3.4 Korpusanalyse – Phraseologie

3.4.1 Übersicht

Modul	Funktionalität	Beschreibung	
SLANG I,2.4	**Korpusanalyse – Phraseologie**	Formelsuche kann automatisch geschehen, sobald zum interessierenden Einzeltext ein passendes Korpus von Texten zur Verfügung steht. Das Programm nimmt jede Sequenz des Einzeltextes vom Textanfang her, interpretiert dies als Suchkriterium im Korpus. Es wird bei den Ergebnissen geprüft, ob sich Ketten bilden lassen, so dass nicht voreingestellt werden muss, wie lange die Ergebnisse / Übereinstimmungen zu sein haben. Auf diese Weise findet der Computer selbstständig Übereinstimmungen, wo eine Wortkette der Länge$_n$ genau so sich im Korpus nachweisen lässt. Ist die Länge z.B. > 5, die gefundene Belegzahl = 1, so könnten eine gezielte Anspielung oder ein Zitat vorliegen. Ist die gefundene Belegzahl hoch, könnte es sich um gängige Alltagsfloskeln handeln. Die Suche kann im Programm noch verfeinert werden, indem Permutationen und / oder Ähnlichkeiten erlaubt werden. Denn mit Formeln / Phrasen kann auch gespielt werden.	xyzI,2.4

3.4.2 Erläuterung

Vor allem wenn große Datenmengen durchsucht werden sollen, können die Fähigkeiten des Rechners besonders gut genutzt werden. Aber es geht nicht nur um *Quantitäten*: Im Fall von Sprachdaten liegen in der Korpusanalyse auch *qualitative* Erkenntnismöglichkeiten. Sie sind im traditionellen Terminus »Phraseologie« schon angedeutet: Über große Korpora von Sprachdaten lassen sich Muster des Sprachgebrauchs erheben, geprägte Wendungen, Formeln, Abwandlungen von festen Mustern – was sich Kabarettisten gern zunutze machen . . .

Es kann über eine solche Analyse bewusst werden, wie sehr jeder Sprachgebrauch geprägt, standardisiert ist. Selbst die kreativsten Schriftsteller müssen sich in einem hohen Maß danach richten, wenn sie verstanden werden wollen. Es bleibt ihnen dann immer noch ein Bereich, in dem sie unkonventionell den Sprachgebrauch verändern, neue Sichtweisen prägen können. Aber die Standards beginnen schon im untersten Grammatikbereich. Wortfolgen wie »auf der Mauer, auf der Lauer« kann man nicht ungestraft variieren: »der auf Mauer, der auf Lauer«.[85] Die Präposition ist dann zwar immer noch Prä-Position, d. h. dem Nomen vorangestellt. Nur hat sie sich zwischen Artikel und Nomen gezwängt. Das darf nicht sein. Das wird auch ein kreativer Schriftsteller nicht ändern.[86]

Die Spannbreite und die Möglichkeiten einer Korpusanalyse sind also wesentlich größer, als sie mit dem Stichwort »Formelsuche« angedeutet ist. Letztere ist sozusagen nur die leuchtende Spitze des Eisbergs. Darunter kann dieses *tool* jedoch den Blick öffnen für die Fülle von grammatischen Standards (Konventionen), die den Sprachgebrauch beherrschen, die nicht weiter begründbar sind, die aber das Thema »Muster« = mehr oder weniger festgelegte Sprachgebräuche schon sehr früh, bei einfachen und alltäglich erscheinenden Wortverbindungen ansiedeln.

Verlängert man die Wortketten, so nimmt natürlich bei der Suche im Korpus die auffindbare Belegzahl ab, die unspektakulär aber wirksam den Sprachgebrauch bestimmen. Sichtbar werden aber immer mehr längere Ketten von Wortformen, die entweder identisch oder mit Variationen in Gebrauch sind und dann als eigene Einheiten wiedererkannt werden. Ein eingestreutes Sprichwort oder eine Begrüßungsfloskel werden erkannt als Klischee, als vorgegebene, als nicht vom aktuellen Sprecher geformte Wortkette (vgl. RIFFATERRE). Solche Befunde und Belege weist der Rechner nach.[87]

---

[85] »der auf See blasende Wind ist stärker als der an Land« – die Variation geht also doch. Ein Beispiel dafür, dass nicht mit strengem »Entweder – Oder« bzw. einem starren Regelbegriff operiert werden kann, sondern mit mehr oder weniger großen Wahrscheinlichkeiten. Das ist ohnehin für den Bereich der Grammatik natürlicher Sprachen geboten.

[86] Die Argumentation folgte nun den traditionellen, immer auch semantik-basierten Wortarten. / der / bleibt zwar *eine* Wortform. Aber im Bedeutungshintergrund ist sie einmal als *Artikel*, einmal als *Relativpronomen* deutbar. Im ersten Fall darf es keine Umstellung geben, im zweiten doch. – Nur zur Erinnerung: in dieser Form werden wir *nicht* argumentieren. Eine Kategorie wie »Nomen« oder »Artikel« spielt aktuell keine Rolle. Dagegen könnte »Prä-position« von Interesse sein. Im Wortsinn ist er bedeutungsfrei, zeigt nur an, dass etwas grundsätzlich *vor* etwas anderem steht. Im Englischen kommt man dann in Schwierigkeiten, wenn man grundsätzlich / in / als »Präposition« klassifiziert. »The change we believe in« – offenkundig widerspricht hier das / in / der traditionellen, aus anderen Sprachen übernommenen Kategorisierung als Prä-position. Hier ist es Post-position. Man sollte also nicht annehmen, eine Wortform, die in verschienenen Sprachen die gleiche zu sein *scheint*, habe überall auch die gleichen distributionellen Merkmale.

[87] SCHINDELE (1995) hat – theoretisch unterfüttert durch die weiteren Arbeiten – sehr schön für die hebräische Josefsgeschichte (mit der gesamten hebräischen Bibel als Vergleichskorpus) Befunde erhoben und damit rechnergestützt im

Wie die mehrfach im Korpus nachweisbaren längeren Wortketten dann benannt und bewertet werden, das bleibt dem Interpreten überlassen.[88] Es werden die schon genannten Begriffe ins Spiel kommen. Entsprechend den Befunden fallen die Aussagen zur Aussageintention aus. Bedient sich ein Sprecher *vorwiegend* solcher vorgeprägter und langer Wortketten? Sein Motiv könnte der Wunsch sein, sich an vorgegebenen Autoritäten auszurichten, oder seinen Bildungsstand zu dokumentieren. Oder der Autor formuliert weitgehend *ohne* solche Bezugnahmen (allerdings auf der Basis dessen, was grammatisch üblich ist), an einer Stelle jedoch bringt er offenkundig ein Zitat, eine erkennbare Anspielung. Das kann einen dramatisierenden Effekt haben. – Die Gestaltungsmöglichkeiten sind in diesem Feld unüberschaubar, immer wieder überraschend und auf jeden Fall spannend.

Das *tool* kann demnach zu zwei unterschiedlichen Zwecken eingesetzt werden:

– Zur genau spezifizierbaren Suche nach einzelnen Wortketten – mit großer Spannbreite bei den Bedingungen (exakte Übereinstimmung bis hin zu verschiedenen Ähnlichkeitsgraden, erlaubten Permutationen usw.).

– Zum Vergleich: Einzeltext in seinen Relationen zu einem großen, umgebenden Textkorpus. Zu den Suchbedingungen gilt das Gleiche wie soeben. Nur wandert jetzt ein Fenster über den Einzeltext, immer wird nach Entsprechungen im großen Korpus gesucht. Die vielen Befunde werden geordnet. Der Benutzer kann bestimmen, ab welcher Kettenlänge ihn die Befunde interessieren (Zweierbefunde werden unüberschaubar sein, ab Dreierbefunden – nach oben offen – werden die Ergebnisse spannend). Wichtig ist, dass man dem Programm nicht mitteilen muss, nach welcher Kettenlänge es suchen soll. Vielmehr merkt es selbst, ab wann es keine Befunde mehr gibt. Der Benutzer kann sich also überraschen lassen über die gefundenen Übereinstimmungen – und für den Fall der Josefsgeschichte kann ich referieren, dass es viele erfreuliche Überraschungen gab.[89]

Wenn für das Korpus bzw. dessen Einzeltexte Entstehungszeiten bekannt sind, können die gefundenen Treffer dazu verwendet werden eine *relative Chronologie* zu erstellen. Entscheidend ist, in welche Textbereiche hinein die gefundenen Querverbindungen gehen. Wichtig ist auch die Quantität: Werden Wortverbindungen in einer Häufigkeit gefunden, dass man sprachliches Allgemeingut anzunehmen hat, oder handelt es sich um Einzelverweise?[90]

---

eindeutigen Sinn »Konkordanzarbeit« vollzogen. Wogegen Konkordanzarbeit in üblicher Weise meist doch wieder Bedeutungswissen einschließt: äusserlich gleiche Wortformen werden doch auch nach unterscheidbaren Bedeutungen sortiert – ein berechtigter Einwand z. B. gegen MUTHMANN, dessen Werk zunächst nach nichts anderem als einer reinen Wortliste aussieht. Der gleiche Einwand gilt für die im hebraistischen Bereich üblichen großen Konkordanzen.

