The editors of the series APPLICATIO are

prof.dr. H. Leene, prof.dr. E. Talstra
Faculty of Theology of the Vrije Universiteit
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

The following titles in the series APPLICATIO are still available:

2. Amos. Concordance and Lexical Surveys
   A. van der Wal, E. Talstra (1984)
3. Amos. A Classified Bibliography
   A. van der Wal (1986)
4. Trito-Isaiah. An exhaustive Concordance of Isa. 56-66
   A. van der Wal (1988)
7. Computer Assisted Analysis of Biblical Texts
9. A Prophet on the Screen. Computerized Description and Literary Interpretation
   of Isaiatic Texts
    in the "Werkgroep Informatica" Computer Texts of the Hebrew Bible
    A.T.C. Verheij (1994)
    of Old Testament Computer Text of the Hebrew Bible
    J.W. Dyk (1994)
    A Computer-assisted Analysis and Textual Interpretation
    Towards a Syntactic Text Segmentation Model for Biblical Hebrew
15. Ad Fontes! Quellen erfassen - lesen - deuten
    Was ist Computerphilologie?

AD FONTES!
QUELLEN ERFASEN - LESEN - DEUTEN

WAS IST COMPUTERPHILOLOGIE?
Ansatzpunkte und Methodologie - Instrumente und Praxis

Contributions to the Conference "Computerphilologie"
November 5-8, 1998
Ernst-Moritz-Arndt-Universität Greifswald

Christof Hardmeier
Wolf-Dieter Syring
Jochen D. Range
Eep Talstra

Adam. VU University Press 2000

pp. 33-68
Genese und Kategorienentwicklung der WIVU-Datenbank

oder: ein Versuch, dem Computer Hebräisch beizubringen

Eep Talstra und Constantijn Sikkel (Amsterdam)

Einleitung

Wie kann die Aufgabe bewältigt werden, eine Text-Datenbank zu erstellen, die für Linguisten und Literaturwissenschaftler, für Exegeten und Übersetzer als Hilfestellung bei ihrer Arbeit mit Sprachen und Texten zur Verfügung stehen soll? Die hebräische Datenbank der WERKGROEP INFORMATICA an der Theologischen Fakultät der VRIJE UNIVERSITEIT in Amsterdam (WIVU) stellt einen der möglichen Ansätze dar, diese Aufgabe zu lösen.


1. Eine Datenbank syntaktisch analysierter biblisch-hebräischer Texte – ein Resultat oder ein Instrument?

Die bisher erhältlichen Computerprogramme zur Untersuchung biblischer Texte stützen sich auf Datenbanken, die linguistische Ana-

- Zum einen ist die Frage zu beantworten, inwieweit eine sprachliche Beschreibung von Textebenen oberhalb der Wortebe möglich ist. Kann man linguistische Kategorien entwickeln, um Satzgefüge und andere Erscheinungen der Textebene in systematischer Weise grammatisch zu analysieren?


2. Textsyntax: die Analyse der sprachlichen Ebenen in Texten


2.1 Sprachsystem und Textkomposition

Theoretisch stellt man sich die linguistischen Ebenen in einem Text häufig als einfache Hierarchie vor: der Text als höchste Einheit gliedert sich in kleinere Einheiten, die wiederum in noch kleinere Einheiten segmentiert werden. Die auf diese Weise definierten linguistischen Ebenen können in einem abstrakten Modell veranschaulicht werden:

<table>
<thead>
<tr>
<th>Segментирование der Textoberfläche</th>
<th>Analyse-Vorschlag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satzteile</td>
<td>einz. Satz</td>
</tr>
<tr>
<td>Satzteile</td>
<td>zus. Satz</td>
</tr>
<tr>
<td>Wortgruppe</td>
<td>Wortgruppe</td>
</tr>
<tr>
<td>Wort</td>
<td>Wort</td>
</tr>
<tr>
<td>Wörter</td>
<td>Wörter</td>
</tr>
<tr>
<td>Text</td>
<td>Text</td>
</tr>
<tr>
<td>Sätze</td>
<td>Sätze</td>
</tr>
<tr>
<td>Worte</td>
<td>Worte</td>
</tr>
<tr>
<td>Wie beschrieben</td>
<td>Wie beschrieben</td>
</tr>
<tr>
<td>Wie formuliert</td>
<td>Wie formuliert</td>
</tr>
<tr>
<td>Wie deuten</td>
<td>Wie deuten</td>
</tr>
</tbody>
</table>

Figuur 1: Idealtypische textuelle Hierarchie

Das praktische Problem besteht nun darin, daß ein „echter“ Text diesem Idealbild kaum entspricht. Sehr häufig läßt sich in einem konkreten Text eine Mischung von Ebenen und Kategorien feststellen:


4 Vgl. W. RICHTER (Satz, 1980) und F. ANDRESEN (Sentence, 1974).

Sätze können eine Funktion auf der Wort- oder Phrasenebene haben; einzelne Wörter können als Sätze fungieren; ein Text (als Segment der direkten Rede) kann in narrativem Kontext als Objekt zu einem Verb angesehen werden, usw. Aus dieser Tatsache folgt unmittelbar, daß die Ebenen der linguistischen Textbeschreibung nicht mit den Ebenen der Textkomposition identisch sind. Die theoretischen linguistischen Ebenen können in einer strenge Hierarchie beschrieben werden, die faktischen textlichen Ebenen jedoch nicht, da sie selten in einer hierarchischen Ordnung auftreten; sie können vielmehr mehrfach ineinander eingebettet werden.

