SOLL-ABLAUF CTX
(MORPHOSYNTAKTISCHE INDEXIERUNG)

SAARBRÜCKEN, JANUAR 1983

UNIVERSITÄT DES SAARLANDES
5,5 INFORMATIONSWISSENSCHAFT
PROF. DR. HARALD H. ZIMMERMANN
F-509-B1-04

ÜBERSICHT

Vorbemerkung

Die folgende Beschreibung bezieht sich auf den Teil des Saarbrücker Systems zur computergestützten Texterschließung (CTX), der inhaltliche Mehrdeutigkeiten nicht berücksichtigt. Mithilfe des Verfahrens werden:

- sinntragende Textwörter (z.B. Wortformen wie SCHLOESSER, GESEHEN, SCHÖNES) in der Grundform (hier z.B. SCHLOSS, SEHEN, SCHÖN) als "Deskriptoren" zu einem Text/Dokument ermittelt (morphologische Analyse);
- aufgrund bestimmter syntaktischer, d.h. satzorientierter Strukturen Präkoordinationen vorgenommen (z.B. "schöne alte Schlösser" - SCHOENES SCHLOSS; ALTES SCHLOSS; "Untersuchungen von Anlagen" - UNTERSUCHUNG ANLAGE G):
- zusammengesetzte Wörter in sinntragende Bestandteile zerlegt (z.B. "Haustür" - HAUSTÜR, TÜR, HAUS)
- morphologische Derivationen einander zugeordnet (z.B. UNTERSUCHUNG - UNTERSUCHEN).

In dieser Materialie wird der (technische) Ablauf des Verfahrens vorgestellt.

1. TEXTAUFNAHME (Basis-Inventar)

Der maschinenlesbare Fremdtext wird üblicherweise auf Magnetband angeliefert. Das Band wird zunächst technisch auf Platte umgesetzt, anschließend an die CTX-Rahmen-Norm angepasst und schliesslich (ggf.) zur Weiterverarbeitung in kleinere "Portionen" zerlegt (Paketierung). Der Output ist unmittelbar Input-Schnittstelle für den Aufbau des Dokumenttexts in der Datenbank (sog. "Dokumentkörper"). Er ist zugleich Inputschnittstelle für die Textanpassung an die Saarbrücker Textanalyse, i.a. aber dazu noch weiter aufzubereiten. Über die einzelnen Schritte wird ein Inventar- und Zustandsprotokoll (Basisinventarbuch) geführt. Dabei wird (je nach Zustand) festgehalten:

(a) Lieferant (z.B. DPA - Deutsches Patentamt)
(b) Datum der Lieferung (JJMMTT)
(c) Inhaltsangabe, z.B. "Offenlegungsschriften" der 51. Woche 1982"
(d) (Vorgesehener) Dateiname für die 1:1 umgesetzte Datei auf der Siemens-Rechenanlage
(e) Datum der Dateierstellung (JJMMTT)
(f) Dateiname nach Umsetzung in die CTX-Rahmen-Norm
(g) Datum der Umsetzung in die CTX-Rahmen-Norm (JJMMTT)
(h) Dateiname(n) für Analyse-Inputdatei(en)
(i) Datum der Erstellung der Input-Pakete (JJMMTT)

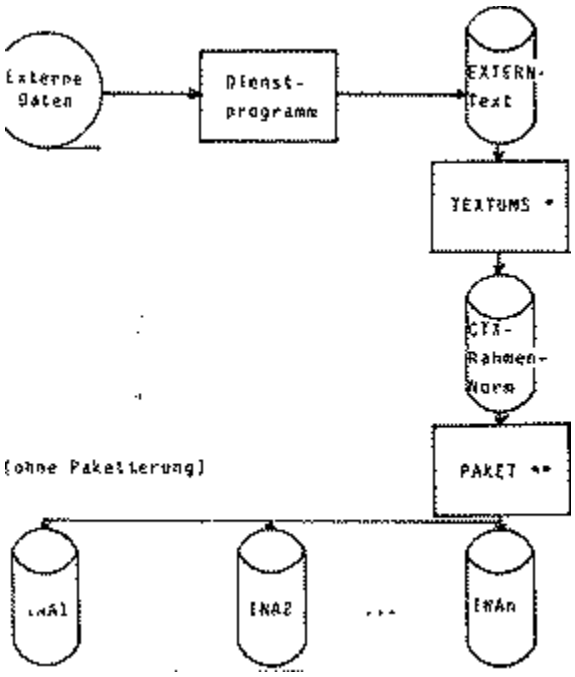

Abb. 1: Textaufnahme

* Je Externformat wird ein entsprechendes Umsetzprogramm benötigt. Bislang liegen Umsetzprogramme vor für Daten von DPA, Deutsche Bibliothek, BAM, WZB.

** Die Paketierung dient der Portionierung von Externdaten. Sie soll eine Erleichterung der Arbeitsorganisation mit sich bringen. In der Regel sollen ca. 1.000 Zeilen Text zu einem "Paket" gebündelt werden, wobei in jedem Falle der Einschnitt an Dokumentgrenzen erfolgt.