88 Beachten sollte man, ob die Funde sich vorwiegend bei ein und dem selben Autor bzw. Presseorgan finden. »Herr | Frau [ ], wir danken Ihnen für dieses Gespräch« – diese Wortkette, mit Varianten am Anfang, lässt mit hoher Sicherheit auf ein SPIEGEL-Interview schließen. Finden sich die Befunde breit gestreut, kennzeichnen sie den gängigen Sprachgebrauch einer Epoche, u. U. soziografisch weiter eingrenzbar: »voll cool« kennzeichnet gegenwärtige Jugendsprache, sicher nicht die Behördensprache.

89 Vgl. dazu den Beitrag SCHINDELE (1995).

90 Im Fall der Josefsgeschichte wurden die von SCHINDELE erhobenen und dann (im selben Band von SCHWEIZER) auch chronologisch ausgewerteten Befunde zusätzlich mit weiteren Erkenntnissen abgeglichen, vgl. SCHWEIZER (1998).

Die Literaturangaben zu diesem Abschnitt – in Ergänzung von einigen Angaben schon bei Ziff. 1.2.6 – enthalten auch viele Beiträge zum Thema: *Autoridentifikation.* Das kann in *literar-historischer* Perspektive wichtig sein, wenn etwa bei fraglichen SHAKESPEARE-Texten deren Autorschaft überprüft werden soll. Das ist auch im Bereich der *Forensik* wichtig, wenn es um Schuld oder Unschuld von Angeklagten geht. So hat man beim letzten in England hingerichteten Verurteilten nachträglich, durch Sprachanalysen festgestellt, dass ihm belastende Aussagen von den Vernehmungsbeamten untergeschoben worden waren, er also unschuldig gewesen war.[91]

Die Literaturangaben sind *methodisch* nicht einheitlich. So arbeitet der Beitrag von MIRANDA GARĆIA mit dem Wissen, was »Funktionswörter« in der betreffenden Sprache sind. Damit können wir *noch nicht* operieren. Uns interessieren auf der aktuellen Ebene Überlegungen zur Autorschaftidentifikation insofern, als damit auf Wortkettenbasis gearbeitet wird: können damit persönliche Eigenarten beim Sprachgebrauch eines Menschen identifiziert werden?[92]

---

[91] See the book of HÄNLEIN 17s.

[92] Es geht z. B. um persönliche Marotten. Ein durchaus gewiefter Student hatte die Angewohnheit, häufig in seine Rede einzustreuen: »weiß nich, keine Ahnung«. Diese 4 Worte widersprachen dem, was er jeweils beitrug: er hatte durchaus Ahnung. Die vier Worte in fester Abfolge hatten eine andere Funktion: sie sollten sicherstellen, dass der Sprecher weiterhin reden darf, das Wort behält, die anderen gefälligst weiterhin zuhören. – Aber eine derartige Analyse gehört bereits zur Pragmatik, zur Dialogbeschreibung. Aktuell interessiert das Merkmal: da streut einer ständig diese Wortkette mit ziemlich fester Frequenz in seine Redebeiträge ein. Daran ist er zu erkennen.

3.5 Korpusanalyse im Web: CoMOn-Programm

3.5.1 Übersicht

Modul	Funktionalität	Beschreibung
SLANG I.2.5	**Korpusanalyse im Web:** **[Angebot verschiedener** **Korpora: Altes Testament** hebräisch, griechisch, **Neues Testament**, griechisch, **Koran**, deutsch, arabisch, – **inzwischen erweitert**] Einzeltext im Verhältnis zum gesamten dazugehörigen Korpus	Angebot, per Java-Applet[93] einen Einzeltext komplett nach Entsprechungen im rest- lichen Korpus zu durchsuchen. Hinter der Web-GUI wird eine eingeschränkte Version des Programmes unter 3.4 aktiviert. Ausblick: Permutationen und Ähnlichkeitsmaße lassen sich einstellen.

Literatur: Korpusanalyse am hebräischen AT im Web (als Exempel)

Vgl. die Publikationen von M. SCHINDELE unter Ziff. 3.4

3.5.2 Erläuterung

Korpora, die aktuell integriert sind: Hebräisches Altes Testament (Masoretentext), Griechisches AT (Septuaginta), Griechisches Neues Testament, Kombination: griechisches NT und griechisches AT, Arabischer Koran, Deutscher Koran, Günter Grass, »Die Blechtrommel«, Mark Twain, »Tom Sawyer«, »Huckleberry Finn«, deutsches Zeitungskorpus NEGRA, Marcel Proust, »À la recherche du temps perdu«, Lew N. Tolstoj, »Anna Karenina«.

---

[93] Der Aufruf über einen gängigen Browser verlangt, dass das aktuelle JAVA Plugin installiert ist. Es kann heruntergeladen werden von *http://www.java.com/de/download* (mindestens Version 1.6.10 !; für volle Funktionalität: JAVA 1.7). Popup-Einstellungen müssen nicht beachtet werden. – Die neueste Entwicklung: ein weiter unten nochmals erwähntes *launching protocol* der Fa. *Sun* überprüft beim Programmstart, ob die benötigte Java-Version auf dem Rechner liegt. Wenn nein, kann durch einfaches Anklicken des roten Warnhinweises das Updaten in die Wege geleitet werden.

3.5.2.1 Hebräisches Altes Testament, zugleich: Einführung in die Arbeitsweise des Programms

Konkordanz:

Mit Hilfe großer gedruckter Konkordanzen (Auflistung jeder Wortform, mit ein wenig Kontext, samt Stellenangabe) ist es in der Forschung immer schon möglich und üblich, für grammatische Fragestellungen oder auch Formelsuche auf Befunde in einem großen Textkorpus zurückzugreifen.[94] Das Problem der gedruckten Wortlisten war bislang aber ein Dreifaches:

(1) Konkordanzen nahmen keineswegs die *äussere Form* allein zum Sortierkriterium, obwohl das standardmäßig behauptet wurde. Vielmehr spielte auch *Inhaltswissen* herein, etwa dann, wenn man – inhaltlich – ein und dieselbe äussere Form auf zwei Lemmata verteilen konnte (Thema der Homonyme).[95] Sobald im Kontext von »Konkordanz« das Thema »Lemmatisierung« verhandelt wird, kann man sicher sein, dass nicht nur die *Signifikanten* der Suche zugrunde liegen, sondern immer auch nach den *Signifikaten*, also die Bedeutungen, geschaut wird. –

(2) Bisherige Konkordanzen hatten Einzelwortformen im Blick. Bisweilen wurde die Trefferdarstellung durch Hinzufügung einiger umgebenden Wortformen anschaulicher gemacht. Aber die Kontextwörter waren nicht Bestandteil der Suche gewesen. Suchkriterium in klassischen Konkordanzen war und blieb die Einzelwortform. Eine Öffnung auf *Phraseologie* hin blieb damit den Schlussfolgerungen des Nutzers überlassen. Er musste sich *Idiom, Zitat, Anspielung usw.* durch mehrfache Suche erst zusammenbauen.[96] Heutzutage kann und muss das Stichwort lauten: *computergestützte, string-orientierte Konkordanz.*

(3) Im Zeitalter der *Text*orientierung der Linguistik kann man den Blick von einzelnen Wortformen, allenfalls Phrasemen, weiten auf einen *gesamten Text*, der insgesamt und lückenlos überprüft werden soll auf seine Verbindungen (gleiche strings) zum umgebenden Korpus. Die Fokussierung auf isolierte Einzelfunde wird abgelöst durch Beiziehung *aller* einschlägigen Daten.[97] Die Befunde werden dem Benutzer automatisch und sehr schnell geliefert. Er kann sich dann ganz auf die Auswertung konzentrieren und darauf, die angemessenen Schlüsse zu ziehen.[98]

---

[94] Natürlich ist dieser 'Handbetrieb' mühsam und zeitaufwändig. Das Durchprüfen einer überschaubaren Profeten-Kurzgeschichte dauerte einmal 2 Wochen.

[95] Diese Vorhaltung gilt nicht nur für die großen hebräischen Konkordanzen, sondern auch für das »Rückläufige deutsche Wörterbuch« von G. MUTHMANN.

