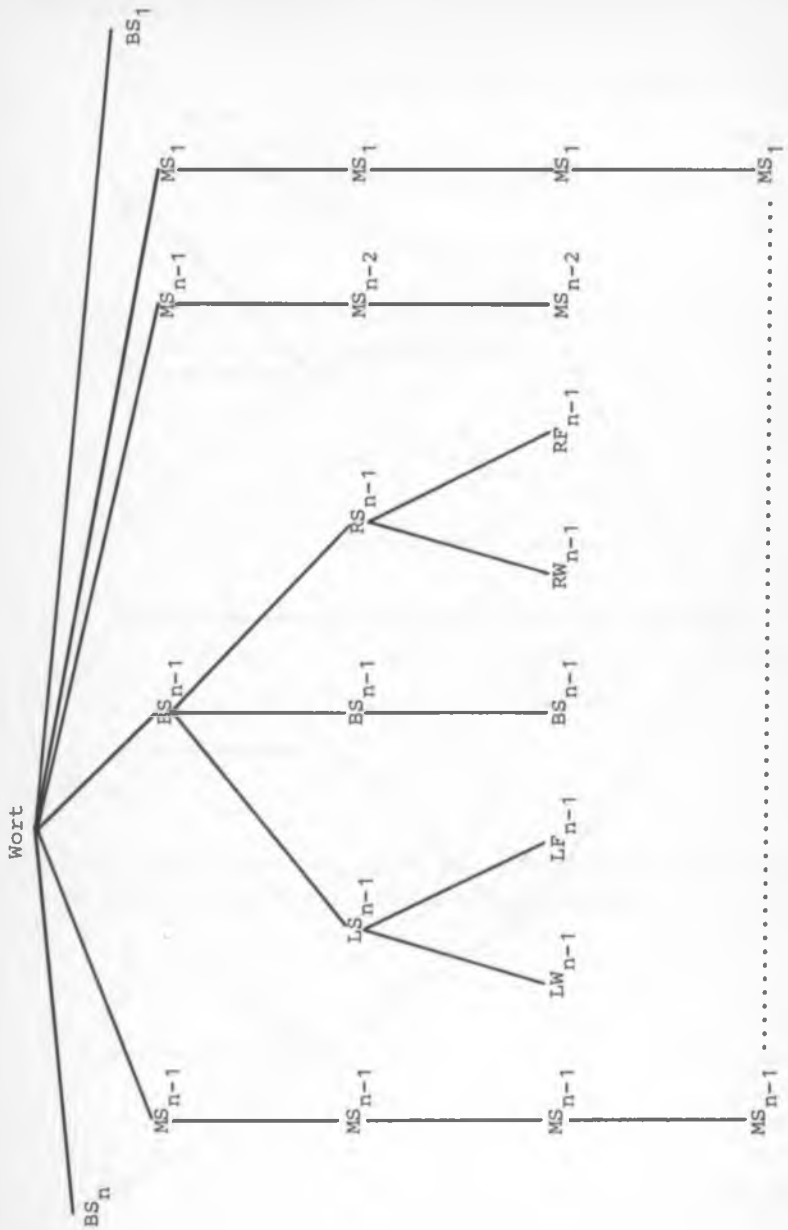
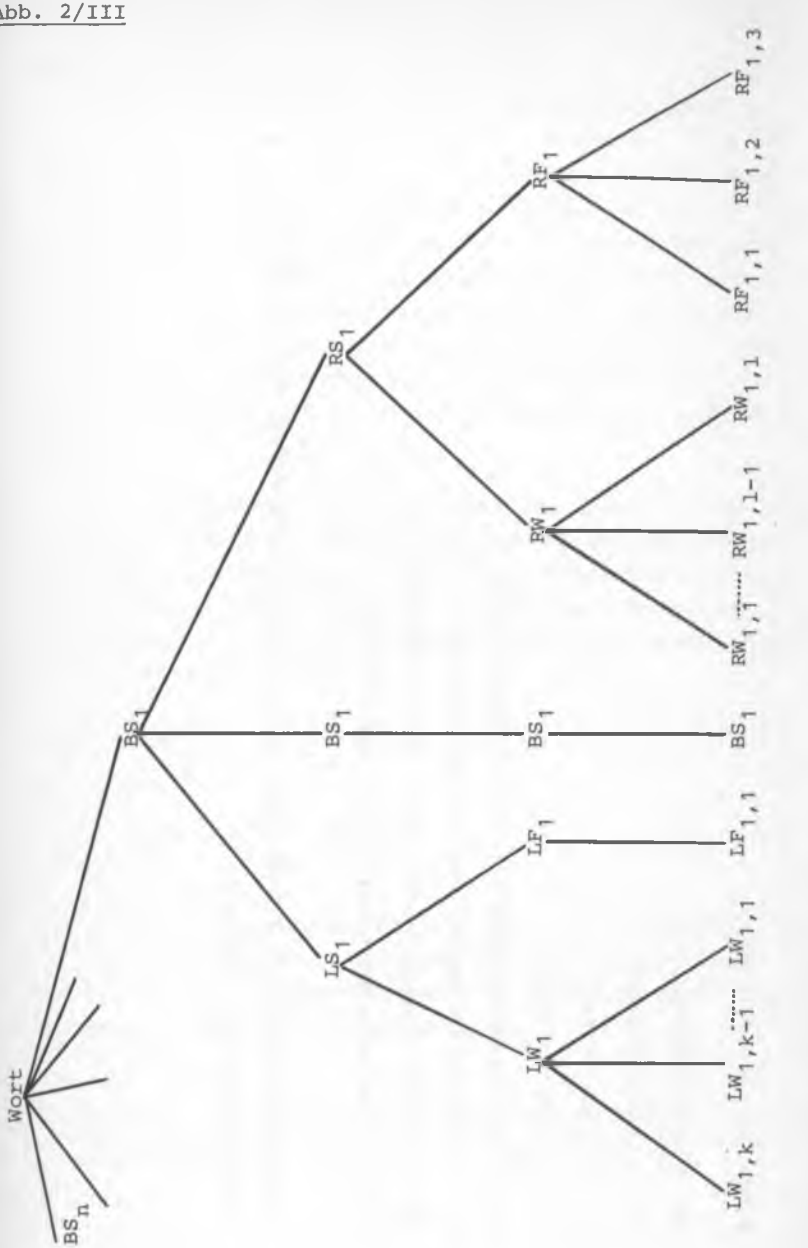