Als Testfall für das Problem der linguistischen Kategorisierung kann folgendes Beispiel dienen: wieviel Sätze enthält die Textpassage Numeri 13,32f.?

Es zeigt sich, daß eine Hierarchie von Ebenen tatsächlich existiert, im syntaktischen Aufbau dieses konkreten Textes jedoch die Hierarchie der Elemente nicht mit der der Textebenen übereinstimmt. Die Ebenen als Elemente des grammatischen Systems lassen sich in einem Text bestimmen, ergeben aber keine eindeutige Hierarchie – etwa nach dem Bild einer Pyramide. Die sprachlichen Ebenen des Systems sind nicht identisch mit den textlichen Ebenen der Komposi-
tion. Bei der Arbeit an einer Textdatenbank besteht daher auch nach der Definition der linguistischen Beschreibungskategorien noch nicht die Möglichkeit, die Syntax konkreter Texte zu speichern. Dieses Problem bietet eine Erklärung dafür, daß die Mehrzahl der heutigen Computerinstrumente zur Textanalyse über die Worte ebene nicht hinausgehen. Auf dieser Ebene ist der Unterschied zwischen langue und parole kaum wahrzunehmen. Wenn man jedoch Texte auf höheren Ebenen analysiert, nehmen die Unterschiede zwischen Sprachsystem und Texthierarchie erheblich zu.


Figur 2b: Num 13,32f. nach der Übersetzung Martin Luthers (Stuttgart 1984)

Daher besteht die wichtigste Frage zunächst nicht darin, wie man einen konkreten Text linguistisch analysieren und möglicherweise erklären kann. Wenn das Ziel der Datenbank-Entwicklung ein Instrument zur linguistischen und literarischen Textanalyse ist, dann sollten gerade nicht Interpretationen, sondern vielmehr Daten gespeichert werden. Vordringlich ist daher die Klärung der Frage, wie die komplexen Verhältnisse in empirischen Texten kategorial beschrieben werden und in einer Datenbank gespeichert werden können: Welche linguistische Segmentierung ist notwendig, welche Eigenschaften zeigen die Segmente, und welche Beziehungen haben sie zueinander?

7 Die Wörter „ein böses Gerücht auf“ wurde – in Parallelen zum hebräischen Text – nicht übersetzt. Wörter wurden in
risierung zueinander in Beziehung gesetzt. Das geschieht aber nur teilweise: die Verbindung der Hauptsätze – wie z.B. die mit „und“ (hebr. ֵוָּלָא) verknüpften Sätze Numeri 13,32c.33a.33b.33c – wird bei der syntaktischen Analyse nicht berücksichtigt.

Dem von Richter verwendeten Verfahren zur Satzsegmentierung ist zuzustimmen: zunächst ist der Text linear zu segmentieren, auch wenn dabei vorläufig unvollständige Sätze auftreten (Numeri 13,32a.32c); anschließend sind die gewonnen (Teil-) Sätze nach funktionalen Kriterien zu kombinieren.9 Die hinter der Edition der BH stehende Datenbank stellt ein vorläufiges Instrument dar, das Material für weitere Untersuchungen zur Verfügung stellt.10


2.2.2 Descendentes Verfahren


9 Diese Vorgehensweise empfiehlt sich auch für die Kombination von Wörtern zu Wortgruppen.
10 W. Eckardt (Analyse, 1987).
11 Vgl. unten Kapitel 4.2.
12 Vgl. J. Sailhamer (Approach, 1990); J. Sailhamer (2 Samuel 13, 1992); E. Gö...

\[\text{Ref. Form} \quad \text{CT} \quad \text{DT ML ATP}\]

<table>
<thead>
<tr>
<th>Ref. Form</th>
<th>CT</th>
<th>DT ML ATP</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAYYIQTOL</td>
<td>V Nar 1 0+=</td>
<td></td>
</tr>
<tr>
<td>NOM ABS</td>
<td>Dis 2</td>
<td></td>
</tr>
<tr>
<td>+ NC</td>
<td>N Dis 2</td>
<td></td>
</tr>
<tr>
<td>W+NC</td>
<td>N Dis 2</td>
<td></td>
</tr>
<tr>
<td>W+X-QAT</td>
<td>V Dis 2 0++</td>
<td></td>
</tr>
<tr>
<td>WITHQAT</td>
<td>V Dis 2</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Dis 2</td>
<td></td>
</tr>
<tr>
<td>Clause Type (V = Verbsatz; N = Nominalsatz)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discourse Type (Nar = narrativ; Dis = diskursiv)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metacommunicative Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actant Time Place</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figur 4: Datenanalyse nach John Sailhamer (Num 13,32f.)

Sailhamers Analyse kombiniert auf diese Weise linguistische und literarische Daten: Sätze und Kommunikationsebenen, „Spieler im Text“, lokale und temporale Verweisungen.13 Diese Kombination tekstlinguistischer und literarischer Interessen hat weitreichende Auswirkungen für die Datenanalyse.


13 Vgl. die Beschreibung der Datenbank von A. Den Exter Blokland (Clause-Analysis, 1990) 75: „A Combination of Linguistic and Literary Data."
14 Vgl. A. Den Exter Blokland (Clause-Analysis, 1990) 79: „Participial and infinitival phrases, except the infinitive absolute ..., are excluded: on the surface level they are not clauses.“
matikalische Wiederaufnahmen achten? Passen z.B. Person und Numerus eines Prädikats zu Person und Numerus eines Pronominalsuffixes in einem vorhergehenden Satz oder nicht?