II. TEXTANPASSUNG

Die Textanpassung soll eine befriedigende Analyse der Texte / Sätze sicherstellen. Hierbei ist zu unterscheiden zwischen rein technischen (automatisch ablaufenden) Verfahren (der sog. Textnormierung) und intellektuellen Eingriffen (Präkodierung, Textkorrektur).

Ein Paket wird zunächst "technisch" auf die Inputanforderungen der Textanalyse hin normiert (TXNORM) und dabei die Inputdatei für das Analysesystem erstellt (INBx). Dabei werden Umlaute (ä, ö, ü) ggf. umgesetzt (in ae, oe, ue); "ß" wird in "ss", der Satzendepunkt in "*" verwandelt. Soweit erkennbar (und ohne Einfluss auf den Dokument"körper"), wurden bereits bei der Umsetzung in die CTX-Rahmen-Norm Textstrukturierungen vorgenommen (z.B. Kennzeichnung eines Dokumentanfangs, ggf. auch Kennzeichnung des Satzendepunktes als "*"). Dennoch kann es erforderlich sein, den Text über spezielle Programme oder auch intellektuell weiter aufzubereiten. So müssen z.B. Spiegelstriche und Klammern besonders behandelt werden; gelegentlich bereitet auch der Bindestrich Probleme (z.B. "1980-1985": hier bedeutet "-" eigentlich "bis"). Auch der Wortbindestrich am Wortanfang und Wortende ist zu bearbeiten. Schliesslich müssen "überlange" Wörter und Sätze behandelt werden, d.h. Wörter mit mehr als 36 Buchstaben und Sätze mit über 100 Wörtern. Ein Großteil dieser Problemfälle wird maschinell (durch Hinweise des Programms TXNORM) überprüft. Mögliche Fehler werden in der Textfehlerliste TLIST

angezeigt. Hierbei können die "Pakete" parallel behandelt werden. Üblicherweise wird der Editor des Betriebssystems für diese Arbeit herangezogen.

Für die Korrektur von Rechtschreibfehlern wie auch für weitergehende Präkodierungen ist stets die INBx-Datei zugrundezulegen. Eine Ausnahme liegt vor, wenn der Original-Input im Rahmen der CTX-Anwendung in die Korrekturphase mit einzubeziehen ist. Dann ist ggf. (d.h. zur Bereinigung von Fehlern im Dokument-"Körper") auf der INAx-Schnittstelle aufzusetzen bzw. dort zusätzlich zu korrigieren.

Der Status der Bearbeitung wird in einem Paket-Inventar festgehalten. Dieses ist eingehend im Abschnitt "Wörterbucherweiterung" (IV) beschrieben.