[96] Aus doppeltem Grund muss dieses Verfahren als veraltet gelten: (a) Seit einigen Jahrzehnten öffnen sich Teile der Sprachwissenschaft zum Text hin, zur Korpuslinguistik. (b) Etwa seit gleicher Zeit bietet der Computer immer mehr an Unterstützung für solche umfangreichen und mühevollen Recherchen.

[97] Durch Definition von Mindestbedingungen wird dafür gesorgt, dass die Suche unter gleichen Parametern abläuft. Im Rahmen heutiger Rechengeschwindigkeit ist das zig-fache und automatische Aufrufen des Suchalgorithmus, bis einmal der gesamte Text durchlaufen ist, keinerlei Problem mehr. Benötigte der Großrechner für diese Aufgabe bei der hebräischen Josefsgeschichte Anfang der 1990er Jahre noch 5 Stunden, erledigt dies ein heutiger PC in weniger als 1 Minute. – Nicht auszudenken der Aufwand, den ein menschlicher Benutzer mit Buchkonkordanz treiben müsste – abgesehen von all den Unsauberkeiten (s.o.), die sich in seine Ergebnisse einschleichen würden.

[98] Nimmt man die kurze erste Sure des arabischen Koran, stellt man fest, dass das letzte Textdrittel im restlichen Koran keine Resonanz hat. Anders gesagt: es ist kreativ formuliert. Der Befund ist merkwürdig, so lange man annimmt, die erste Sure sei doch wohl auch die älteste, denn dann müsste man annehmen, dass ihr Sprachgebrauch vielfältig in den weiteren Suren gespiegelt worden wäre. Plausibler wird der Befund, wenn man annimmt, dass die *Eröffnungssure* bei der Zusammenstellung des Buches *als letzte* der schon fertigen Sammlung vorangestellt wurde.

Die maschinelle Suche kann und braucht sich Beschränkungen wie die beschriebenen nicht erlauben. Die 'Dummheit der Maschine' – bei ungeheurem Fleiß – bietet die Chance, die *Ebene der Ausdrücke als einheitliches Suchfeld* zu verwenden. Ausdrücke vergleichen, Übereinstimmungen festzustellen, zu sortieren, aufzubereiten – das ist es, was ein Computer unschlagbar schnell durchführen kann. Verschonen muss man ihn jedoch von jeder Sorte von Inhaltswissen. Das bedürfte eines qualitativ völlig anderen Zugangs.[99] Aber die vermeintliche Beschränkung hat zwei Vorteile im Schlepptau:

Der erste: Die Fokussierung auf die Ausdrucksseite ist durch die Zeichentheorie bestens gedeckt. Die Technik unterstützt die fundamentale (zeichen-)theoretische Unterscheidung: Sprache beruht auf *arbiträrem* Zueinander von Ausdrucksebene und Bedeutungsebene. Die Technik erzwingt geradezu – im Gefolge der Theorie – eine unzweideutige Praxis: Ausschluss von jeglichem Bedeutungswissen, Konzentration allein auf die Ausdrücke. Es ist also kein Zufall, wenn damit und so erst das technische Hilfsmittel höchst effizient zur Geltung kommen kann.[100]

Der zweite Vorteil besteht darin, dass die Homogenität der Ergebnisse (Ausdrucksebene) den Blick des Betrachters unverstellt auf die Dimension des sprachlichen Aktes lenkt, die physisch realisiert wird, also die Sinne anspricht, somit unmittelbar zugänglich ist.[101] Da Sprachbenutzer ohnehin die Bedeutungsebene *privilegieren*, die Beiträge der gestalteten Ausdrucksseite eher übergehen und übersehen, also im Unbewussten belassen, rücken die verschiedenen Schritte der (Ausdrucks-)SYNTAX, darunter die Korpusanalyse, den Beitrag der *strings* zur Kommunikation ins Bewusstsein.

Textbasis:

Im Fall der hebräischen Bibel ist das Textmaterial aus historischen Gründen zu differenzieren:

Der maschinenlesbare Text des Alten Testaments liegt in Form einer Transkription vor: die von rechts nach links zu schreibenden hebräischen Schriftzeichen wurden in von links nach rechts zu lesende lateinische Buchstaben umgesetzt. Die Konventionen, die bei uns gelten, werden in Form einer *Transkriptions-Tabelle* beigefügt.[102]

---

99 Deswegen kommen alle, die »Lemmatisierung« für nötig halten, in große methodische Schwierigkeiten.

100 Dagegen ist die in der traditionellen Grammatik übliche Mixtur von Ausdrucks- und Bedeutungsebene informatisch gesehen vom Übel: es können dabei nur schwer zu bedienende und ineffiziente elektronische Konkordanzen entstehen. Um sie zu bedienen ist linguistisches Expertenwissen nötig, wogegen die Arbeit an der Ausdrucksebene von jedem vollzogen werden kann.

101 Dagegen verlangt alles, was mit *Bedeutungen* zusammenhängt, spezifisches, erlerntes Wissen, erschließt sich erst durch diese Schranke hindurch.

102 Das Thema »Transkription« des Hebräischen ist umfangreich. Es gibt verschiedene, z. T. für verschiedene Zwecke ausgerichtete Systeme. Vgl. die Auflistung in SCHWEIZER (1995,iii) 16. Um mnemotechnisch Hilfestellungen zu bieten, wurden hebräische Schriftzeichen bisweilen durch ähnlich aussehende wiedergegeben. Das CATAB-Institut, Lyon, benutzte für א das »%«. Mnemotechnisch eine gute Lösung, informatisch für unsere Zwecke unbrauchbar (s.u.). – Bei uns gilt folgende Liste – kursiv und in *[ ]* wird der jeweilige *Unicode*-Wert beigefügt:

א = A	*[05D0]*		
ב = B	*[05D1]*		
ג = G	*[05D2]*		
ד = D	*[05D3]*		
ה = H	*[05D4]*		
ו = W	*[05D5]*		
ז = Z	*[05D6]*		
ח = X	*[05D7]*		
ט = J	*[05D8]*		
י = Y	*[05D9]*		
כ = K	*[05DB]*	am Wortende: ך	*[05DA]*
ל = L	*[05DC]*		
מ = M	*[05DE]*	am Wortende: ם	*[05DD]*
נ = N	*[05E0]*	am Wortende: ן	*[05DF]*

Der hebräische Text besteht aus mehreren, aus historischen Gründen zu unterscheidenden Komponenten:

Die *Ebene der Konsonanten.* Sie ist die eigentlich interessante, da diese Textschicht nicht nur konstitutiv für die Vermittlung der Bedeutungen ist, sondern auch hohes Alter beansprucht (vor der Zeitenwende).[103]

Die *Ebene der Vokale*. Sie klärt natürlich viele Mehrdeutigkeiten, die der reine Konsonantentext zulassen würde. Die Vokalzeichen gehören aber der späteren Bearbeitung durch die sog. *Masoreten* an, sind also bis zu ca. 1000 Jahre jünger als der Konsonantenbestand.[104]

Die *Ebene der Akzente*. Sie dienen dem Vortrag im Gottesdienst. Auch sie sind eine Einfügung der Masoreten und für uns im linguistischen Sinn vollkommen irrelevant.[105]

Die *Ebene der biblischen Zählung*. Sie dient der Orientierung, der schnellen Zitierbarkeit, ist ebenfalls eine sekundäre Zutat zum Text, greift aber nicht in diesen ein und ist zudem hilfreich für die Orientierung im großen Korpus.

Festlegung: Dem Programm wird der reine Konsonantentext zur Verfügung gestellt.[106] An ihm werden die Suchläufe durchgeführt. Die biblische Zählung wird mitgezogen, aber von der eigentlichen Suche ausgeschlossen. – Dagegen würden die Vokale, erst recht die Akzente, die Suchergebnisse häufig verfälschen, da

---

ס = S *[05E1]*
ע = I *[05E2]*
פ = P *[05E4]* am Wortende: ף *[05E3]*
צ = C *[05E6]* am Wortende: ץ *[05E5]*
ק = Q *[05E7]*
ר = R *[05E8]*
שׁ = F *[05E9 05C2]*
שׂ = E *[05E9 05C1]*
ת = T *[05EA]*

Bei der Bemerkung »am Wortende« gilt es die Leserichtung zu beachten: von rechts her! Programmiertechnisch, also bei Leserichtung 'links → rechts', gelten folgende Bedingungen:
Blank+»K« → *05DA*
Blank+»M« → *05DD*
Blank+»N« → *05DF*
Blank+»P« → *05E3*
Blank+»C« → *05E5*

Wer die Liste vergleicht mit der in der gen. Literatur »USA-Transkription« genannten, wird 5 Differenzen feststellen. Die Zeichen) (+ & $ wurden ersetzt, da sie – informatisch gesehen – nicht zur Gruppe der »Buchstaben« gehören. Um die Suchmöglichkeiten flexibel zu gestalten, ist die Verwendung von *Regulären Ausdrücken* notwendig. Darunter sind Konventionen zu verstehen, mit denen Zeichengruppen, Umgebungs- oder Häufigkeitsbedingungen abstrahiert definiert werden können. Dadurch kann man eine Suche flexibel gestalten (nicht nur nach identischen Entsprechungen suchen). Dafür darf es im Suchtext aber keine Vermischung der Buchstabenebene (Suchzeichenfolge) mit Zeichen geben, die der Gruppe der Steuer- oder Sonderzeichen angehören. Die Suchergebnisse würden verfälscht.