Zweitens führt das literarische Interesse dazu, dass die Textebenen nach literarischen und nicht nach linguistischen Kriterien unterschieden werden. Wenn z.B. innerhalb der direkten Rede – vornicht in narrativem Kontext verwendete – Wayyiqtol-Sätze vorkommen, kann man aus linguistischen Gründen durchaus annehmen, dass hier ein narrativer Textteil in eine direkte Rede eingebettet ist.16 Nach SAILHAMERS Auffassung werden jedoch auch diese Sätze als diskursiv gekennzeichnet, weil sich die Kommunikationsebene nicht geändert hat.17

Drittens bieten SAILHAMERS Daten auch Informationen über die Nominalisierung und die Renominalisierung von Aktanten, was zunächst als Vorteil anzusehen ist. Es handelt sich dabei jedoch nicht um textgrammatische, also textinterne Markierungen; vielmehr gehören die Aktanten bei SAILHAMER – ganz im Sinne der textlinguistischen Analysen von Elisabeth Gulich und Wolfgang Raible – zu den Markierungen der Welt außerhalb des Textes, d.h. zu den Verweisen auf die Figuren sowie auf Ort und Zeit der Handlung. Daher repräsentiert SAILHAMERS Datenbank ein eher literarisches und exegetisches Anliegen. Ohne Zweifel sind Angaben zu den Aktanten im Text für Exegese und Übersetzung wertvoll, aber es bleibt zu fragen, ob sich diese Art der Textanalyse, eben weil sie einer „top-down“-Vorgehensweise folgt, nicht von vorneherein zu sehr auf die (Gesamt-) Interpretation des Textes stützt und zu wenig auf die textgrammatischen (Detail-) Analyse.

3. Zur textlinguistischen Analyse in der WIVU-Datenbank

Anhand eines kurzen Textes aus Numeri 13,32 wird im folgenden gezeigt, welche Entscheidungen bei der Bestimmung der Kategorien einer Datenbank für grammatische und linguistische Forschung getroffen werden müssen.

16 S. unten Figur 6 zu Num 13,32-33.

---

Die fünf Wörter in Figur 5 bilden ein ziemlich komplexes sprachliches Gefüge. Die Teile haben eine Reihe von Eigenschaften und Beziehungen, die ein aufmerksamer Leser jedoch ohne besondere Schwierigkeiten feststellen kann. Daher sind zunächst zwei allgemeine Fragen zu klären. Welche linguistischen Daten, die wir als Leser hier feststellen können, sollen in der Datenbank gespeichert werden? Und inwiefern kann der Computer die weitere Systematisierung der Daten nach linguistischen Kriterien unterstützen, so dass bei fortschreitender Analyse eine Möglichkeit zur Konsistenzüberprüfung zur Verfügung steht?


Das sich daraus ergebende Problem der Datenstrukturen wird in der WIVU-Datenbank gelöst durch eine präzise Unterscheidung zwischen Einheiten der linguistischen Distribution (als Beschreibung ihres Aufbaus und ihrer Verteilung) und Einheiten der linguistischen Funktion. Dabei steht der Gedanke im Vordergrund, daß auf allen Ebenen die Ermittlung linguistischer Funktionen (z.B. „Attribut“, „Subjekt“, „Nebensatz“) erst nach der Segmentierung des Textes nach distributionalen Kriterien erfolgen kann. Denn eine Abgrenzung von Wortfolgen oder Wortgruppen ist möglich, auch ohne deren Funktion (z.B. als Satzkonstituenten oder als Satz) festzulegen. Im folgenden (3.1 - 3.4) wird diese Art von Textanalyse schrittweise dargestellt.

3.1 Analyse der Wortgruppenabgrenzungen

Auf der Grundlage eines bereits morphologisch analysierten Textes kann man mit Hilfe eines Computerprogramms nachprüfen, ob in diesem Text z.B. eine Wortfolge mit folgenden Eigenschaften existiert:

<table>
<thead>
<tr>
<th>Nomen</th>
<th>Artikel</th>
<th>Nomen</th>
</tr>
</thead>
</table>

Wenn eine solche Wortfolge auftritt (z.B. מְרָא הָעֹלָה = „böses Ge-rückt von dem Land“ in Numeri 13,32), ist vom Benutzer des Programms zu entscheiden, ob die drei Wörter eine grammatisch akzeptable Konstruktion darstellen. Im positiven Fall muß das Programm die gefundene Kombination nicht nur im Text als Wortgruppe markieren, sondern auch festhalten, welche zur Bildung der Wortgruppe relevanten morphologischen Eigenschaften die Bestandteile haben. Was als relevant anzusehen ist, wird nach einem – vom Benutzer entworfenen – vorläufigen morphologischen Paradigma entschieden.

So bedeuten die Feststellungen „הָעֹלָה hat die Nominalendung נ and kein pronominales Suffix“ und „מְרָא hat die Nominalendung ein Nullmorphem und kein pronominales Suffix“ folgendes: der Numerus von מְרָא ist 'singular', der Status ist 'unkannt'; der Numerus von הָעֹלָה ist 'singular', der Status ist 'unkannt'; beide Wörter tragen 'kein Suffix'. Auf diese Weise kann der Benutzer eine erste Regel zum Aufbau einer Nominalphrase formulieren lassen:

| Nomen (Status: unbekannt; Numerus: singular; Suffix: 0) | Artikel |
| Nomen (Status: unbekannt; Numerus: singular; Suffix: 0) |

→ NP (determiniert)

18 Der sog. Leningradensis liegt der Biblia Hebraica Stuttgartensia (BHS, 1877–1975) und der im Erscheinen befindlichen Biblia Hebraica Quinta

19 Die Kategorien des von Ferenc Postma und Arián J. C. Verheij erstellten, morphologisch analysierten Textes der „Werkgroep Informatica“ werden
Durch die Verwendung einer „Rückkopplung“ kann das Programm nun auf weitere linguistische Werte schließen: der 'unbekannte' Status des ersten Nomens kann als 'status constructus', der des zweiten Nomens als 'status absolutus' qualifiziert werden; eine Nominalphrase mit Artikel muß als 'determiniert' gelten. Streng genommen hat dieses Ergebnis nur hypothetischen Wert. Es stellt noch keine allgemeine grammatische Regel dar, sondern eine vorläufige Hypothese zu den zugrunde liegenden Daten, die sich nur anhand weiterer Textdaten bestätigen oder korrigieren läßt.