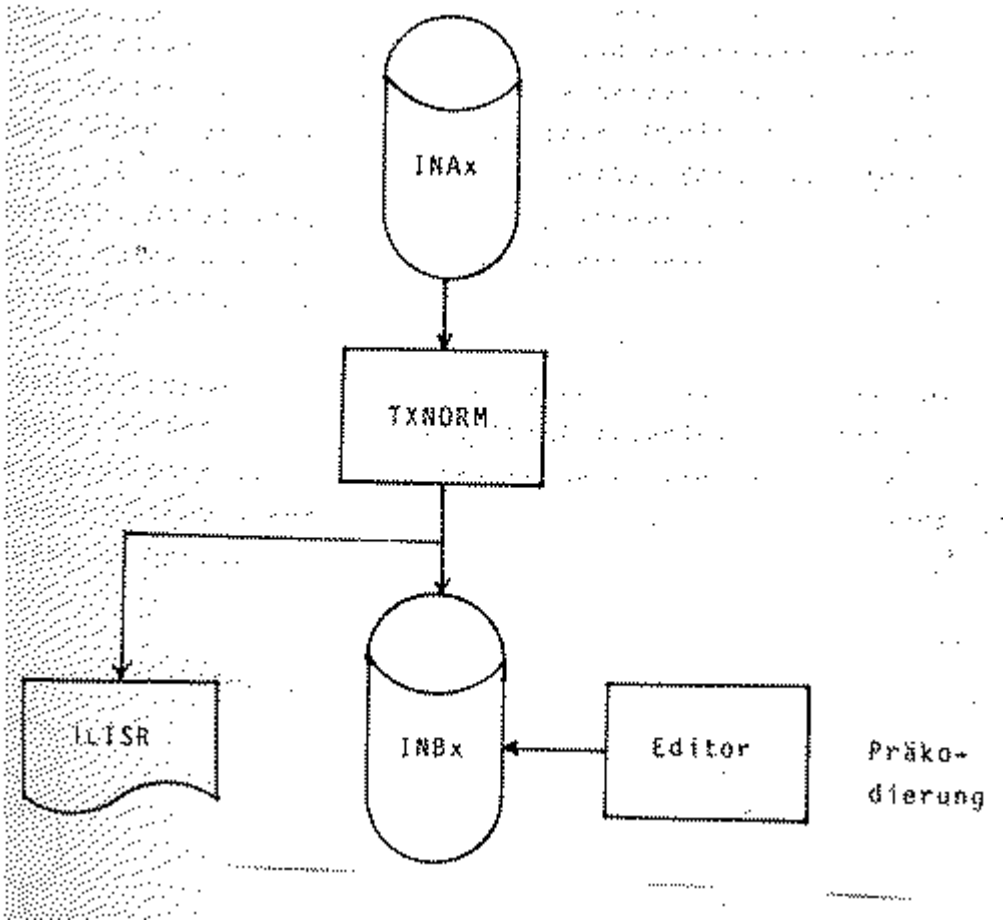

Abb. 2: Textanpassung

III. SATZSEGMENTIERUNG

Das der Verarbeitung zunächst zugrundeliegende System "Saarbrücker Automatische Textanalyse" des Sonderforschungsbereichs 100 ist satzorientiert, d.h. es werden nur syntaktische Strukturen in einem Satz ermittelt. Hierzu wird ein fortlaufender Text zunächst in solche "Sätze" zerlegt: Sätze werden voneinander abgegrenzt durch die Satzendezeichen Punkt (dargestellt als '*'), Ausrufezeichen, Fragezeichen oder künstliches Satzende (dargestellt als ' '). Dabei werden Sätze innerhalb eines Analyselaufs fortlaufend aufsteigend nummeriert und die Wörter im Satz eben-

falls. Ein Satzzeichen wird als eigenes 'Wort' gezählt; so hat der Satz 'HEUTE IST SCHÖNES WETTER* ' 5 'Wörter', der Satz 'ICH WEISS, DASS DU KOMMST' 7 'Wörter'. Anführungszeichen und Apostroph sind keine Satzzeichen, sondern Wortzeichen (und werden dementsprechend nicht mitgezählt).

Das Analyseergebnis wird in einer Datei (LESx) festgehalten, die Sätze werden in einer Liste SLIST ausgedruckt. Falls beim Lesevorgang zusätzliche Fehler auftreten, wird eine selbsterklärende Hinweisliste (T2LIST) erstellt, die analog zu TLIST zu behandeln ist.

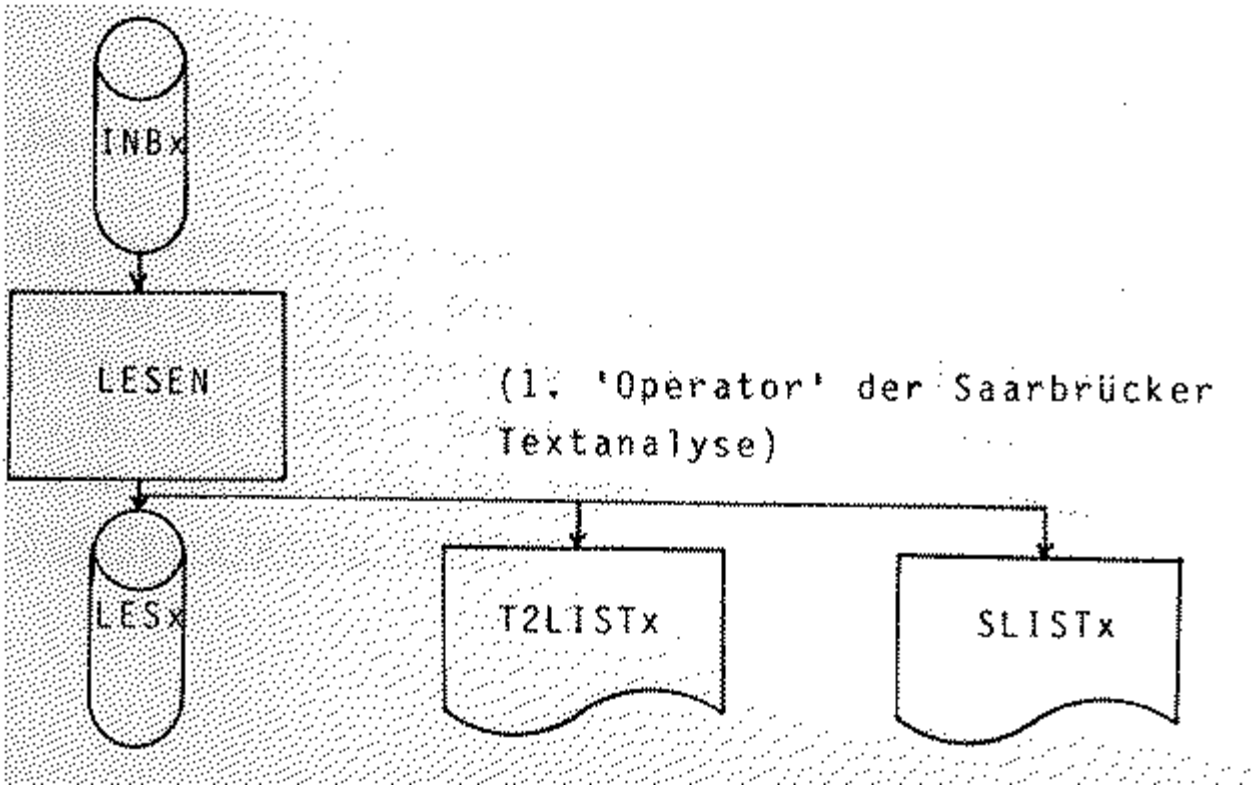

Abb. 3: Satzsegmentierung

IV. WÖRTERBUCHERWEITERUNG (ohne Semantik)

Diese Phase setzt eine abgeschlossene Satzsegmentierung (vgl. III) voraus. In Phase IV kann - wenn zuvor in Pakete geteilt wurde - stets nur ein Paket bearbeitet werden.