Wort
BS1
MS1
MSn-1
BSn-1
RSn-1
RFn-1
RWn-1
LSn-1
LFn-1
LWn-1
MS1
MSn-2
BSn-1
MS1
MSn-1
BSn

Abb. 2/III


Wort
BS n
BS 1
LS 1
RS 1
BS 1
BS 1
BS 1
LW 1
LF 1
RW 1
RF 1
LW 1,k
LW 1,k-1
LW 1,1
LF 1,1
RW 1,1
RW 1,1-1
RW 1,1
RF 1,1
RF 1,2
RF 1,3

Verwendete Abkürzungen:

BS = Basissegment

MS = Mittelsegment

LS = Linkssegment

RS = Rechtssegment

LW = Wortbildungssegment links von Basissegment

LF = Flexionssegment links von Basissegment

RW = Wortbildungssegment rechts von Basissegment

RF = Flexionssegment rechts von Basissegment

(Die Aufeinanderfolge von mehreren rechten Flexionssegmenten scheint nur beim Basissegment BS_1 zulässig)

3. Zur Darstellung der einzelnen Segmenttypen

Wenn man nicht nur eine Segmentierung - wie heute schon mit etwa 70 % Treffsicherheit in den Truncationverfahren der Dokumentation angewendet - von Wörtern vornehmen will, sondern diese Segmentierung auch als Ausgangsmaterial für weitere wissenschaftliche Untersuchungen brauchbar sein soll, dann muß die Trefferquote erhöht werden. Um dies zu erreichen, müssen die o.a. Segmenttypen sauber getrennt und auch entsprechend dieser Trennung wieder auffindbar abgespeichert werden, damit man mit einem Regelapparat auf die Segmenttypen selbst und außerdem auf einige Unterklassen zugreifen kann. Hierzu soll die folgende Nomenklatur verwendet werden:

Da die Basissegmente der Angelpunkt der Segmentierung sind, sollen sie als erste beschrieben werden. Bei Wörtern mit mehr als einem Basissegment werden die Basissegmente mit einem Index versehen, und zwar läuft dieser Index von rechts nach links[1]. *Landflucht* wäre demnach nach erfolgter Segmentierung notiert als:

Land	*flucht*
BS_2	BS_1

oder auch:

Plan	*ziel*	,	*Groß*	*mutter*
BS_2	BS_1		BS_2	BS_1

Bei den Wortbildungssegmenten (WS) als Linkssegmenten (LW) erfolgt die Indexierung wiederum von rechts nach links, so daß das dem Basissegment am nächsten stehende Wortbildungssegment mit dem Index 1 notiert wird.

---

1) Warum die Indexierung hauptsächlich von rechts nach links und nicht umgekehrt läuft, ergibt sich aus dem Abarbeitungsmechanismus, der unter 4. beschrieben ist.

Beispiele:

Aus	*ver*	*kauf*	,	*Mit*	*be*	*sitz*
LW_2	LW_1	BS_1		LW_2	LW_1	BS_1

Un	*ver*	*stand*
LW_2	LW_1	BS_1

Bei den Wortbildungssegmenten als Rechtssegmenten erfolgt die Indexierung jedoch von links nach rechts, so daß jeweils das dem Basissegment am nächsten stehende Segment die niedrigste Indexzahl aufweist.