[103] Man kann zumindest in Teilen das heute gebräuchliche Standardmanuskript der Masoretenfamilie Ben Ascher, entstanden 1004 n. Chr., hinsichtlich des Konsonantenbestandes überprüfen an Textfunden aus Qumran. Diese sind ein Jahrtausend älter und beweisen im wesentlichen die hohe Konstanz des überlieferten Konsonantenbestandes. Auch weitere textkritische Analysen untermauern, dass der Konsonantenbestand – seltene Schreibfehler ausgenommen – sehr zuverlässig überliefert worden war. So waren am langen Erzähltext der Josefsgeschichte nur etwa 10 Korrekturen notwendig – eine Quote, die oft auch von Manuskripten der Jetztzeit erreicht wird.

[104] Sie sind – angestachelt durch die mittlerweile entstandene »Konkurrenz« des Koran – vom Bestreben beherrscht, einen *verbindlichen* und *eindeutigen* Text zu kreieren. (Hinweis STEFAN SCHREINER). Natürlich lässt ein reiner Konsonantentext dann, wenn man ihn bei der Rezitation mit Vokalen versieht, bisweilen unterschiedliche, ja widersprechende Deutungen zu. Diesen »Wildwuchs« wollte man eingrenzen.

[105] Beließe man sie in unserer Textgrundlage, so würde dadurch jedes Suchergebnis komplett verfälscht – ohne dass mit den variierenden Akzenten eine Variation nach Ausdruck und Bedeutung verbunden wäre. Schon die Modifikation eines einzelnen Akzents (aus der Liste von ca. 50 verschiedenen) würde einen Treffer aussortieren.

[106] Zwei Einschränkungen sind zu machen: (1) Die Kennzeichnungen für *Dageš forte* wurden belassen. Dieser Befund wird in der U SA-Transkription durch Verdoppelung des jeweiligen Konsonanten angezeigt. Derartige Verdoppelungen zu reduzieren, wäre technisch kein Problem. Man würde sich zugleich aber auch eine Reihe von Verfälschungen des Konsonantentextes einfangen – nicht jeder verdoppelt geschriebene Konsonant geht auf *Dageš forte* zurück. Daher wurde auf diesen Schritt verzichtet. – (2) Die Unterscheidung via *diakritische Punkte,* durch die die Buchstaben שׁ = F bzw. שׂ = E unterschieden werden, wurde beibehalten. Sobald die Suche nach *Ähnlichkeiten* aktiviert ist, kann, wer will, diese Unterscheidung aufheben.

es im Fall der Vokale um ein äusserst elaboriertes System geht, erst recht bei den ca. 50 verschiedenen Akzenten. Beides hat mit dem ursprünglichen Textbestand nichts zu tun, kann kein vergleichbares Maß an Authentizität beanspruchen.

Die Bereitstellung von Suchtext (Ausgang der Suche, Suchkriterium) wie Korpus (als Suchraum) muss noch das *Ketib* – *Qere*-Problem lösen. Der zur Verfügung stehende, elektronisch lesbare AT-Text enthält – korrekt – die Informationen der hebräischen Bibel (1317), wo zum Konsonantentext und dessen Vokalisierung am Manuskriptrand eine alternative Vokalisierung vorgeschlagen wird.[107] Uns interessiert dabei zweierlei: Die durch das *Qere* (den Kommentar am Rand) vorgeschlagene Änderung betrifft die Lesung im Gottesdienst, die Ebene der Vokale, also nicht den Bereich, auf den wir uns festgelegt haben: den Konsonantentext. – Informatisch gesehen wiederholt sich die Aufgabe, Nicht-Buchstaben zu eliminieren. Denn im vorliegenden Korpus gilt die Praxis, dass mit »*« das Wort markiert wird, zu dem es eine alternative Vokalisierung gibt. Letztere wird in der elektronischen Fassung unmittelbar anschließend, eingeleitet durch »**«, genannt. – Aus beiden, ganz unterschiedlichen Gründen werden wir das *Qere* eliminieren, denn wir sind am viel älteren Konsonantenbestand allein interessiert; außerdem müssen die als Sonderzeichen geltenden »*« eliminiert werden (sie kollidieren mit den *Regulären Ausdrücken*, haben also die Potenz, den gesamten Suchalgorithmus durcheinander zu bringen).[108]

All die vorbereitenden Schritte am Text der hebräischen Bibel sind *Entscheidungen*, d.h. es werden immer auch Möglichkeiten den Text zu konzipieren ausgeschlossen. Daher ist es nur konsequent hier zu betonen: Der Zuschnitt des hebräischen Textes, wie er hier im Programm bereit gestellt wird, kann und will nicht die Wiedergaben in wissenschaftlichen Ausgaben ersetzen. Auch ist es gut, wenn erzielte Treffer immer wieder überprüft werden in wissenschaftlichen Editionen.

Suchprozedur:

Nach den Bemerkungen zur Arbeitsgrundlage »hebräischer Text« kommen wir zum Thema »Schnelle Suche in Texten«:[109] User werden zunächst drei Entscheidungen an der Benutzeroberfläche (GUI) treffen:

(1) Auswahl des Korpus, an dem gearbeitet werden soll. Sobald das getätigt ist, erscheint eine differenziertere Eingabemaske. Mit ihr kann man

(2) klären, wie CoMOn genutzt werden soll. Es gibt 3 Möglichkeiten den Suchtext zu bestimmen und dem Programm mitzuteilen:
   (a) Auswahl durch Anfangs- und Endposition und zusätzlich die Möglichkeit, den Suchtext im rechten Fenster (oben) zu bearbeiten (= **editable corpus text**); durch Drücken der linken Maustaste und Cursor-Bewegung kann man Textstellen *markieren*; durch Betätigen der Löschtaste – »←« – wird der markierte Text entfernt.
   (b) Auswahl *nur* durch Anfangs- und Endposition ohne Editionsmöglichkeit (= **corpus text only**);

---

[107] Z. B. durch Einsetzen einer für einen bestimmten Vokal gebräuchlicheren *mater lectionis*.

[108] Eine weitere, das Korpus vorbereitende Aktion bezieht sich auf den Namen »Issachar«. Er enthielt in der Codierung ein »#«, das auch nicht zum Konsonantenbestand gehört. Es wurde entfernt.

[109] Das Konkordanzprogramm hat mehrere Komponenten und AutorInnen: (1) Ausgangspunkt und im Hintergrund die Arbeit verrichtend ist das Suchprogramm *KoMA* – »Korpus Matching Analysis« – von MICHAEL PACH. Der Suchtext wird darin in »Reguläre Ausdrücke« umgesetzt, was den Wechsel zulässt von der Suche nach *identischen* Entsprechungen und *ähnlichen*. (2) Diese Suchgrundfunktionen wurden von SANDRA LAUCKNER erweitert in eine *Volltextsuche*. Nun arbeitet der Algorithmus sukzessive die Wortketten eines Einzeltextes ab und sucht nach Treffern – je nach Suchbedingungen – im Korpus. (3) SERHIJ BYKH schuf ein *Java-Applet*, mit dessen Hilfe man an einem Internetbrowser seinen Suchauftrag definieren kann, und einen *Mediator*, über den der Suchwunsch dem Programm im Hintergrund mitgeteilt wird. Beides zusammen läuft unter *CoMOn* = Corpus Matching Online. Via Internet sind nur eingeschränkte Möglichkeiten des KoMA-Programms zugänglich.