Lexika

Die morphosyntaktische Analyse verwendet folgende (strukturgleichen) Wörterbücher:

(1) ein morphosyntaktisches Wörterbuch (SADAW)
(2) ein Kompositum-Lexikon (KOMPLEX)
(3) ein Lexikon der Abkürzungen (ABKLEX)

Das CTX-System benutzt darüber hinaus (ohne Semantik-Teil) folgende Wörterbücher:

(4) ein Partizipien-Lexikon (PTZLEX)
(5) ein Derivations-Lexikon (DERLEX)
(6) ein "Fachlexikon" (FALEX).

zu 1: (SADAW)

Das morphosyntaktische Lexikon (SADAW) umfasst alle Funktionswörter (z.B. DER, UND, JENER) und mindestens alle Simplizia (d.h. Einfachwörter) zu Adjektiven (z.B. "SCHÖN"), Verben (z.B. "GEH"(EN)), Substantiven (z.B. "HAUS") und Adverbien (z.B. "HIER"). Es ist immer zu ergänzen, wenn ein derartiger Simplex-Eintrag fehlt.

Das Analysesystem ist in der Lage, in den meisten Fällen Wortzusammensetzungen (z.B. "TELEFONKABEL") und Wortableitungen (z.B. "SINGBAR") auch dann zu erkennen, wenn nur die Wortbestandteile (hier: "TELEFON", "KABEL", "SING(EN)") im Wörterbuch vorhanden sind. Daher müssen solche Einträge nicht notwendig in SADAW aufgenommen werden. Ähnliches gilt für Abkürzungen, wenn diese als "groß geschrieben" gekennzeichnet sind und als "Substantive" behandelt. werden können.

In beiden Fällen empfiehlt sich aber eine (weitgehend automatisierte) Eintragung. Bei Abkürzungen ist z.B. die (intellektuelle) Markierung von Genus und Numerus nützlich (die bei der automatischen Erkennung ansonsten "offen bleibt"); bei korrekt zerlegten Komposita und Derivationen kann bei einer Lexikalisierung erheblich Rechenzeit (ca. 2/3) gegenüber der automatischen Dekomposition bei Wiederauftreten in den Texten eingespart werden. In jedem Falle sind dazu die Zusatzlexika KOMPLEX und ABKLEX zu erweitern (vgl. 2 und 3)..

zu 2: (KOMPLEX)

Das Kompositum-Zerlegungslexikon umfasst die Komposita (bzw. Derivationen) (in der Grundform) und ihre "sinnvollen" morphologischen Bestandteile. Diese werden im CTX-System zur Bildung von Teil-Relationen bzw. Derivationsrelationen herangezogen. Während der Wörterbuchsuche wird KOMPLEX überprüft, soweit kein direkter SADAW-Eintrag vorliegt. Die SADAW-Einträge, soweit es sich um Komposita handelt, sind bereits vollständig in KOMPLEX enthalten. Nur wenn auch KOMPLEX bei Dekompositionen und Derivationen keinen Treffer ausweist, wird der Eintrag zur intellektuellen Kontrolle und Bearbeitung in die Lexikalische Kontrollliste "KLIST" eingetragen. Nach intellektueller Kontrolle (und ggf. Korrektur) derartiger Einträge wird KOMPLEX anschließend erweitert.
Bezogen auf die sinnvollen Zerlegungen stellt KOMPLEX zugleich einen Bestandteil des CTX-Thesaurus dar. Insofern ist nicht nur eine korrekte Identifikation, sondern auch eine korrekte Zerlegung zu überprüfen und sicherzustellen.

zu 3: (ABKLEX)
Das Abkürzungslexikon ABKLEX wird analog zu KOMPLEX bearbeitet (vgl. 2). Bei Bedarf und nach Möglichkeit können hier die Langformen zu einer Abkürzung als "Synonyme" eingetragen werden; ABKLEX ist insofern Bestandteil des CTX-Thesaurus. Durch Vergleich eines nicht identifizierten "Substantivs" mit ABKLEX während der Wörterbuchsuche wird sichergestellt, dass die Lexikalische Kontrollliste KLIST nur Wörter enthält, die zuvor (d.h. bei früheren Bearbeitungen) nicht schon einer entsprechenden Kontrolle unterworfen waren.

zu 4: (PTZLEX)

Das Partizipienlexikon umfasst alle Partizipien von Verben. Die SADAW-Einträge sind bereits vollständig enthalten, so dass nur dann noch Ergänzungen vorgenommen werden müssen, wenn ein (Simplex-)Verb nicht in SADAW enthalten war. Das Partizipien-Lexikon wird zum Aufbau komplexerer attributiver Deskriptoren herangezogen. Es wird automatisch beim Neueintrag eines Verbs erweitert.