In Analogie zu dem bereits behandelten Beispiel kann das Programm die Gruppe אלברב חישה (= „bei den Söhnen Israels“) in Numeri 13,32 als neue Präpositionalphrase in die Liste einfügen:

| Präposition (Suffix: 0) |
| Artikel |
| Nomen (Status: constructus; Numerus: plural; Suffix: 0) |
| Eigenname |
| → PP (determiniert) |

Ein weiteres Beispiel in Numeri 13,33 אלֶהָ הָעֵשֶׁנָה יַעֲקֹב (= „die Riesen, die Söhne Enaks vor den Riesen“) zeigt die rekursive Vorgehensweise. Bei der distributiven Segmentierung wird noch nicht festgelegt, daß die acht Wörter dieser Gruppe einen Satzbestandteilen bilden, in diesem Falle ein Objekt (ן), vielmehr werden zunächst drei Gruppen unterschieden und zwar nach zwei Strukturmustern:

| Präposition (Suffix: 0) |
| Artikel |
| Nomen (Status: absolutus; Numerus: plural; Suffix: 0) |
| Eigenname |
| → PP (determiniert) |

Erst im Anschluß daran werden - unter Anwendung desselben Verfahrens - die Gruppen mit weiteren Mustern verglichen. Im genannten Beispiel kann man ein neues Strukturmuster definieren, das aus einer Kombination der beiden verwendeten Muster besteht. Von diesem neuen, erweiterten Muster ausgehend kann das Programm die zweite Wortgruppe als Apposition qualifizieren:

| Präposition (Suffix: 0) |
| Artikel |
| Nomen (Status: absolutus; Numerus: plural; Suffix: 0) |
| Nomen (Status: constructus; Numerus: plural; Suffix: 0) |
| Eigenname |
| → PP (determiniert) + NP (determiniert; Apposition) |

Dieses Beispiel zeigt, daß in bestimmten Fällen mit dem distributiven Verfahren - aufgrund der Informationen über Wortart und morphologische Eigenschaften - nicht nur eine Segmentierung des Textes durchgeführt, sondern auch spezifische Beziehungen zwischen den neu gebildeten Einheiten festgestellt werden können. In den meisten Fällen gelingt das nicht vollständig. So ist die Beziehung zur dritten Gruppe (הָעֵשֶׁנָה) auf dieser Ebene der Analyse noch nicht entscheidbar, weil man dazu die Information benötigt, ob diese Gruppe auf Satzebene mit dem Prädikat (d.h. als präpositionales Objekt zu רָאִים) oder auf Phrasenebene mit der Appositionsgruppe verbunden ist (d.h. als Spezifizierung der רָאִים). Diese Entscheidung kann erst im Zuge der Analyse der Satzkonstituenten (Prädikat, Subjekt, Objekt, Komplement, usw.) getroffen werden.

Die bei der distributiven Segmentierung gewonnenen Einheiten sind somit grammatisch durchaus korrekt, stellen aber nicht immer vollständige linguistische Größen dar. Daher werden diese Segmente in der WIVU-Datenbank als „Atome“ bezeichnet. Mit diesem

Dieses Verfahren erlaubt die korrekte Beschreibung eingeschobener Elemente (embedding) und dadurch entstehender Lücken (gapping) in funktionalen Einheiten: während Atome stets lineare, unterbrochene Wortfolgen sind, lassen die aus ihnen zusammengesetzten funktionalen Einheiten Lücken bzw. Einschübe zu, wie das Beispiel einer der geteilten Wortgruppe in Numeri 13,26 zeigt:

<table>
<thead>
<tr>
<th>[אָדָמֶלִירְדָו]</th>
<th>[ירָבָר]</th>
<th>[אֵמוֹת]</th>
<th>[ישָׁב]</th>
<th>[יִ]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obj2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obj1  - - - - - Obj1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2 Analyse der Satzabgrenzungen


Genese und Kategorienentwicklung der WIVU-Datenbank


Beim einem Satz sind prinzipiell alle – im Rahmen der jeweiligen Verbalen zugelassenen – Verbindungen möglich, wegen der nominalen Form des Partizips kommen die Möglichkeiten der constructus-Verbindung hinzu. Partizipien eröffnen daher Konstruktionsmöglichkeiten auf zwei Ebenen. Im Fall einer verbalen Valenzverbindung resultieren Formen auf der Ebene der Satzsyntax; bei einer nominalen Verbindung ergeben sich Formen auf Phrasenebene.

Ein Problem entsteht insbesondere dann, wenn aufgrund der Nominalendungen nicht über den Status (absolutus oder constructus) eines Partizips entschieden werden kann. Wie kann man in diesem Fall eine Hypothese zum Satzbau bzw. zum Phrasenbau begründen? In der Anfangsphase der Textcodierung kann man nur aufgrund einer naheliegenden Interpretation eine erste Hypothese formulieren. Liegen bereits mehrere vergleichbare Fälle in den Textdaten vor, kann der Benutzer diese für seine Hypothesenbildung heranziehen.