(c) Verzicht auf die Auswahl durch Anfangs- und Endposition, stattdessen schreibt man *direkt* in das rechte Fenster, wonach gesucht werden soll (= **direct input**).[110]

(3) Im Fall von (1) oder (2): definieren, *was* User überprüft haben wollen. Sie müssen definieren, welches ihr *Suchtext* ist, zu dem weitere Belegstellen, *strings*, gefunden werden sollen. Verschiedene Modalitäten sind durchführbar:

1. Man wählt aus dem Korpus-Text via Stellenangaben – von xx – bis yy – eine Passage aus, die dann zum Suchtext bestimmt wird.[111] Maximale Länge: ca. 70 Verse bzw. 6000 Zeichen.[112] Oben rechts in der grafischen Oberfläche läuft ein Zähler mit, der, sobald die empfohlene Textlänge überschritten ist, einen Warnhinweis ausgibt: mehr wird vom Programm nicht akzeptiert.[113] Der Suchvorgang kann sehr komplex werden, vor allem wenn später noch »Ähnlichkeiten« und / oder »Permutation« verlangt werden. Mit der Längenbegrenzung wird die Wahrscheinlichkeit von Programmabstürzen (wegen Speicherproblemen) minimiert. – **Praxishinweis:** Es ist aus Laufzeitgründen günstiger, einen langen Suchtext durch *mehrere* Suchabfragen zu behandeln. Das Ausschöpfen der maximal möglichen Suchtextlänge verlängert die Suchdauer unverhältnismäßig. Etwa beim griechischen Alten Testament macht sich bemerkbar, dass es deutlich mehr Texte enthält als das hebräische, d.h. der Speicherbedarf ist entsprechend größer. Manche Suchtexte liefern eine besonders große Fülle an Ergebnissen = Querverbindungen (die gleichen umfangreichen Befunde oft mehrmals).[114] – Der gewählte Text erscheint als Suchtext im rechten Programmfenster. Wenn zuvor *text- / literarkritisch* gearbeitet worden war, kann der Suchtext diesen Erkenntnissen angepasst werden – was bei Literarkritik in der Regel heißen wird: gekürzt werden. Auch neue *textkritische* Erkenntnisse können von Hand integriert werden.[115] – Damit ist definiert, welche Wortformenkette vom Programm untersucht werden soll.[116] Der Suchtext wird insgesamt als *eine* Wortkette verstanden und ausgewertet. Dazu gehörende Kapitel-Vers-Angaben sind im Suchtext-Fenster nicht sichtbar. Sie werden bei der Suche ignoriert. Damit besteht die Chance auch Wortketten zu finden, die im Suchtext durch eine biblische Zählung unterbrochen werden.

   Manuelle Änderungen des Suchtextes – aus text- oder literarkritischen Gründen – bewirken, dass die **Stellenangabe erlischt**. Der Benutzer hat durch die Änderung einen Suchtext geschaffen, der via Stellenangabe nicht mehr exakt gleich in der Bibel anzutreffen ist.

---

[110] In diesem Fall springt unten die Voreinstellung für »minimal length« auf den Wert »1«. Damit hat man die Möglichkeit, mit dem Programm wie bei der klassischen Konkordanz lediglich nach Einzelwörtern zu suchen. – Da Anfangs- und Endbestimmung in diesem Fall funktionslos sind, werden sie ausgegraut und deaktiviert.

[111] Per Programm werden die pro Buch variierenden Kapitel und Verse angeboten; ist eine Anfangsposition bestimmt, kann hinter sie bei der Endposition nicht zurückgegangen werden. Durch diese Restriktionen werden versehentliche Nonsense-Eingaben unterbunden.

[112] Das sind zwei unterschiedliche Längenmaße, die näherungsweise etwa die gleiche Textmenge angeben. Es handelt sich um eine *kursorische* Einschätzung für den *biblischen* Rahmen. Wir übertragen später das Zählsystem auch auf neuzeitlich-moderne Texte, z.B. Romane. Dabei gilt in der Regel, dass die Größe »Vers« = Paragraf viel umfangreicher ist. Die Maximalzahl von erlaubten 6000 Zeichen wird dann also mit deutlich weniger Versen bereits erreicht.

[113] Entscheidendes Kriterium sind die 6000 Zeichen, gerade auch dann, wenn sie schon *vor* der Verszahl »70« erreicht werden, was vorkommen kann.

[114] Es muss in einem Text nur mehrfach κυριος ο θεος vorkommen, so gilt es jeweils 550 Belegstellen zu finden und aufzulisten.). Das verlangt auch eine leistungsfähige *hardware*. – Dem bei *dual core*-Rechnern ohnehin erhöhten Speicherbedarf versuchten wir durch erweiterte Speicherzuweisung gerecht zu werden. – Bei Laufzeitproblemen oder gar Programmabstürzen sollte also der Suchauftrag gesplittet werden. Erzielbarer Zeitgewinn und Präzision der Funde übertreffen auf jeden Fall alles 'um Lichtjahre', was durch Konkordanzarbeit 'im Handbetrieb' erreichbar wäre.

[115] Durch diese Möglichkeiten wird der Suchauftrag sehr flexibel, Man kann die Flexibilität allerdings auch für einen fehlerhaften Suchauftrag nutzen. Daher gilt es folgende Hinweise zu beachten: Mehr als *ein* **Leerzeichen** in Folge wird vom Programm ignoriert. Ansonsten ist klar, dass bei Hinzufügungen man sich an die hier geltende **Transkription** hält. Alle **Zeilenumbrüche, Tabulatoren, Formatier-, Sonderzeichen, Satzzeichen** bringen den Suchalgorithmus durcheinander. Also auf derartiges verzichten! Erlaubt sind nur **Buchstaben und Ziffern**. Wer nur kürzen muss, tut sich leichter.

[116] Die Begrenzung rührt vom Rechenaufwand her. Der ist bei manchen Einstellungen zu »Ähnlichkeit« oder »Permutation« beträchtlich. Es geht auch darum, die Fülle von Treffern anschließend noch sinnvoll zu verarbeiten. – Auch rechtliche Überlegungen spielen mit. Das Programm soll nicht missbraucht werden, damit der Text als ganzer kopiert wird – aber Vokale und Akzente fehlen ohnehin.

2. Der *Suchraum* (das Korpus, aktuell besprechen wir das Beispiel: hebräische Bibel) wird schon beim Aufruf des Suchprogramms vom Benutzer gewählt.

3. Die Funktion des Programms: die Wortketten des Suchtextes werden sukzessive als *Suchkriterium* genommen.[117] Damit wird der Suchraum = das gewählte Korpus durchlaufen. Übereinstimmungen (Identitäten oder Ähnlichkeiten – entsprechend Voreinstellung) werden aufgelistet.[118]

4. Voreingestellt ist, dass das Programm nur Fundstellen ausgibt, die *identische* Wort-/Ausdrucksfolgen aufweisen. Man kann aber auch mit *Ähnlichkeiten* (*tolerance*) spielen, *Umstellungen* (*permutation*) erlauben. Beides ist eigens festzulegen. [In der Startphase sind *permutation* und *tolerance* noch deaktiviert. – Es wird einen eigenen Hinweis auf der GUI geben, sobald auch diese Funktionen zur Verfügung stehen.]

5. Eine *Minimallänge* der Wortketten ist anzugeben. Gibt man »2« an, kann man sicher sein, in der Fülle der Befunde zu ertrinken. Aber für manche Fragestellung mag die Einstellung sinnvoll sein. Einstellung »1« gibt sämtliche Einzelworte, die anderswo belegt sind, aus. Aus der Flut von Querverweisen werden *Hapaxlegomena* hervorstechen, die also nur im aktuellen Suchtext vorkommen und im restlichen Korpus nicht.[119] Als Minimalbefund bietet sich »3« an. Das Programm erkennt selbst, bis zu welchem Maß sich Übereinstimmungen finden lassen. Im einen Fall wird es Wortketten der Länge 4 ausgeben, im anderen solche der Länge 8. Man muss also keine gewünschte Obergrenze von Hand eingeben.

6. Zunächst wird summarisch die Anzahl der Treffer in der Liste genannt. Durch »List all« kann man die einzelnen Fundstellen aufgelistet bekommen.

7. Durch Anklicken einer einzelnen Stelle = Zeile in der Tabelle wird – farbig unterlegt – zweierlei angezeigt: (1) Im Suchtextfenster ist die Textpassage zu sehen, nach der in diesem Fall gesucht worden war und zu der es einen Treffer gibt. Sieht man im Suchtextfenster keine farbige Markierung, ist entweder nach oben oder nach unten zu Scrollen, da der Suchtext nicht zur Gänze im Fenster angezeigt werden kann. – (2) Es geht ein neues Fenster auf, in dem der Treffer aus dem Korpus angezeigt wird: mit biblischer Zählung, sowie einigem vom literarischen Kontext. – Wer mit den Transkriptionskonventionen vertraut ist, wird so schon eine Vorstellung entwickeln, was der gefundene Text aussagt. Andernfalls kann man via Stellenangabe in seiner eigenen Bibel nachschlagen. – Als Verfeinerung und zur Weiterverarbeitung der Ergebnisse kann durch eigenen Button der Suchtext in hebräischer Klarschrift angezeigt werden, wobei die Stellenangaben der einschlägigen Treffer an passender Stelle in einer parallelen Spalte aufgeführt werden.