Bsp.: Text: die gesungenen Lieder

Analyse: SING (Verb) LIED (Substantiv)

Deskriptoren: o SINGEN
o LIED
o GESUNGENES LIED

zu 5: (DERLEX)

Das Derivationslexikon dient der Zuordnung von Verben, Adjektiven und Substantiven mit "gleichem" Stamm (im CTX-Thesaurus), also etwa

BERUHIGUNG - BERUHIGEN
SCHUTZ - SCHÜTZEN - SCHÜTZBAR.

Das Lexikon ist "technisch" Teil des FALEX (vgl. 6), da es sich nur um besondere (morpho-syntaktische) Relationstypen handelt. Es muss intellektuell aufgebaut und gepflegt werden. Im Prinzip wäre auch hier - analog zu dem Aufbau des Partizipien-Lexikons - eine systematische Umsetzung der SATAN-Einträge als erster Schritt sinnvoll.

Unter Verwendung von DERLEX kann dies jedoch auch textbezogen erfolgen. Alle Simplizia des Textes, die in SADAW gefunden wurden, werden mit DERLEX/FALEX verglichen. Sind sie dort nicht aufgeführt, so werden sie in einer speziellen Liste RLIST aufgenommen und intellektuell überprüft.

Wenn sie nicht morphosyntaktisch relationierbar erscheinen, werden sie in DERLEX/FALEX mit entsprechender Markierung eingetragen. Können sie relationiert werden, so werden sie in DERLEX/FALEX mit den Relationen übernommen.

Die Relata werden anschliessend mit SADAW abgeglichen und ggf. dort ebenfalls ergänzt.

zu 6: (FALEX)

Für die nicht-semantische Analyse ist FALEX weitgehend identisch mit DERLEX. Daneben können noch Synonym-Relationen eingebracht werden (wobei allerdings Probleme bei mehrdeutigen Wörtern auftreten können).

Es ist hier deutlich darauf hinzuweisen, dass FALEX bei der Verwendung der semantischen

Analyse (CTX-II) einen anderen Umfang aufweist und dabei auch weitere Funktionen übernimmt.