Beispiele:

Klein	*lich*	*keit*	,	*Stein*	*ig*	*ung*
BS_1	RW_1	RW_2		BS_1	RW_1	RW_2

Links vom Basissegment stehende sowie rechts vom Basissegment bzw. vor rechten Wortbildungssegmenten stehende Flexive erhalten jeweils derart einen Index, daß der kleinere Index die größere Nähe zum Basissegment ausdrückt.

Beispiele:

aus	*zu*	*arbeit*	*en*	
LW_1	LF_1	BS_1	RF_1	
ab	*ge*	*spiel*	*t*	
LW_1	LF_1	BS_1	RF_1	
Mit	*ver*	*schwor*	*en*	*er*
LW_2	LW_1	BS_1	RF_1	RW_1
Ver	*arbeit*	*ung*	*en*	
LW_1	BS_1	RW_1	RF_1	

Bei Wörtern mit mehr als einem Basissegment verkompliziert sich die Darstellung insofern, als die verschiedenen Wortbildungs- und Flexionssegmente über einen zweiten Index an das Basissegment angebunden werden müssen, auf das sie sich beziehen, um einen leichteren Zugriff für die Weiterverarbeitung zu gestatten.

Beispiel:

Besoldungstarifvertragsparteien wird segmentiert in:

Be sold ung s tarif ver trag s part ei en
$LW_{4,1}$ BS_4 $RW_{4,1}$ MS_2 BS_3 $LW_{2,1}$ BS_2 MS_1 BS_1 $RW_{1,1}$ $RF_{1,1}$

Hierbei wird deutlich, daß nur die Wortbildungssegmente und Flexionssegmente über einen doppelten Index verfügen, der sie an das jeweilige Basissegment anbindet. Die Basissegmente selbst und die Mittelsegmente verfügen nur über einen Index.

4. Das Analyseverfahren

4.1. Ziel des Verfahrens

Eine beliebige in einem zu analysierenden Text auftretende Wortform soll in Segmente zerlegt werden. Das Verfahren soll so gut wie möglich sowohl an die bereits im IdS vorliegenden Daten als auch an die Lösungsverfahren ähnlicher Probleme angepaßt werden. Als Ergebnis des Analysevorgangs einer einzelnen Wortform erhält man ein oder mehrere mögliche Zerlegungen in Segmente. Das Ergebnis soll darüberhinaus in eine zu definierende standardisierte Form gebracht werden, die optimale Weiterverarbeitung ermöglicht.

4.2. Zugrundeliegendes Datenmaterial

Dabei sind zwei unterschiedliche Typen von Daten, über denen das Verfahren operiert, zu unterscheiden. Zunächst werden unterscheidbare Klassen von Segmenten definiert (verschiedene Rechtssegmente, Mittelsegmente, Linkssegmente, Basissegmente etc.). Für diese Segmente werden ein oder mehrere Lexika (SELEX) angelegt, die pro Eintrag neben dem Segment selbst weitere Beschreibungsmerkmale enthalten. Diese Lexika sind entweder bereits vorhanden oder aber ihre Erzeugung ist zum großen Teil automatisch

durch Programme durchführbar, die über schon vorliegenden Lexika operieren.

Der zweite Typ von Daten sind Regeln über die mögliche Aufeinanderfolge der verschiedenen Unterklassen der Segmenttypen. Diese Regeln werden nicht als Regeln für einzelne Segmente geschrieben, sondern als Regeln, die ganze Unterklassen von Segmenttypen erfassen. Dadurch ist es möglich, daß die Regeln auf bestehende Lexika angewandt werden können und bei der Hinzunahme einzelner neuer Segmente in das Segmentlexikon nicht umgeschrieben werden müssen. Dieses Regelwerk soll in Form eines Netzwerkes realisiert werden. Das Netzwerk ist darstellbar durch Ecken und Wege. Eine Ecke steht jeweils für eine bestimmte Klasse von Segmenttypen. Ein Weg geht von einer Ecke zu einer anderen. Das Netzwerk ist nun so beschaffen, daß jeder zulässigen Aneinanderreihung von Segmenten zu einer Wortform ein möglicher Weg im Netzwerk entspricht.