8. **Neu ab Juni 2012:** Durch die Studienarbeit von JOHANNES VENZKE sind einige Verbesserungen im *Suchtextfenster* nachgetragen:
   (a) Hatte man mit Kapitel | Vers einen Suchtext definiert, so werden nun die Verse auch angezeigt. Bislang fehlte die Nummerierung, die einzelnen Verse waren lediglich durch Leerzeile getrennt. Das ist nun übersichtlicher.
   (b) Möchte man aus dem ausgewählten Suchtext einige Verse löschen, löscht man also auch die Leerzeile vor dem zu löschenden Text, so verschwindet die Versangabe ebenfalls. Die restlichen bleiben aber erhalten.
   (c) Was gelöscht ist, ist weg. Möchte man doch wieder dem ursprünglichen Suchtext näherkommen, muss er nochmals links definiert werden.

---

[117] Links unten auf der GUI ist per Häkchen die Voreinstellung *a = A* bei Korpora moderner Sprachen sichtbar. Das bedeutet, dass Klein- und Großbuchstaben *nicht* unterschieden werden. Ein *aber* im Satz und ein *Aber* am Satzanfang werden als identisch erkannt. Entfernt man das Häkchen, werden die Unterschiede beachtet. – *Satzzeichen* werden in der Textdarstellung beibehalten, bei der Suche jedoch ignoriert. – In seltenen Fällen kann das *Apostrof* Probleme bereiten, u.z. dann, wenn es als Typ von Anführungszeichen verwendet wird.

[118] Schon das Programm auf TUSTEP-Basis (entworfen von MARTIN SCHINDELE), mit dem wir Anfang der 1990er Jahre die Befunde zur Josefsgeschichte erarbeiten ließen, konnte Einbettungen erkennen: Zur gegebenen Kette A – B – C – D seien Treffer gefunden worden. Es werden dann aber auch Treffer zu A – B – C und B – C – D gefunden (bei Mindestlänge 3). Diese Fähigkeit hat auch das CoMOn-tool.

[119] Man kann von Einzeltext zu Einzeltext sehr schnell den Anteil an *Hapaxlegomena* erheben und die Werte vergleichen. Die Folgefragestellung wäre: wo treten Zweier-Verbindungen nur in bestimmten Textbereichen auf, so dass sie auf einen spezifischen Sprachstil etwa eines Autors oder einer Gruppe (theologische Schule) verweisen. – Für bestimmte Fragestellungen grammatischer oder phraseologischer Art braucht man Befunde bereits ab Minimallänge »2«.

(d) Jede Änderung – löschen, addieren – im Suchtext verlangt, dass intern der Index für den gesamten restlichen Suchtext (ab Änderungsstelle) angepasst werden muss. Dies nur als Information. Für den Benutzer hat dies keine Auswirkungen, die er beachten sollte.
(e) **Transkription**: Bei nicht-lateinischen Schriften arbeiten wir mit Transkriptionen. Da es ganz unterschiedliche Konventionen gibt, kann es dem Benutzer bisweilen unklar sein, welcher lateinische Buchstabe für welchen Buchstaben in der Originalschrift steht. Zur Information wird daher für die jeweils aktuelle Schrift *oberhalb des Suchtextes die passende Transkriptionskonvention eingeblendet.*

Das Korpus, in dem gesucht wird, ist fest eingebaut. Es handelt sich um eine aus den USA stammende Textversion.[120] Der Suchtext wird bei der Volltextsuche vollständig auf Entsprechungen im Korpus überprüft. Die *strings* – die Minimallänge kann eingestellt werden – können nach oben offen sein: die Länge der gefundenen Entsprechungen muss nicht vorher festgelegt sein. Voreinstellung ist, dass nach exakten Entsprechungen gesucht wird. [Später wird man auch Ähnlichkeitsgrade einstellen können, oder auch Permutationen erlauben].[121]

Suchergebnisse:

Über die Tabelle der Treffer (rechts unten)[122] können die einzelnen Funde *in ihrem jeweiligen Kontext* im Rahmen des Suchtextes (=Ausgangstextes): im Suchfenster wird die aktuelle Wortkette farblich markiert. Dadurch kann man sich leicht einen Eindruck verschaffen, in welchem Maß der Suchtext durch *strings* geprägt ist, die auch anderswo in der hebräischen Bibel (und wie häufig) vorkommen. Bzw. umgekehrt: Inwiefern der aktuelle Einzeltext kreativ gestaltet, also weitgehend frei von Formeln und Floskeln ist.

Die *Trefferauswertung* kann / muss mehrstufig erfolgen. Hierzu folgende Hinweise:

1. Der Suchalgorithmus durchläuft das gewählte Korpus. Wird an einer Stelle eine Übereinstimmung mit einer Stelle des Suchtextes gefunden, wird der Treffer notiert. *corpus position* bzw. *corpus range* als Stichwörter in der Treffer-Tabelle besagen also: Fundort innerhalb des Korpus – nach Buch | Kapitel | Vers-Zählung oder nach Buchstabenzahl (letzteres ab Korpusbeginn gerechnet). Diese beiden Spalten bieten in *informatisch* relevanter Form die Ergebnisse.

---

[120] Es handelt sich um THE MICHIGAN OLD TESTAMENT, das anfänglich unter der Leitung von H. VAN DYKE PARUNAK (später an der Universität MICHIGAN) und RICHARD E. WHITAKER (Claremont Grad. Schools) angelegt wurde, und hier in einer Kopie des Center for Computer Analysis of Texts (CCAT) der University of Pennsylvania unter dem Namen MICHIGAN-CLAREMONT BHS vorliegt. Allerdings – wie erläutert – haben wir das File bearbeitet, die Codierung geändert.
ALAN GROVES zufolge ist der hier benutzte maschinenlesbare Text überprüft worden durch Vergleich mit einer unabhängig gespeicherten Version der Abtei Maredsous (Belgien), ebenso mit einer Version der Bar Ilan-Universität (Israel). Eine Korrektur fand auch auf der Basis von BHS 1983 statt (*Biblia Hebraica Stuttgartensia*). – ALAN GROVES, dem viel zu früh Verstorbenen, der sich um den elektronisch aufgezeichneten hebräischen Text verdient gemacht hat, an dieser Stelle ein dankbares Gedenken.

[121] Der Suchaufwand wird dadurch gewaltig steigen, so dass entsprechend der Suchtext kleiner zu dimensionieren ist.

[122] Die Tabelle enthält die Spalten
source = das Korpus, das der Suche zugrunde lag.
corpus position (= erste Wortform des gefundenen Treffers im Korpus: nach Kapitel und Vers),
length (= ausgehend von der ersten Wortform besteht die gesamte Suchkette, zu der Entsprechungen gefunden wurden, aus *x* Wortformen),
corpus range (= von-bis-Angabe: der Treffer im Korpus wird durch absolute Buchstabenzählung im Korpus identifiziert),
searchtext range (= wie *corpus range*, nur dass der *string* im Suchtext identifiziert wird, zu dem es im Korpus Treffer gab: Angabe von Buchstabenposition$_x$ bis Buchstabenposition$_y$ ab Suchtextbeginn gerechnet). – Durch *Anklicken* eines dieser Stichwörter tut man den Willen kund, dass man die betreffende Spalte sortieren möchte. Rechts neben dem Stichwort (Pfeil oben | unten) kann man bestimmen, ob man aufwärts oder abwärts sortieren möchte, was faktisch auf zwei Möglichkeiten führt: die Treffer vom Korpusanfang oder -ende her (*corpus position* und *corpus range* laufen parallel) oder ab Mindestlänge bzw. von größter Länge her geboten.

2. *Benutzer* benötigen die Ergebnisse in anderer Form: durch Nennung via *Stellenangabe* – sie wird in eigener Spalte geboten und möglichst mit *Kontextangabe*. Da Bibelverse immer etwas länger sind als die meisten Suchergebnisse, ist diese Ergebnispräsentation ungenauer, genügt aber für die praktische Analysearbeit. Durch Anklicken der betreffenden **Zeile + Button »show context«** bekommt man den Kontext des jeweiligen Treffers im Korpus im Klartext angezeigt.