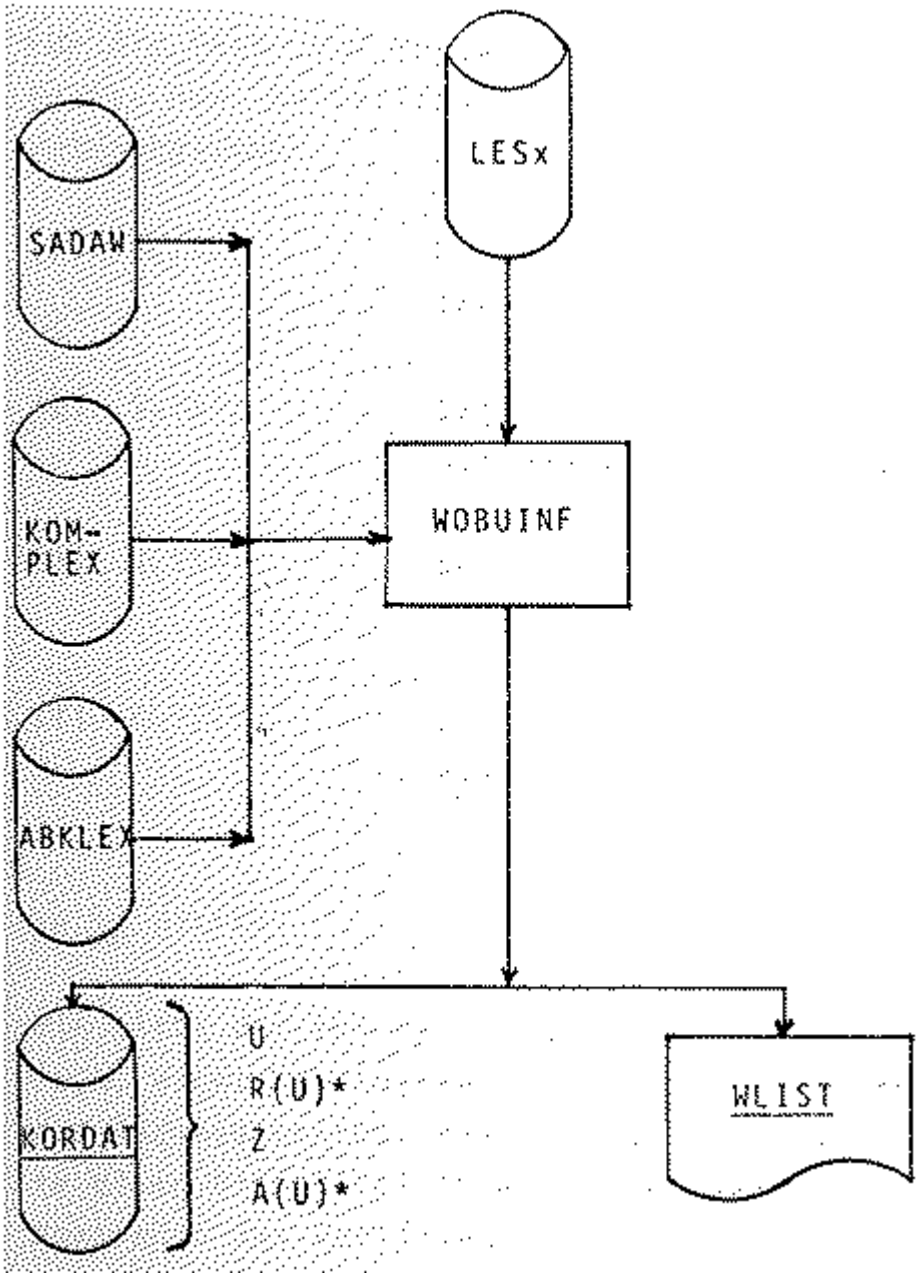

Abb. 4: Wörterbuchabgleich zu Korrekturzwecken

U = "echt" unbekanntes Wort
(R) = "Rechtschreibfehler"
Z = Kompositumzerlegung **
(A) = Abkürzung *

---

* Zunächst ist in der Datei KORDAT nur differenziert nach "U" (nicht identifiziertes Wort) und "Z" (Kompositumzerlegung). Nach Durchsicht der Probeliste werden die Einträge in

KORDAT intellektuell weiter differenziert, wobei R für Rechtschreibfehler und A für Abkürzungen steht.

** Die durch Zerlegung ermittelten Komposita werden einem speziellen Korrekturverfahren unterzogen (vgl. Materialie "Kompositumkorrektur").

V. KORREKTURLISTEN

Für die Text- und Wörterbuchkorrektur werden entsprechende Hilfs- bzw. Korrekturlisten erzeugt. Sie werden wie folgt unterschieden:

(1) TLIST: Liste der bei der Texteingabe ermittelten "fehlerhaften" Wörter

(2) SLIST: Satzliste (Textkennung und Satznummer der bearbeiteten Sätze)

(3) WLIST: Wortliste der lexikalisch zu überprüfenden Wörter

(4) RLIST: Liste potentieller Relationswörter nach DERLEX/FALEX-Abgleich

zu 1: (TLIST)

Die Liste der bei der Texteingabe ermittelten fehlerhaften (d.h. nicht inputnormgerechten) Wörter dient zur Korrektur des Ausgangstextes. Die korrigierten Wörter werden parallel in eine Hilfsdatei TKORD geschrieben und einer separaten Kontrolle des Wortinventars unterzogen. (Vgl. dazu im Einzelnen: Materialie "Textkorrektur") (Ggf., wenn vorhanden, muss in gleicher Weise die Hilfsliste T2LIST bearbeitet werden).

zu 2: (SLIST)

Die Satzliste kann bei der Kontrolle auf korrekte Satzsegmentierung herangezogen werden. Sie ist zugleich Hilfsliste zu den Korrekturarbeiten. Ggf. muss der Ausgangstext (INBx) korrigiert werden. (Vgl. dazu im einzelnen: Materialie "Satzsegmentierung")

zu 3: (WLIST)

Die Liste der zu überprüfenden Wörter umfasst

(a) durch Dekomposition/Derivation identifizierte Wörter (entsprechend markiert)

(b) Rechtschreibfehler u.ä. (hier ist der Text (in INBx) zu korrigieren)

(c) "echt" unbekannte Wörter, wobei es sich um Simplizia, Komposita oder Abkürzungen handeln kann.

Vor der eigentlichen Korrektur ist diese Liste (mit Programmunterstützung) entsprechend zu differenzieren.

Dabei werden

(d) die Komposita in KOMPLEX übertragen (diese können aus (a) oder (c) stammen)

(e) Rechtschreibfehler in korrekte Textwortformen umgesetzt

(f) SADAW-Einträge vorgenommen

(g) das Abkürzungslexikon aufgebaut.

zu 4: (RLIST)

Die Liste der Wörter, die auf syntaktische Relationierung hin überprüft werden müssen, umfasst Simplizia von Verben, Adjektiven und Substantiven, die entsprechend zu überprüfen sind (vgl. die Materialie "Erstellung morphosyntaktischer Relationen").

VI. ERGEBNISLISTEN

Daneben werden textbezogene Ergebnislisten erstellt. Sie werden wie folgt unterschieden:

1) EWL Ergebnisliste: Im Wörterbuch SADAW identifizierte Einträge

2) EKL Ergebnisliste: Im Wörterbuch KOMPZER identifizierte Einträge (samt erlaubten Zerlegungen)

3) EAL Ergebnisliste: Im Abkürzungswörterbuch ABKLEX identifizierte Einträge

4) EDL Ergebnisliste: Im Derivationslexikon identifizierte Einträge (samt vorgenommenen Zerlegungen)

Diese Listen (nebst einer Statistik) sollen die Wörterbuchpflege verbessern helfen; sie können ggf. zu einer stichprobenartigen Nachkorrektur herangezogen werden.