4.3. Das Verfahren

Die Analyse kann dann mittels eines sogenannten Netzwerkparsers erfolgen. Dies ist ein Programm, das zunächst ein Segment in der Wortform zu finden versucht. Wenn der Segmenttyp anhand des Lexikons festgestellt ist, wird die entsprechende Ecke im Netzwerk aufgesucht. Von dieser Ecke aus können nun in der Regel mehrere Wege zu einer nächsten Ecke begangen werden, die wieder einem neuen oder demselben Segmenttyp entspricht. Ob ein Weg begangen werden kann, hängt natürlich davon ab, ob das nächste in der Wortform auftretende Segment in der Klasse liegt, die der Ecke entspricht, zu der der gegangene Weg führt. Ist dies nicht der Fall, wird ein anderer Weg ausprobiert, usw.. Die Ecken tragen weiterhin eine Kennzeichnung, ob sie Anfangs- oder Endecken sein können, d.h. ob die zugehörige Klasse von Segmenten am Anfang oder am Schluß einer Wortform stehen kann.

Die Analyse kann also nur an einer Ecke beginnen, die als Anfangsecke zugelassen ist, und kann nur bei einer Ecke enden, die als Endecke zugelassen ist. Endet die Analyse bei einer Ecke, die nicht als Endecke zugelassen ist, so muß das Ergebnis verworfen werden. Es wird dann ein neuer Versuch unternommen. Das geschieht so: An jeder Ecke gehen, wie erwähnt, in dem Netzwerk mehrere Wege ab. Wenn ein Weg gar nicht begehbar ist, kann er ausgeschlossen werden. Ergibt sich erst bei der weiteren Verfolgung dieses Weges eine "Sackgasse", so wird zu der letzten Ecke zurückgegangen, bei der eine andere Entscheidung noch möglich ist, usw..

Das Ergebnis wird dann z.B. sein, daß genau ein zulässiger Weg zwischen mehreren Ecken gefunden wird. Es können auch mehrere zulässige Wege gefunden werden, was dann ein bzw. mehreren möglichen Zerlegungen entspricht. Tatsächlich wird dieser Vorgang noch dadurch kompliziert, daß nicht, wie etwa bei der syntaktischen Analyse eines Satzes (die im IdS mithilfe eines Netzwerkparsers durchgeführt wird) von vorneherein feststeht, welche Buchstabengruppen der Wortform Segmente sind. Dann ginge es ja nur noch darum, die Klasse dieser Segmente zu ermitteln. So aber müssen alle möglichen Buchstabengruppen nacheinander daraufhin geprüft werden, ob sie Segment sein können. Die diesbezügliche Entscheidung kann vor dem zeitaufwendigen Zugriff in ein Lexikon bei Basissegmenten dadurch erleichtert werden, indem ohne Lexikon geprüft wird, ob die betreffende Buchstabenfolge überhaupt ein Segment sein kann. Da Basissegmente Wortanfänge sein können, müssen die Buchstabengruppen von links gelesen einen sinnvollen Silbenanlaut darstellen, da die Wortanfänge einsegmentiger Wörter zugleich auch Anfänge der ersten Silbe sind. Für diese Prüfung steht bereits ein Programm zur Verfügung, das für eine Buchstabenfolge ermittelt, ob sie Silbenanlaut sein kann. Wenn dies nicht der Fall ist,

kann selbstverständlich die Prüfung im Lexikon entfallen.

Bei der Analyse der Wortform erscheint es aus einigen Gründen sinnvoll, von rechts zu beginnen, also von rechts nach links zu analysieren. Die Klasse mit der weitaus größten Anzahl unterschiedlicher Einzelsegmente ist die Klasse der Basissegmente. Eine Wortform beginnt häufig mit einem Basissegment, endet aber meist nicht mit einem Basissegment, sondern mit einem Rechtssegment. Es wird also in der Regel schneller gehen, das erste Rechtssegment zu finden. Über dieses erste Rechtssegment können zudem wichtige Informationen erschlossen werden, z.B. welche Segmentklassen als nächstes auftreten können etc.. Außerdem wird sich die Unvollständigkeit der Lexika am ehesten bei den Basissegmenten auswirken. In einem solchen Fall könnte man Rechts- und Linkssegmente suchen und den Rest versuchsweise als Basissegment ohne lexikalische Bestätigung analysieren.

Beispiel: Wortform *INFORMATIONSERSCHLIESSUNG*
(verkürzte Darstellung)

1. Ist *G* ein Rechtssegment? NEIN
2. Ist *NG* ein Rechtssegment? NEIN
3. Ist *UNG* ein Rechtssegment? JA RS1 = *UNG*
4. *S* könnte ein weiteres Rechtssegment oder ein Mittelsegment sein, dies ist aber ausgeschlossen, weil links von *UNG* nur ein Basissegment stehen kann
5. Ist *S* oder *SS* ein Basissegment? NEIN (kein Silbenanlaut)
6. Ist *ESS* ein Basissegment? JA (Stamm von *ESSEN*) BS1 = *ESS*
7. Jetzt kann, falls das Wort nicht zu Ende ist, ein Rechts-, Links- oder weiteres Basissegment oder ein Mittelsegment folgen.
8. Die Prüfung auf Linkssegment fällt negativ aus