Das aus der WIVU-Datenbank hervorgegangene Programm Quest 21 bietet zwar noch nicht die Möglichkeit, nach Subjekt und Prädikat zu suchen, aber wenn man nach einem einfachen Muster (wie z.B. „Nomen + Partizip mit femininer Endung + Nomen mit Suffix“) fragt, findet man genügend Stellen, die zur weiteren linguistischen Analyse hilfreich sein können.

Einerseits finden sich Belege, die eine Interpretation auf Satzenebene nahelegen, bei der eine Partizipia-Konstruktion alsattributiver oder Adverbial-Satz verwendet wird und mit Subjekt, Objekt oder auch mit anderen Konstituenten verbunden sein kann.

21 E. Talstra/C. Hardmeier/A. Grovis (Ouest. 1992): vel. E. Talstra (Demon-
3.3 Analyse der Texthierarchie

Die Datenanalyse auf höheren Ebenen (Satzgefüge und Textsyntax) wird mit vergleichbaren Prozeduren durchgeführt. Da die Entwicklung der verschiedenen Stufen dieses Verfahrens mehrfach (meist im Zusammenhang mit der Diskussion aktueller Fragen der Exegese oder der Grammatik) dargestellt wurde, werden wir uns in den nächsten Abschnitten auf eine Präsentation der aus diesem Verfahren erwachsenen Daten beschränken.

3.4 Distributionale und funktionale Einheiten

Die oben angeführten Beispiele der Phrasen- und der Satzanalyse basieren auf einem zweistufigen Verfahren der linguistischen Analyse.


Die „Atome“ als elementare Kategorien sind, auch wenn sie teilweise keine vollständigen linguistischen Größen repräsentieren, für zwei Zwecke hilfreich. Zum einen fügen sie sich in die Idealvorstellung einer strengen linguistischen Hierarchie: Atome passen genau ineinander und aneinander; sie bilden Einheiten, die sich nicht überschneiden und den Text linear und lückenlos segmentieren. Auf diese Weise wird die textuelle Hierarchie für ein Computerprogramm bzw. eine Datenbank darstellbar. Zum anderen dienen die Atome als eine Art Metasprache: sie ermöglichen es, in einem Text die funktionalen linguistischen Einheiten als Kombination von Atomen zu definieren und so auch lückenhafte Einheiten zu erfassen; so kann z.B. ein Satz Nr. 10 aus zwei Satz-Atomen (Nr. 12 und Nr. 14) bestehen, die durch ein eingebettetes Element (Satz-Atom Nr. 13) voneinander getrennt sind.

- In einem zweiten Arbeitsgang werden die Atome zu funktionalen sprachlichen Einheiten kombiniert. Diese repräsentieren die praktische

---

22 Vgl. C. HARDMEIER/E. TALSTRA (Spachgestalt, 1989); E. TALSTRA (Hebrew Syntax, 1991); E. TALSTRA (Menora, 1996); F. TALSTRA (Hierarch. 1997).

Aufgrund der oben ausgeführten Beispiele können nun einige vorläufige Ergebnisse und die sich daraus ergebenden Aufgaben formuliert werden.

Es ist deutlich, daß die erste Phase der Erstellung einer Textdatenbank an sich noch keine für Theologen oder Literaturwissenschaftler ertragreiche Forschung darstellt. Sie ist jedoch insofern eine grundlegen- de Arbeit, als in der Kombination von distributionalen und funktionalen Verfahren ein Modell geschaffen wird, das die Kategorien zur Beschreibung und Speicherung der unterschiedlichen linguisti- schen Daten überhaupt erst bereitstellt.

Anschließend werden auf der Basis der so entstandenen Datenbank weitere Analysen durchgeführt, um die Daten zu präzisieren bzw. zu korrigieren und auf diesem Wege die Konsistenz der analysierten Daten und der dazugehörigen Grammatik zu erhöhen. Erst auf dieser Grundlage einer konsistenten Datenanalyse kann die Diskussion mit Exegeten und Übersetzern über die Interpretation der Texte beginnen. 24


24 Vgl. die Bemerkung N. Lohfink (Bewegung, 1995) 97 über die „morali- sche Vernichtung", mit den neuen Instrumenten die Daten zu den klas-
4. Die Kategorien der WIVU Datenbank

Anhand einiger Verse aus Numeri 13 werden im folgenden die in der WIVU-Datenbank verwendeten Kategorien beschrieben und erläutert, aus welchen Gründen bzw. mit welchem Ziel sie entwickelt wurden. Aus Kap. 3 ging bereits hervor, daß Kategorien aus drei Beschreibungsebenen kombiniert werden; es gibt

- Kategorien, die einen Text als Dokument (hierarchisch) gliedern,
- Kategorien, die einen Text als (lineare) Folge distributionaler linguistischer Einheiten ("Atome") darstellen, und schließlich
- Kategorien, die einen Text als ein (komplexes) Gefüge funktionaler linguistischer Einheiten beschreiben.