VII. SYNTAKTISCHE ANALYSE

Nach der Korrektur der Wörterbücher und einer Kontrolle der ergänzten Wörter kann die Satzanalyse durchgeführt werden. Hierzu ist der Testlauf ab INBx erneut zu starten.

(In einer späteren Version soll vorgesehen werden, dass "nur" noch solche Sätze ab INBx bearbeitet werden, für die eine Update-Operation (Text- oder Wörterbuchkorrektur) vorgenommen wurde. Die restlichen - zuvor 'korrigierten' - Sätze könnten ab LESx bearbeitet werden.)

Die syntaktische Analyse durchläuft mehrere Stadien, die u.a. von Parametern beeinflusst werden, die extern durch den Bearbeiter vorgegeben werden. Üblicherweise gelten folgende Regelungen:

(a) Die Analyse berücksichtigt die Groß-/Kleinschreibungsmarkierung, d.h. Wörter, die im

Satzinnern beim ersten Buchstaben eines Wortes großgeschrieben sind, gelten als Substantive (oder Substantivierungen); Wörter, die an entsprechender Stelle kleingeschrieben sind, stellen "Nicht-Substantive" dar. Diese Verfahrensweise verkürzt den Analyseprozess und verbessert (leicht) die Ergebnisse (grundsätzlich wäre das Saarbrücker Verfahren auch in der Lage, diese Information nicht auszunutzen.)

(b) Statt der möglichen 12 "Lesartenketten" für Wortklassenfolgen werden nur die beiden "wahrscheinlichsten" weiter verfolgt. Dies senkt erheblich den Rechenzeitbedarf, kann aber für Teilstrukturen zu Strukturfehlern führen.

(c) Als "Lösung" für die Weiterverarbeitung in CTX wird die 1. Struktur herangezogen, die die Satzanalyse als kompletteste ermittelt hat. Etwaige alternative strukturelle Möglichkeiten, die "grammatikalisch" im Sinne des Systems möglich sind, werden nicht beachtet.

d) Bei sehr komplexen Satzstrukturen kann eine korrekte Struktur wegen Überschreitens von Rechenzeitschranken eines Analysebausteins ggf. nicht ermittelt werden.

(e) Indiz für "unzureichend" ermittelte Satzstrukturen ist die Statusdatei, die für jeden Satz entsprechende Zwischenzustände festhält. Ein Satz ist mit großer Wahrscheinlichkeit korrekt strukturiert, wenn die Analysetiefe 6 (Komplementanalyse) erreicht wurde. Konnte keine Nominalgruppenanalyse erfolgen (Analysetiefe kleiner als 4), so werden für einen Satz auch keine präkoordinierten Begriffe (komplexe Deskriptoren) ermittelt.

Bei Bedarf können die Kriterien (Parameter) allgemein oder im Einzelfalll modifiziert werden, im Extremfall ist auch (z.B. durch Präkodierung) eine Verbesserung der Strukturergebnisse möglich. Für die "Produktversion" ist jedoch der Standardfall zu empfehlen, da aus verschiedensten Gründen (z.B. "defekte" Ausgangssätze) kaum eine "hundertprozentige" Lösung erreicht werden kann.