(*I*, *LI*, *SCHLI* etc. sind keine Linkssegmente)

9. Die Prüfung auf Rechtssegmente fällt ebenfalls negativ aus
10. Die Prüfung auf Mittelsegmente fällt negativ aus
11. Zu prüfen ist also ein Basissegment.
    Auch hier fallen die Prüfungen negativ aus
    (*LI*, *SCHLI*, *ERSCHLI*, *SERSCHLI* etc.)
12. Es muß daher nun bei 6. erneut angefangen werden.
    BS1 $\neq$ *ESS*
13. Nach weiteren Prüfungen wird das Basissegment
    SCHLIESS gefunden BS1 = *SCHLIESS*
14. Ist *ER* ein Linkssegment? JA LS1 = *ER*
15. Jetzt können nur weitere Linkssegmente, ein Mittelsegment oder ein Basissegment folgen. Linkssegmentprüfung geht fehl
16. Ist *S* ein Mittelsegment? JA MS1 = *S*
17. Jetzt können Rechtssegmente oder Basissegmente folgen
18. Ist *ION* ein Rechtssegment? JA MS2 = *ION*
19. Ist *T* ein Rechtssegment? JA RS3 = *T*
20. Jetzt können Rechtssegmente oder Basissegmente folgen.
21. Die folgenden Prüfungen zeigen, daß *T* allein kein Rechtssegment sein kann. Es muß bei 19. erneut angefangen werden.
22. Ist *AT* ein Rechtssegment? JA RS3 = *AT*
23. Prüfung auf weitere Rechtssegmente geht fehl.
24. Prüfung auf Basissegmente *RM*, *ORM* geht fehl.
25. Ist *FORM* ein Basissegment? JA BS2 = *FORM*
26. Ist *IN* ein Linkssegment? JA LS2 = *IN*

Es ergibt sich somit folgendes Ergebnis:

IN	*FORM*	*AT*	*ION*	*S*	*ER*	*SCHLIESS*	*UNG*
LS_2	BS_2	RS_3	RS_2	MS_1	LS_1	BS_1	RS_1

Im obigen Beispiel wurde nur auf die Unterscheidung in Rechts- und Linkssegmente abgehoben. Feinere Unterschei-

dungen ebenso wie die Zuordnung der Links- und Rechtssegmente zu den beiden Basissegmenten unterblieben der Übersichtlichkeit wegen.

Dieses Beispiel sollte ansatzweise verdeutlichen, wie das Verfahren arbeitet. Nicht alle zu gehenden Wege sind vollständig aufgezeigt, einige nur angedeutet. Außerdem wird davon ausgegangen, daß die entsprechenden Segmente auch im Lexikon enthalten sind.

Man sieht an dem Beispiel bereits, daß der kritische Teil des Verfahrens, was Rechenzeit und Analyseergebnis betrifft, die Segmentbildung aus den einzelnen Buchstaben ist. Hierbei können in dem Netzwerk praktischerweise Stoppvorgaben gegeben werden. Z.B. ist es nicht sinnvoll, Buchstabenketten, die länger als 4 bzw. 5 sind, auf Links- oder Rechtssegment zu überprüfen. Hier ist im ersten Schritt die Prüfung abzubrechen und auf Basissegment zu prüfen etc. Genauso muß die Prüfung der möglichen Wege von einer Ecke zu den nächsten in eine Reihenfolge gebracht werden, bei der der häufigste Fall (Erfahrungswert) oder der am wenigsten zeitaufwendige Fall (kann berechnet werden) zuerst geprüft wird. Solche Änderungen können interaktiv während des Tests erfolgen, und bei der Erarbeitung des Netzwerkes wird man sich neben präzisen Kenntnissen und Erfahrungen von heuristischen Ergebnissen leiten lassen. Der Vorteil dieser Lösung ist, daß an dem einmal erstellten eigentlichen Programm (Netzwerkparser) nichts mehr geändert werden muß, sondern nur an dem Netzwerk selbst, über dem das Programm operiert. Dies ist auch für die allmähliche Erweiterung der qualitativen Leistungsfähigkeit des Programms wichtig, was ebenfalls durch Erweiterungen des Netzwerkes durchgeführt werden kann.