Das Beispiel einer komplexen Wortgruppe in Numeri 13,26 zeigt, daß es sich bei der linguistischen Beschreibung nicht nur um die Segmentierung des Textes in linguistische Einheiten geht, sondern auch um die Bestimmung der inneren und äußeren Beziehungen dieser Einheiten:

Numeri 13,26 ... וְבִּשַׁלְחֵנָהּ אָלָמָּהּ אֵלָכֶלֶתָה בָּניָּירָא ... 
Diese Wortfolge besteht – gemäß der zunächst durchzuführenden distributionalen Analyse – aus sechs Segmenten ("phrase atoms"):

<table>
<thead>
<tr>
<th>וְבִּשַׁלְחֵנָהּ</th>
<th>אָלָמָּהּ</th>
<th>אֵלָכֶלֶתָה</th>
<th>בָּниָּירָא</th>
</tr>
</thead>
<tbody>
<tr>
<td>[וְבִּשַׁלְחֵנָהּ]</td>
<td>[אָלָמָּהּ]</td>
<td>[אֵלָכֶלֶתָה]</td>
<td>[בָּниָּירָא]</td>
</tr>
</tbody>
</table>

In einem weiteren Bearbeitungsgang dienen zwei weiteren Analyseprozesse der Bestimmung der linguistischen Beziehungen und Funktionen: jedes Segment kann innere (z.B. Genitiv- oder Attribut-) Beziehungen und äußere Verbindungen (z.B. zu parallelen Einheiten oder zum Prädikat des Satzes) realisieren:

| Innere Beziehungen (reg = regens; rec = rectum):
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>... וְבִּשַׁלְחֵנָהּ [rec→] [e-reg] וְ בִּישַׁלְחֵנָהּ שָׁתָה [rec→] [e-reg] ...</td>
</tr>
</tbody>
</table>
| Äußere Verbindungen (p = parallel):
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>... וְבִּשַׁלְחֵנָהּ [p→] [e-p] וְ בִּישַׁלְחֵנָהּ בָּניָּירָא [p→] ...</td>
</tr>
</tbody>
</table>
| Äußere Verbindungen (cmpl = Komplement; pred = Prädikat):
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>... וְבִּשַׁלְחֵנָהּ [cmpl] וְ בִּישַׁלְחֵנָהּ בָּניָּירָא [pred] ...</td>
</tr>
</tbody>
</table>

Die Beschreibung der Textdaten erfolgt also einerseits mit Dokument-orientierten Kategorien, die die Textgliederung – z.B. nach Kapiteln und Abschnitten bzw. Versen – festhalten, und andererseits mit linguistischen Kategorien, die auf jeder linguistischen Ebene die

<table>
<thead>
<tr>
<th>distributionale Einheiten</th>
<th>funktionale Einheiten</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. Ebene: Paragraph</td>
<td></td>
</tr>
<tr>
<td>Analyse: Satztypen und Akte</td>
<td></td>
</tr>
<tr>
<td>↑</td>
<td>↑ → → Analyse: Hauptsätze und Konstituent-Sätze ↓</td>
</tr>
<tr>
<td>5. Ebene: Satz-Hierarchie</td>
<td></td>
</tr>
<tr>
<td>↑</td>
<td>↑ Analyse: Textsyntax → → → → → → → → → → → → → → → → → → → → → → → → → → → ↑</td>
</tr>
<tr>
<td>4. Ebene: „clause atoms“</td>
<td></td>
</tr>
<tr>
<td>↑</td>
<td>↑ Analyse: Satzsyntax → → → → → → → → → → → → → → → → ↑</td>
</tr>
<tr>
<td>3. Ebene: „phrase atoms“</td>
<td></td>
</tr>
<tr>
<td>↑</td>
<td>↑ Analyse: Wortgruppensyntax → → → → → ↑</td>
</tr>
<tr>
<td>2. Ebene: Lexeme Status constr. oder abs./funktionale Wortart</td>
<td></td>
</tr>
<tr>
<td>↑</td>
<td>↑ Analyse: Morphologie/Lexikon → → → → → Worterhebung ↑</td>
</tr>
<tr>
<td>1. Ebene: Morpheme</td>
<td></td>
</tr>
</tbody>
</table>
| ↑ | ↑ erstes, aszendentes Verfahren ↑ zweites, deszendentes Verfahren ('bottom-up') ↓ ('top-down')

Figuur 7: Übersicht über den Analyse-Prozeß
4.1 In welcher Reihenfolge gewinnt man linguistische Daten?

Das Verhältnis distributionaler und funktionaler Daten


Wie in Kap. 3 gezeigt wurde, basiert das Verfahren auf dem Prinzip der Musterverwiedererkennung. Auf jeder Ebene bietet das Analyseprogramm Vorschläge an, die aus einer Systematisierung der bereits analysierten Daten gewonnen wurden. Dieses Vorgehen wird erst sekundär und nur dann, wenn genügend Daten für eine Konsistenzkontrolle vorliegen, durch daraus ableitbare grammatische Regeln oder lexikalische Informationen ergänzt, z.B. durch ein "Valenzwörterbuch" im Fall der Satzanalyse.

Die Daten werden in fünf Arbeitsgänge - von der Morphemebene bis zur Textebene aufsteigend - produziert. Der Schlußsatz in Numeri

5. Schlußsatz:

5.1 Satzstapelstyp:

[Stapel] i.
Verbalsatz/Narrativer Text

4.1 Satzanalyse:

[Obj] [Art] [Noun] [Prep] [Verb]

3.1 Phrasenanalyse:

[Art] [Noun] [Prep] [Verb]

2.1 Lexikalische Analyse:

(Art) (Noun) (Prep) (Verb)

1.1 Morphologische Analyse:

(Stabstätte) (Stabstätte) (Stabstätte) (Stabstätte)

Figur 8: Beispiel der "bottom-up"-Analyse von Ebene 1–5

13,26 illustriert einige der Kategorien, die in jedem Arbeitsgang hinzugefügt werden (Stufe 1–5 in Figur 8, vgl. auch Figur 7).