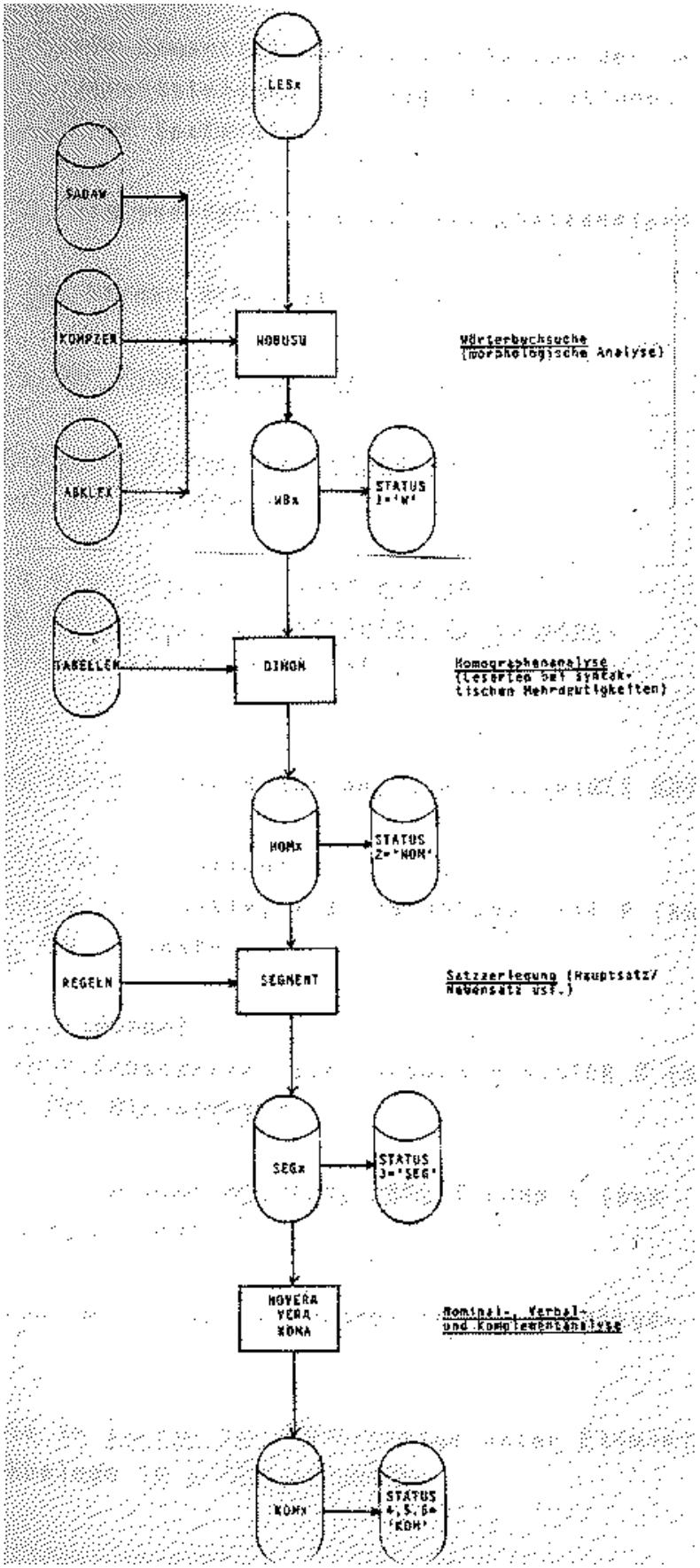
LESx
SADAW
KOMPZER
ABKLEX
HOBUSU
Wörterbuchsuche (morphologische Analyse)
HBx
STATUS 1='W'
TABELLEN
DIHOM
Homographenanalyse (Lesarten bei syntak. tischen Mehrdeutigkeiten)
HOMx
STATUS 2='HOM'
REGELN
SEGMENT
Satzzerlegung (Hauptsatz/ Nebensatz usf.)
SEGx
STATUS 3='SEG'
NOVERA VERA KONA
Nominal-, Verbal- und Komplexentanalyse
KOMx
STATUS 4,5,6= 'KOM'

Abb. 5 Satzanalyse Grobdarstellung)

VIII. DESKRIPTORERSTELLUNG

Die letzte Phase (ohne Semantik) umfasst den Aufbau der Deskriptorenliste und die Bereitstellung von Begriffsrelationen aus den Ergebnissen der Textanalyse. Grundlage sind die Analyseergebnisse aus VII (Satzanalyse). Folgende Funktionen werden erbracht:

(1) Eliminierung der Funktionswörter
(2) Bereitstellung der Einwortdeskriptoren in der jeweiligen Grundform
(3) Bereitstellung der Mehrwortdeskriptoren in der Form Deskriptor-1, Deskriptor-2, Relator, wobei als Relator verzeichnet ist:
    - 0 (leer)
      Adjektiv-Substantiv-Relation, z.B. KOERNIGES GUT
    - P (Präpositionalrelation)
      Substantiv-Substantiv, z.B. ANPASSUNG TEXT P (aus "Anpassung an Texte")
    - G (Genitivrelation)
      Substantiv-Substantiv, z.B. BESUCH MINISTER G (aus "Besuch des Ministers")
    - K (Konjunktionsrelation, z.B. ARBEIT LOHN K (aus "Arbeit oder Lohn"))

Die Deskriptoren werden wiederum in zwei Formen bereitgestellt:

(a) Dokumentweise in sortierter Reihenfolge unter Eliminierung der Mehrfachbelege in einem Dokument

(b) Dokumentweise in der Reihenfolge des Auftretens im Text, d.h. ggf. unter mehrfacher Aufführung des gleichen Deskriptors.

In allen Fällen von Typ (b) werden zusätzlich mitgeliefert:

- die Satznummer
- die Wortnummer im Satz
- die Kennzeichnung der Analysetiefe

(vgl. die Dateibeschreibung "Deskriptordatei")

(4) Bereitstellung von Derivationen und Synonymen

Für alle Einfachwörter werden die (in DERLEX/FALEX verzeichneten) Derivationen und Synonyme (z.B. bei Abkürzungen) explizit aufgeführt, wobei über die gesamte bearbeitete Dokumentmenge sortiert wird und Mehrfachbelege getilgt sind. Neben dem jeweiligen Wortlaut ist die Relation angegeben. Zusätzlich ist die Satz- und Wortnummer des Erstbeleges angeführt (vgl. die Dateibeschreibung "Derivationsdatei").

(5) Bereitstellung von Dekompositionsergebnissen

Für alle Einfachwörter, die in "sinnvolle" Bestandteile zerlegt sind, werden anhand des Zerlegungslexikons KOMPZER die Elemente in der Form Kompositum - Zerlegungselement aufge-

führt. Doppelbelege werden eliminiert, die Datei ist nach dem Kompositum alphabetisch sortiert. Zusätzlich wird die Satz- und Wortnummer des Erstbeleges mitgeführt (vgl. die Dateibeschreibung "Zerlegungsdatei").