Einfaches Netzwerk

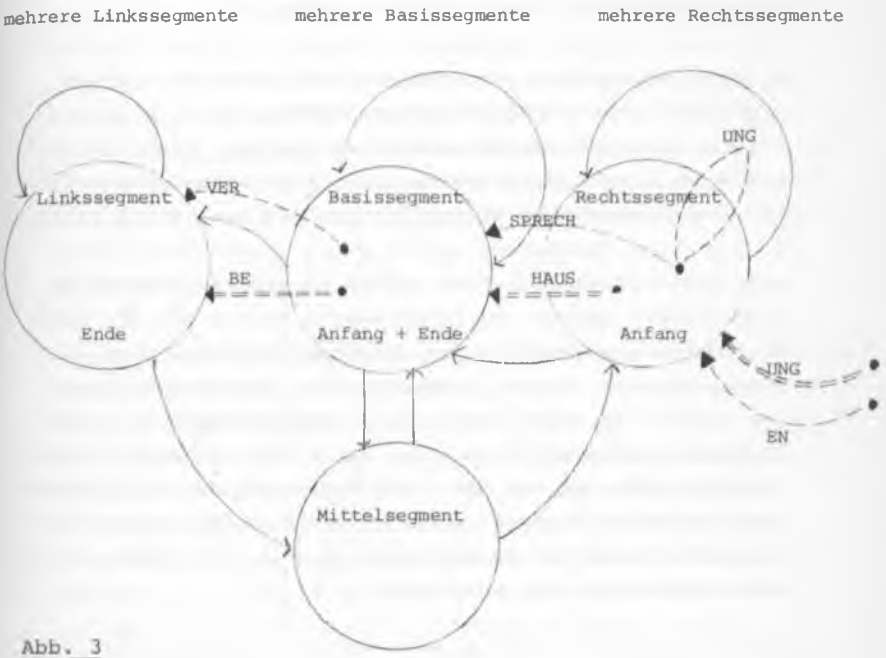


Abb. 3

Benötigte Linkssegmente:	Analysierbare Wortformen:
BE	BE-HAUS-UNG (= = =)
VER	VER-SPRECH-UNG-EN (- - -)
Benötigte Mittelsegmente:	HAUS-EN
S	SPRECH-EN
Benötigte Rechtssegmente:	⋮
UNG	
EN	

5. Zur Gewinnung der Segmente

Die obige Darstellung mag für manchen den Schluß zulassen, daß eine solche Segmentierung ohne weiteres leistbar und die verschiedenen Segmente sehr leicht abtrennbar seien. Dieser Schluß ist jedoch nicht richtig.

Da diese Segmentierung maschinell und nicht wie bisher größtenteils von Hand vorgenommen werden soll, ergeben sich einige grundsätzliche Schwierigkeiten, nämlich, wie man selbst unter der Forderung der lückenlosen Segmentierung zu "sinnvollen" Segmentierungen im o.a. Sinne gelangt.

Auch diese Schwierigkeiten sollen an einigen Beispielen verdeutlicht werden: zu segmentieren seien, aus dem Satzzusammenhang gegriffen, die folgenden Wörter: *Leben, Lebensgewohnheit, Leber, Lebergeschwür, Spieler, Spielergewohnheit*. In allen vier ersten Beispielwörtern taucht die Buchstabenkombination *-Leb* auf. Für ein maschinelles Verfahren wäre es nun das einfachste und auch einleuchtendste, die Beispielwörter in der folgenden Form abzutrennen (auf die Angabe der Segmenttypen wird im folgenden, da sie hier irrelevant ist, verzichtet):

Leb-en, Leb-en-s-ge-wohn-heit, Leb-er, Leb-er-ge-schwür, Spiel-er, Spiel-er-ge-wohn-heit.

Da das angestrebte Verfahren jedoch nicht nur nach bestimmten Regeln Segmentierungen vornehmen, sondern als Parser unter Benutzung des Segmentlexikons arbeiten soll, würde in diesem Fall diese Abtrennung nicht als Ergebnis akzeptabel sein können, und zwar aus den folgenden Gründen:

Im Lexikon der Segmente (SELEX) wird zwar *Leb-* als Basissegment aufgefunden, ebenso jedoch auch *Leber-*. Darüberhinaus wird *-er* als Wortbildungssegment ebenfalls in diesem Lexikon verzeichnet sein.

Über das Regelnetzwerk des Parsers wird nun auch maschinell die Entscheidung möglich, daß eine Abtrennung von *Leb-er* unzulässig ist, obwohl die gleiche Abtrennung bei *Spiel-er* zulässig ist und dies, obwohl sowohl *Leb-* als auch *Spiel-* Segmente sind, die in präfigierten Verben als Basissegmente auftreten, nämlich *spiel-en* und *leb-en*. Da jedoch eine Regel dieses ATN-Parsers besagen wird, daß für den Fall, daß ein Segment auf *-er* mit einem auch als Basissegment einzeln belegten Segment in das SELEX als Einheit aufgenommen wurde, eine Segmentierung ausgeschlossen ist, wird mithilfe des angestrebten Verfahrens in den Beispielen die folgende Segmentierung erfolgen: *Leb-en, Leb-en-s-ge-wohn-heit, Leber, Leber-ge-schwür, Spiel-er, Spiel-er-ge-wohn-heit.*