3. Beurteilt aus der Sicht einer Stelle des Suchtextes heißt das: die verschiedenen Treffer, die alle mit der einen Passage (= Wortkette) des Suchtextes übereinstimmen, finden sich in der Tabelle nicht kompakt beieinander stehend, sondern verstreut. Umgekehrt kann *ein* Korpustreffer sich auf *zwei* (oder mehr) *strings* im Suchtext beziehen. Jede Verbindung eines *strings* im Suchtext zu einem Treffer im Korpus wird in einer eigenen Zeile markiert.[123]

4. Folglich wurde ein zusätzlicher **Ergebnisbutton** eingefügt, der die Treffer *mit dem Leitfaden: Suchtext* sortiert. Aufsteigend, vom Beginn es Suchtextes an, werden die gefundenen Treffer via Stellenangabe an passender Stelle aufgelistet.[124]

   Das Ergebnisfenster enthält alle Metadaten der Suche (Bedingungen, Zuschnitt des Suchtextes), dann den kompletten Text mit den jeweiligen Treffern. Durch Aktivieren/Deaktivieren des Kästchens links unten kann man hin- und herschalten: der Suchtext wird in Transkription (Schreibrichtung: links → rechts) bzw. in hebräischer Klarschrift (Schreibrichtung: links ← rechts) angezeigt.[125]

5. Wer zu den Treffern den jeweiligen Korpuskontext betrachten will, soll weiterhin über die Tabelle per *show* sich in einem eigenen Fenster die Textpassage anzeigen lassen. In die dort geltende Transkription hat man sich schnell eingelesen. Außerdem gibt die Stellenangabe die nötige Orientierung.

6. Das Ergebnisfenster mit allen sortierten Ergebnissen, aufgereiht am sequentiell wiedergegebenen Suchtext kann man via *Save* als HTML-Datei im eigenen Homeverzeichnis speichern. Diese Datei kann mittels Browser geöffnet, angeschaut und in anderer Codierung exportiert werden.[126]

7. Die Trefferanzeige in ihrer ersten Version (Tabelle) stellt Rohdaten dar. Da der Suchtext in unserer Versuchsanordnung immer auch Bestandteil des umgebenden Korpus ist, findet das Programm sich auch selbst (erkennbar an entsprechend hoher *length*-Angabe). Dies ist dann zugleich der Nachweis, dass das Programm korrekt arbeitet. – Darin kann ein wesentliches Merkmal des Algorithmus abgelesen werden: Vor dem Starten muss der Benutzer nur die »Mindestlänge« der gesuchten *strings* bestimmen (oder die Voreinstellung »3« übernehmen). Nach oben ist die Länge der auffindbaren *strings* nur begrenzt durch die Länge des Suchtextes. Zwischen Mindestlänge und Suchtextlänge wird der Algorithmus alle Entsprechungen im Korpus finden. Heuristisch liegt darin ein entscheidender Vorteil: Der Benutzer muss nicht vorab schon wissen, wonach er sucht (indem er die Suchbedingung eng definiert), sondern kann sich überraschen lassen von den gefundenen Entsprechungen.[127]

---

[123] Wenn etwa ein Beleg der Botenformel irgendwo im Korpus im aktuellen Suchtext ein *zweimaliges* Vorkommen als Auslöser hat, so wird die Korpusstelle in *zwei* Tabellenzeilen genannt.

[124] Erst damit wird die Perspektive des Textwissenschaftlers wieder gewonnen: Er will – (a) – *im Textverlauf* sehen, an welchen Stellen Weiterverweise auf andere Texte vorkommen, und – (b) – sollten alle *gleichen Verweise* (mit gleichem *string* als Suchkriterium) zusammengefasst werden, so dass die u.U. vielen Zeilen in der Tabelle kompakt aufgelistet werden.

[125] Suchtechnisch ist es so, dass das Programm seinen Suchtext (vorausgesetzt, er war nicht verändert worden) im Korpus wiederfindet. Aber *diesen Komplett-Treffer haben wir aus der Ergebnisdarstellung entfernt.* – Als Konzession sollten Benutzer hinnehmen, dass – wie schon erwähnt – *Dageš forte* als Doppelkonsonanz verarbeitet wurde. Das zeigt sich nun auch jetzt in der Ergebnisausgabe. Es wäre zwar ein Leichtes, verdoppelte Konsonanten in einen einfachen Konsonanten + *Dageš forte*-Punkt zurückzuverwandeln. Aber damit würden auch viele Fälle erfasst, die so gerade nicht behandelt werden dürfen. Daher bleibe die Herstellung der korrekt druckbaren Version in diesem letzten Punkt dem Benutzer überlassen.

[126] Der informatische Hintergrund ist, dass *Java-Applets* so angelegt sind, dass sie keine lokalen Speicherungen durchführen. Es könnte sonst sein, dass unversehens mit allen möglichen Hilfsdateien die eigene Festplatte vollgeschrieben wird. – Allerdings stellt die Firma *Sun* seit neuestem ein *launching protocol* zur Verfügung: es ist in punkto Sicherheit entsprechend zertifiziert und fragt bei Speicherwunsch zurück. Wird der Speicherwunsch bestätigt, ist eine direkte lokale Speicherung möglich – unter Beibehaltung der Formatierung von CoMOn. – Letzter Hinweis: Durch Betätigen des buttons »use native characters« kann man zwischen Klarschrift und Transkription hin- und herschalten. In der gewählten Form wird der Text dann gespeichert.

[127] Darin liegt denn auch der qualitative Unterschied dieser maschinell unterstützten Suche zu herkömmlicher Konkordanzarbeit.

8. Beim Lesen der Ergebnisdatei (erzeugt via »Ergebnisbutton« = »Generate Conclusion«) sollte man folglich achten auf *Einbettungen*. Sie sind separat dargestellt. Der Algorithmus kann z.B. Belegstellen für einen langen *string* finden, von dem sich Teile noch irgendwo anders finden. Die Stellenangaben der kürzeren Teile muss man also zu den längeren addieren. Erst so erhält man die Gesamtzahl (der kürzeren *strings*).

aaa	*bbb*	*ccc*	*ddd*	*eee*	*fff*	*ggg*	*hhh*	*iii*	belegt in [Stellen]
		ccc	*ddd*	*eee*	*fff*				belegt in [Stellen]
				eee	*fff*	*ggg*	*hhh*	*iii*	belegt in [Stellen]
						ggg	*hhh*	*iii*	belegt in [Stellen]

   Es ist der Algorithmus, der sich dynamisch, vom Suchtext ausgehend, anpasst.

9. Hatte der Benutzer von der Möglichkeit Gebrauch gemacht, den Suchtext (zunächst bestimmt durch Anfangs- und Endposition mittels BUCH/KAPITEL/VERS) zu kürzen, z.B. weil er *literarkritisch* überzeugt ist, dass ein Teil in der Mitte sekundär sei, So heißt das für das Programm zweierlei: (a) Der Suchtext hat einen neuen Zuschnitt bekommen, entspricht also nicht mehr fraglos den Bestimmungen via traditioneller Zählung. Daher wird diese deaktiviert. (b) Der Suchtext besteht – im angenommenen Beispiel – aus zwei Teilen. Der Algorithmus wird diese Teile im großen Korpus wiederfinden, d.h. die Suchtextteile werden auch auf der Trefferseite aufgeführt werden.[128]

   Ist somit nach einer Suchtextveränderung ein Suchtextteil in voller Länge auch auf Trefferseite erwähnt worden,[129] so folgen auf diese längst mögliche Kette – wie soeben unter Ziff. 7 erläutert – Entsprechungen für kürzere Ketten, die Teil der langen sind, und für die es separate Treffer gibt.

10. **Neu ab Juli 2011:** In der Ergebnisdarstellung mit den Stellenangaben für die einzelnen Korpus-Treffer sieht man durch blaue Farbe, dass die Stellenangaben *verlinkt* sind. Durch Anklicken wird somit der Korpustreffer samt einigem Kontext angezeigt. Dies ist eine sehr gute Möglichkeit, hinter den bloßen Stellenangaben die Texte, zu denen eine Beziehung besteht, besser in den Blick zu bekommen.

11. **Neu ab Februar 2013:** Die u.U. sehr zahlreichen Ergebnisse können nun via **Heatmap** (Programmierer OMAR EL GHARBI) durchforstet werden: Die Grafik zeigt links die Kapitel des Untersuchungskorpus. Nach rechts folgen Säulen beginnend mit der geringsten gewählten Wortkettenlänge, aufsteigend bis zur längsten belegten. In diesen Säulen wird pro Kapitel mit Farbe angezeigt, wie intensiv das Kapitel bei der Suche Treffer aufwies. Scharfes Hellrot ist demnach besonders auffällig. Die selben Befunde kann man sich auch **sortiert** ausgeben lassen – 'sortiert' im üblichen Verständnis: die gleichen Kapitel stehen beisammen. Das heißt nicht zugleich, dass die Kapitel in der Reihenfolge dargestellt werden, wie sie im Such-Korpus stehen (v.a. bei der Bibel ist die Ordnung deutlich anders – aber wenigstens hat man die Kapitel des selben Buches in aufsteigender Reihenfolge vor sich).