4.2 Welche Kategorien sind zur Textdaten-Speicherung nötig?
Ein Beispiel aus der WIVU-Datenbank

Die folgenden Figuren 9a–d zeigen anhand von vier Wörtern aus Numeri 13,32 die in der WIVU-Datenbank enthaltenen Daten auf Wort-, Phrasen- und Satz-Ebene.

Beispiel Numeri 13,32: analyzer for lexicon: lexicographical properties (verbs)

I. Dokument:

<table>
<thead>
<tr>
<th>Buch</th>
<th>Numeri</th>
<th>Numeri</th>
<th>Numeri</th>
<th>Numeri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kapitel</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Vers</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Halbvers</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Wortform</td>
<td>analyzer for lexicon: lexicographical properties (verbs)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wort-Nummer (absolut)</td>
<td>8765</td>
<td>8766</td>
<td>8767</td>
<td>8768</td>
</tr>
</tbody>
</table>

Figur 9a/9b: Dokument Eigenschaften / linguist. Eigenschaften auf Wortebene

II.1. Wort

II.1.1. Morphologie (p = paradigmatische Form; r = realisierte Form)

Präformativ (p) empty empty
Stammform (p) empty empty
Stammform (r) empty empty
Verbalform (p) empty empty
Verbalform (r) empty empty
Nominalform (p) empty empty
Nominalform (r) empty empty
SUFFIX (p) empty empty
SUFFIX (r) empty empty
LOCATIV (p) empty empty
LOCATIV (r) empty empty

II.1.2. Lexikon:

Lexem (p) analyzer for lexicon: lexicographical properties (verbs)
Lexem (r) analyzer for lexicon: lexicographical properties (verbs)

II.2. Wortfunktion:

Verbform function Ptz. Ptz.
Verbalstamm Qal Qal
Person empty empty empty 3.
Numerus sing. sing. plur. sing.
Status empty st.abs. st.abs. st.abs. st.abs. st.abs. st.abs. st.abs. st.abs.
SUFFIX empty empty 3.abs.
Genese und Kategorienentwicklung der WIVU-Datenbank

4.3 Wie können Textdaten für Abfragen genutzt werden? Möglichkeiten für und Erwartungen an ein Retrieval-Programm


4.3.1 Beispiele zur Nutzung paradigmatischer Daten

a) Wie findet man Sätze?

Für einen Hebraisten oder für einen Exegeten ist die Suche nach Sätzen eine - vielleicht doch nur scheinbar - schlichte Frage, für eine einfache Machine jedoch eine komplizierte Angelegenheit. Ein Retrieval-Programm muß bei dieser Frage bereits die Vorentscheidung treffen, daß im Bereich der linguistischen Daten nach funktionalen Sprachünheiten zu suchen ist und unterbrechene und unterbrochene Einheiten bei der Suche einkalkuliert werden müssen. Bei den in Figur 12 präsentierten Daten treten Sätze nicht als selbständige Größen auf, sondern nur als Teile größerer linguistischer Einheiten, z.B. als Konstituenten zusammengesetzter Sätze.25

Für einen Benutzer der Datenbank behält die Suchaufgabe die Gestalt einer einfachen Abfrage, z.B.: „Suche die in Numeri 13 enthaltenen Sätze“. Die Abarbeitung dieser Suchaufgabe geschieht auf der Basis der WIVU-Daten folgermaßen:

Zur „Treffermenge“ werden alle Wortfolgen gezählt, bei denen die Satz-Nummer und die Satzgefüge-Nummer übereinstimmen, z.B. in Numeri 13,32; Satz = 1 und zusammengesetzter Satz = 1, u.s.w. So erhält der Benutzer auch die Sätze, die durch Nebensätze in zwei oder mehr Abschnitte unterteilt worden sind. Die Suche nach spezifischen Satzformen oder Satzteilen ergibt sich als qualifizierte Teilmenge der so gefundenen Sätze:

---

25 Auf der Ebene der zusammengesetzten Sätze sind davon die größten...
Die Suche nach Nominalsätzen in Numeri beschränkt die Abfrage auf Sätze mit der Kenzeichnung „NS“(Nominalsatz), wobei wieder die Nummer des Satzes mit der des zusammengesetzten Satzes übereinstimmt muß.


b) Beispiele der Phrasen-Struktur

Die Analyseergebnis der Phrasen-internen Beziehungen wird als Zahlenkombination gespeichert. Ein Benutzer kann anhand dieser Daten sowohl die innere Struktur von Phrasen wie auch deren Teile untersuchen. So kann beispielsweise die vollständige Wortgruppe aus Numeri 5,19 dem Benutzer auf folgende Weise präsentiert werden, wobei die intern verwendeten Zahlenkombinationen als „Label“ dargestellt werden:

\[\text{[[\text{דוע} dem--]} \text{[[ו[\text{אט} rec--]} +\text{at[\text{דומ rec--}]} +\text{reg \{+\text{dem\} r\}}] \text{[m]} \text{[m]}\]

Weiterhin ergibt sich die Möglichkeit, nach Phrasen mit einer bestimmten inneren Beziehung zu suchen, z.B. nach Wortgruppen mit einem demonstrativen Element. Zum Ergebnis gehört dann die folgende Wortgruppe:

\[\text{[[\text{דוע} dem--]} \text{[[ו[\text{אט} rec--]} +\text{at[\text{דומ rec--}]} +\text{reg \{+\text{dem\} r\}}] \text{[m]} \text{[m]}\]