Daß die Segmentierung nicht rein formal *-er-, -el* als Rechtssegment abtrennt, muß verhindert werden, da z.B. eine Segmentierung von *Hamm-er, Vat-er, Mutt-er, Vog-el, Spieg-el-n* usw. keine im o.a. Sinne "sinnvolle" Segmentierung darstellt, da weder *Hamm-, Vat-, Mutt-, Vog-* oder *spieg-* in anderen Wörtern der deutschen Sprache ihre Baustein-Funktion erfüllen und somit nicht als "sinnvoll" erkannte Segmente klassifiziert werden können, obwohl"Wörter" auf *-el* und *-er* so häufig in der deutschen Sprache sind, daß man eine Abtrennung von *-el* und *-er* fast schon als geläufige Segmentierung ansehen könnte.

Eine weitere Schwierigkeit bilden die Flexionssegmente, die in Zusammensetzungen auftauchen und hier ihre Flexionseigenschaften formal (d.h. maschinell) nicht mehr erkennbar aufweisen. Im obigen Beispiel wurde *-ge-schwür* in dieser Form segmentiert, was wohl auch die "sinnvollste" Abtrennung darstellt, obwohl dafür eigentlich keine formalen Kriterien vorliegen. Es gibt zwar ein Basissegment *-schwür-*, das als Imperfekt Konjunktiv von *schwören* oder als Plural

von *Schwur* in *Schwür-e* belegt ist.[1)]

In diesem Fall ist jedoch die Entscheidung, ob *-ge-* Flexionssegment oder Wortbildungssegment ist, nach linguistischen Kriterien zu treffen. Diese Entscheidung schlägt sich dann einmal in der Beschreibung für *-ge-* im SELEX nieder, zum andern auch im Regelteil des Netzwerkes; sollte z.B. die Entscheidung dahin gehen, daß *-ge-* in diesem Fall Wortbildungssegment ist, dann müßte *-schwür-* in die Ausnahme-Liste aufgenommen werden, die besagt, daß bei einem direkten Zusammentreffen von *-ge-* und *-schwür- -ge-* nicht als linkes Flexionssegment zu notieren ist.

Ein weiteres Beispiel soll nun zum Schluß noch zeigen, daß bei einem Teil der Mehrdeutigkeiten, die entstehen werden, Wortart, Genus und sonstige morphologische Informationen von großer Bedeutung sein werden. Deshalb soll das Verfahren auch an die am IdS verfügbaren Verfahren zur Morphologie- und Syntaxanalyse angebunden werden. Wie das im einzelnen geschehen soll, wird unter 6. kurz skizziert.

Es seien zu analysieren Spielerei und Hühnerei, Spielereien und Hühnereier.

In beiden Fällen wird die Segmentierung keine Schwierigkeiten bringen: *Spiel-er-ei, Hühn-er-ei, Spiel-er-ei-en, Huhn-er-ei-er*. Schwierigkeiten bereitet in diesem Fall das Segment *-ei-* bezüglich seiner Segmenttyp-Zuordnung. Im

---

1) Die Großschreibung der Nomen soll für das beschriebene Verfahren ohne Bedeutung sein, da ein Großteil der zu verarbeitenden Texte sowieso nur in Großbuchstaben vorliegt und außerdem die am IdS vorhandene Syntaxanalyse, an die das Verfahren angeschlossen werden soll, Nomina über den Parser und nicht über die Großschreibung erkennt.

SELEX wird *-ei-* einmal als Basissegment und einmal als rechtes Wortbildungssegment vorgefunden. Verfügt man nun neben der reinen Worteingabe über einen syntaktisch annotierten Text, so wird man *Spielerei* als Nomen femininum, *Hühnerei* als Nomen neutrum eingetragen finden. Im SELEX wird man bei *-ei-* die Information finden, daß, wenn *-ei-* als Wortbildungssegment nur noch gefolgt von rechten Flexionssegmenten auftritt, es sich um ein Nomen femininum handeln muß, daß bei *-ei-* als Basissegment im gleichen Fall nur Nomen neutrum in Frage kommt. In diesem Fall, der durch die noch weiter rechts vorhandenen Flexivsegmente in den angeführten Plural-Wörtern noch weiter verdeutlicht wird, ist eine eindeutige Trennung von *-ei-* in Basis- oder Wortbildungssegment auch maschinell möglich.

Maschinell nicht mehr lösbar sind die Fälle *Hühn-er-ei-enzym* und *Spiel-er-ei-er-find-er*. In diesen beiden Fällen wird mit Sicherheit eine Mehrfachsegmentierung das Ergebnis sein, die dann nachbehandelt werden muß.

Ebenfalls nicht maschinell lösbar und auch für den menschlichen Sachbearbeiter nur über den Satzzusammenhang zu entscheiden sind *unter-min-ier-t* und *un-termin-ier-t*.

6. Zur Einbindung des Verfahrens in die am IdS vorhandenen Verfahren

Am IdS wurde in der ehemaligen Abteilung Linguistische Datenverarbeitung im Rahmen des Projekts PLIDIS eine morphologische und eine syntaktische Komponente erarbeitet, die es erlaubt, beliebige Sätze syntaktisch zu analysieren. Diese Komponenten wurden realisiert als PASSØ (morphologische Analyse, d.h. Extraktion sämtlicher möglicher morphologischer Beschreibungen zu einer Wortform aus einem Vollformenlexikon) und PASS1, Syntaxanalyse mithilfe eines ATN-Parsers.[1] Der PASSØ, der eigentlich nur in der Aufsuche des Wortes in einem Vollformenlexi-

kon (MOLEX) besteht, wurde ermöglicht durch den MOLEX-Generator, der aus der Angabe der Grundform eines Wortes und seiner Flexionsklasse sämtliche MOLEX-Einträge generiert.[2)]

Vergleichbar der Funktion des MOLEX für die Syntaxanalyse wird das Segmentlexikon (SELEX) für die Segmentanalyse fungieren. Dieses Lexikon wird während der Arbeiten an dem Verfahren sowie im späteren Einsatz sukzessive aufgebaut. Es enthält sämtliche die Segmente betreffenden Informationen und soll in seinem Aufbau dem MOLEX weitmöglichst angeglichen werden, da dadurch ein Großteil der für das MOLEX schon programmierten Zugriffs- und Abspeicherungsroutinen erhalten bleibt.

Über das SELEX soll - vergleichbar der Angabe der Normalform bei jedem Vollformeneintrag im MOLEX - die Zuordnung der Basissegmente zu einem Normsegment geleistet werden.

Bei unregelmäßigen Verben wechselt häufig der Stammvokal des Basissegments,bei den Verbformen z.B. des Imperfekts oder Konjunktivs. Für eine spätere linguistische Auswertung der Ergebnisse des Segmentierungsprogramms wäre es jedoch nützlich, sämtliche Basissegmente, die auf das gleiche Verb zurückgehen, in ihren Belegungen gemeinsam zu erhalten. Bei *fahr-en* wären dies etwa *fuhr-en*, *führ-en*. Im SELEX sollte deshalb für die Segmente *fahr- fuhr- führ-* ein Normsegment angegeben werden, das von den un-

---

1) Vgl. hierzu Lutz, H.D., Kolvenbach, M., Zifonun, G.: PLIDIS-Dokumentation Version 3.0, Mannheim 1980. Hier finden sich auch die Einschränkungen zu "beliebig".

2) Vgl. hierzu M. Kolvenbach: Das morphologische Lexikon (MOLEX) des Systems PLIDIS, Mitteilungen des Instituts für deutsche Sprache, Nr. 7, S. 60ff. und M. Kolvenbach, W. Teubert: Zur Generierung der Nomen, Mitteilungen des Instituts für deutsche Sprache, Nr. 8, Mannheim 1981; LDV-Info 2, 1982

terschiedlichen Vokalen absieht. Für das Segment *führ-* gäbe es damit zwei Normsegmente, einmal wenn es von *fahren* und zum andern, wenn es von *führen* herzuleiten ist. Eine Angabe des zutreffenden Normsegments z.B. für die Wortform *führ-t* wird erst in einem sehr späten Stadium des Verfahrens möglich sein. Zu Anfang wird nur eine Zusammenordnung so unterschiedlicher Segmente wie *iß, aß, äß, ess, iss* unter dasselbe Normsegment leistbar, da sie nicht auf mehrere unterschiedliche Normsegmente rückführbar sind.

Das gleiche Problem tritt auf bei im Plural umlautenden Nomen und im Komparativ umlautenden Adjektiven. Hier sollte ebenfalls eine Darstellung gefunden werden, die sich nicht rein an der Oberfläche orientiert.

Das oben vorgestellte Verfahren soll einmal in seiner Arbeitsweise ähnlich dem PASSØ (SELEX) und PASS1 (Netzwerk-Parser) arbeiten und eine Schnittstelle zu diesen ehemaligen PLIDIS-Komponenten erhalten. Dadurch wird es möglich, in Zweifelsfällen für die Segmentierung morphologische Informationen aus dem MOLEX oder syntaktische Informationen über die Syntaxanalyse heranzuziehen. Darüberhinaus werden die Ergebnisse der Segmentanalyse als Zusatzinformationen in das MOLEX eingebracht, so daß für einen Großteil der Wortformen in neuen zu segmentierenden Texten schon die Ergebnisse früherer Segmentierungen über eine einfache Aufsuche im Lexikon eingebracht werden können und nur die bisher noch nicht - in anderen Texten - segmentierten Wortformen das Verfahren durchlaufen müssen.