Die Daten erlauben auch, die Struktur der Wortgruppenbildung im Hebräischen zu untersuchen, indem man z.B. eine Folge von Wortarten deﬁniert und die diesem Muster entsprechenden und im Suchbereich vorhandenen Phrasen als Ergebnis erhält. Einige Beispiele:

Nomen - Nomen

| Numeri 35,14 | [reg : rec--] [+reg \{+\text{reg} \\text{rec--}\}] [+reg \{+\text{reg} \\text{rec--}\}] |
| Numeri 28,6 | [reg : rec--] [+reg \{+\text{reg} \\text{rec--}\}] |
| Numeri 28,3 | (aj = „adjective“) |

Nomen - Adjektiv

4.3.2 Beispiele zur Nutzung syntagmatischer Daten

a) Syntax und komplexe Strukturen

Für einige Untersuchungen werden nicht nur Sammlungen paralleler Fälle benötigt, sondern eher ein Überblick über bestimmte Kombinationen linguistischer Daten in einem Text, z.B. unerwartete Übergänge von narrativen Sätzen zu direkter oder indirekter Rede wie in Numeri 5,2 (.„Gebiete den Israeliten, daß sie ... hinausschicken.“) oder 9,1 (.„JHWH redete ... und sagte, daß die Israeliten ... feiern sollen.“).
Zum Auffinden solcher Übergänge sind Daten zur Satzhierarchie erforderlich, z.B. auf die Satztyp-Folge 'Imperativ + WeYiqtol' bzw. 'Infinitiv „sagen“ + WeYiqtol'. Als Ergebnis einer Abfrage zur Satzhierarchie kann die grammatische Struktur eines Textabschnittes auf dem Bildschirm gezeigt werden (vgl. Figur 10).

b) Syntax und Texttytpkombinationen

Das Programm zur Erstellung der Texthierarchie ordnet beim Aufbau der Datenbank allen Sätzen in einem Text Texttypmarkierungen zu:26
- ein Wayyiqtol-Satz erhält die Texttypmarkierung „N“ (für „narrative“);
- eine von יֵשָׁן oder רַע ה eingeleitete direkte Rede wird mit „Q“ (für „quotation“),
- ein Imperativ- oder ein Yiqtol-Satz ohne Redeinleitung mit „D“ (für „discourse“) gekennzeichnet;
- fehlen grammatische Hinweise, wird die Kennzeichnung des nächsthöheren Satzes übernommen.

Bei der Untersuchung des Textes können diese Daten dazu verwendet werden, Textstellen mit unerwarteten Kombinationen ausfindig zu machen, wie z.B. unerwartete diskursive Texte in einem narrativen Text.


Die Kennzeichnung „QN“ verweist auf das ebenso unerwartete Auftreten (narrativer) Wayyiqtol-Sätze innerhalb einer direkten Rede. Beispiele finden sich u.a. in Numeri 5,20,27; 8,17-19; 14,16,22,24; 16,14; 20,15.

26 Die hier verwendete Klassifizierung folgt den textgrammatischen Anreihungen von H. Weinhöch (Tempus, 1994) und W. Schneider (Grammatik,

5. Ergebnis

Eine brauchbare Datenbank zur Analyse von Texten der antiken Literatur sollte mindestens drei Typen sprachlicher Daten verwenden, deren Kategorien zur Darstellung und Speicherung der Textstrukturen dienen, der Dokumentenstruktur einerseits und der distributionalen sowie der funktionalen linguistischen Struktur andererseits.
Im Blick auf die Dokumentenkennung müssen die klassischen Gliederungs-Kategorien (bei bibliischen Texten: Buch, Kapitel, Vers, Halbvers) stets zugänglich sein.


Dieser Weg linguistischer Textcodierung bietet ein offenes System in dem Sinne, daß die so erstellte Textdatenbank für viele Zwecke, von phonologischen und morphologischen Untersuchungen angefangen bis zur Analyse textlicher Makrostrukturen, eingesetzt und daher von Linguisten und Literaturwissenschaftlern, von Übersetzern und Exegeten genutzt werden kann.

Nach der Datenproduktion mit Hilfe von Musterwiedererkennungsprogrammen ist die Weiterentwicklung der Retrievalsoftware notwendig. Denn diese muß einerseits die Suche nach paradigmatischen Daten (die Sammlung von möglichst vielen Parallelstellen zu dem untersuchten Phänomen), sondern auch die Suche nach syntaktischen Textstrukturen (die Untersuchung möglichst vieler Phänomene in einer Textkomposition) ermöglichen. Auch die Programme zur Datenauswertung werden weiter entwickelt, insbesondere mit dem Ziel der Textsegmentierung auf höheren Ebenen (z.B. die linguistische Definition von Textparagraphen) und Verfahren zur Wiedererkennung von Aktanten in einem Text (participant tracking).

**Literatur**

Biblia Hebraica Stuttgartensia [BHS] (Stuttgart 1967-1977)
Biblia Hebraica Quinta [BHQ] (Stuttgart 1999ff.)
FRIEDRICH COSERIU, Textlinguistik. Eine Einführung (Tübingen 1972)

JOHN H. SAILHAMER, A Database approach to the analysis of Hebrew Narrative, Maarav 5-6 (Santa Monica, Calif. 1990) 319-335
WOLFGANG SCHNEIDER, Grammatik des biblischen Hebräisch. Ein Lehrbuch (München 1993)
Anhang

Die Struktur der Textdaten (vgl. die beiden folgenden Seiten)
Computerlinguistik und Computerphilologie
Kann der „einfältige Leser“ dazulernen?
Johann Haller (Saarbrücken)


1. Elektronische Sprachverarbeitung in Saarbrücken