Eine Anschlussfragestellung auf der Basis derartiger Daten kann darin bestehen, dass aufgrund der Distribution eine *relative Chronologie* der Texte entworfen wird. Das könnte letztlich zu einer – unter Entstehungsgesichtspunkten betrachtet – völligen Neuordnung der hebräischen Bibel führen.[130]

Eine gegenläufige Beobachtung liegt dann vor, wenn es zu einem längeren *string* im Suchtext nur *eine* bzw. sehr wenige weitere Entsprechungen im Korpus gibt. Es könnte sich dann um direkte Übernahmen / Zitate bzw. Anspielungen handeln.[131]

[128] Das ist der Unterschied zum Fall, dass der Suchtext nicht verändert worden war. Dann wird nämlich der Selbsttreffer unterdrückt.

[129] Mit gleicher Bibelstelle, die zunächst auch für den Suchtext gegolten hatte, und wenn die Suchtextveränderungen nur in Kürzungen bestanden hatten. Weiter gehende Eingriffe in den Buchstabenbestand sind zwar möglich, blockieren aber 'Selbsttreffer', von denen aktuell die Rede ist.

[130] Es war genau dieser Typ von Suche, der uns dazu führte, für die Josefsgeschichte (Gen 37–50 – ohne literakritische Ergänzungen) eine Entstehungszeit um 400 v. Chr. anzunehmen. Herrschende Meinung machte den Text um 3 – 5 Jahrhunderte älter, THOMAS MANN gar 1000 Jahre. Vgl. H. SCHWEIZER (1995).

[131] Damit ist die Richtung der Übernahme noch nicht festgelegt. Sie muss durch weitergehende Indizien erarbeitet werden.

Bezüge zwischen zwei Texten sollten über den übereinstimmenden *string* hinaus immer auch als Einladung verstanden werden, neben der Identität der Ausdrücke auch die *inhaltliche* Struktur der beiden Kontexte zu berücksichtigen. Dann verlässt man zwar die Ausdrucksebene. Letztere hatte aber einen 'objektiven', also starken Hinweis auf eine bestehende intertextuelle Verbindung geliefert. Oft stellt man bei Anspielungen fest, dass der jeweilige Kontext inhaltlich viele ähnliche Merkmale aufweist – was den Eindruck weiter unterstreicht, hier liege eine gewollte Verbindung vor: indem man den Suchtext liest, soll der Leser sich an den zweiten Text erinnern.[132]

Ein **ausführliches Beispiel** wie mit dem CoMOn-Programm gearbeitet werden kann, ist im Web zugänglich unter:
http://www-ct.informatik.uni-tuebingen.de/daten/jguebers.pdf
Vgl. dazu im ANHANG den Abschnitt: **Phraseologie der Bearbeitungen**. Darin wird untersucht, welche Verbindung zur originalen Josefsgeschichte, zum restlichen AT die *sekundären Bearbeitungen* des Textes aufweisen bzw. welche Wortketten analogielos sind.

3.5.2.2 Griechisches Altes Testament

Mit dem selben JAVA-Applet kann auch die sog. »Septuaginta (LXX)« als Suchraum zur Verfügung gestellt werden, die griechische Version des Alten Testaments.[133] Das bedeutet wie im Fall des hebräischen Korpus: Das *Tool* kann und will eine gedruckte Ausgabe, samt ihren wissenschaftlichen Anmerkungen nicht ersetzen. Es interessiert bei uns der Text in einer Form, die geeignet ist für den Suchalgorithmus.[134]

---

[132] *Ähnlich* kann heißen: Gleichheit in der inhaltlichen Anlage, kann dabei aber auch expliziten Kontrast bedeuten. – In einer von beiden Bedeutungen von *ähnlich* zu reden, setzt voraus, dass man von der Wortbedeutung absieht – in dieser Hinsicht dominiert ja die Unähnlichkeit –, stattdessen nimmt man die Bedeutungsstruktur auf pragmatischer Ebene in den Blick, arbeitet also mit einer Abstraktion. – Wieder die Josefsgeschichte betrachtet: dort gab es auffallend viele derartige gezielte und sinnvolle Anspielungen. Der Text ist in eine Wolke von mitklingenden weiteren Texten, die man kennen sollte, eingebettet. Es ist heutzutage der Computer, der das Wissensdefizit ausgleichen kann – die profunde Kenntnis der hebräischen Bibel, die auf technische Hilfe verzichten könnte, ist in Mitteleuropa sehr selten anzutreffen. Kleines Beispiel für Ähnlichkeit in Form von Kontrast: die mehrere hundertmal im AT vorkommende Botenformel wird in Gen 45 einmal aufgerufen, aber kontrastierend verändert: Josef ist Subjekt, nicht Jahwe. Das fällt auf angesichts des restlichen quantitativen Befundes.

[133] Wir benutzen die Fassung von CATSS (COMPUTER ASSISTED TOOLS FOR SEPTUAGINT STUDIES), hergestellt an der University of California, freigegeben für Forschungszwecke, nicht jedoch für kommerzielle Nutzung. Als Tradenten sind für uns wichtig: Dr. WINFRIED BADER, Anfang der 1990er Jahre Mitarbeiter am Zentrum für Datenverarbeitung der Universität Tübingen, und MARTIN SCHINDELE, damals Mitarbeiter am Arbeitsbereich »Textwissenschaft« der Fakultät für Informatik.

[134] Das hebräische Korpus umfasst 39 Bücher, das griechische 60, z. T. in (partieller) Doppelüberlieferung bzw. anderer Bündelung. Die Besonderheiten der LXX und unsere Abkürzung in ( ): Joshua B-Text (JOS-B), Joshua 15,21*–62.18,21*–19,45, A-Text (JOS-A), Richter, B-Text (RI-B), Richter, A-Text (RI-A), 3 Könige (1KON), 4 Könige (2KON), 1 Esra (1ESDR), 2 Esra (2ESDR(ESR-NEH)), Tobit, BA-Text (TOB-BA), Tobit, S-Text (TOB-S), 1 Makkabäer (1MAC), 2 Makkabäer (2MAC), 3 Makkabäer (3MAC), 4 Makkabäer (4MAC), Oden (OD), Weisheit (WIS), Jesus Sirach (SIR), Psalmen Salomons (PSSOL), Baruch (BAR), Brief des Jeremia (EPJER), Susanna, ursprünglicher LXX-Text (SUS), Susanna, Text des Theodotion (SUSTH), Daniel, ursprünglicher LXX-Text (DAN), Daniel, Text des Theodotion (DANTH), Bel und der Drache, ursprünglicher LXX-Text (BEL), Bel und der Drache, Text des Theodotion (BELTH).
Zunächst in 1KON, dann aber noch an einer Reihe weiterer Stellen, entdeckten wir gegenüber dem hebräischen Text einen Zusatz. Um ihn unterzubringen und zugleich die Verszählung nach dem Hebräischen nicht zu stören, wurde »V.24« erweitert durch Kleinbuchstaben; zusätzlich wurde durch « angezeigt, dass es sich um zusätzliches Material handelt. Die Zählung sieht somit wie folgt aus:
1KON012,»024a;
1KON012,»024b;
. . .
Wie angedeutet: Die gleiche Adressierung (mit « und Kleinbuchstaben) gibt an einer Reihe weiterer Stellen den

Folglich muss auch das griechische Korpus zuvor – analog dem hebräischen (und jedem weiteren) – aufbereitet werden. Eine Reihe von Aktionen und Entscheidungen ist dafür notwendig:

Der Blick in frühe Codices zeigt,[135] dass in den frühesten uns erreichbaren Manuskripten MAJUSKELN[136] benutzt worden waren, die gewohnten griechischen Akzente ebenso fehlen wie die Kapitel-Verszählung. Außerdem wird der *scriptio continua* gefolgt, d.h. es gibt keine Wortzwischenräume oder gliedernde Satzzeichen. Nicht nur im Blick auf das technische Funktionieren des Programms, sondern auch aus textkritischen Gründen werden folgende Aktionen notwendig:

– **Aktion 1:** Wir integrieren in unseren Text die Kapitel-Vers-Zählung analog dem hebräischen Korpus. D. h. auch die Buchbenennungen sind denen unserer hebräischen Fassung angepasst (faktisch spielen darin auch Konventionen der deutschen Einheitsübersetzung eine Rolle). Nur bei Eigengut der *Septuaginta* wurden die Buchbezeichnungen aus dem CATTS-Projekt belassen.[137]

– **Aktion 2:** Im Gegensatz zu den alten Codices (aus dem 4. Jahrhundert n. Chr.) integrieren wir Wortzwischenräume, machen also die einzelnen Wortformen sichtbar. Anders kann das Programm nicht arbeiten.

– **Aktion 3:** Die Akzente, die als spätere »Zutat« zu beurteilen sind, werden eliminiert.[138]

– **Aktion 4:** Die Satzzeichen werden als spätere »Zutat« beurteilt und eliminiert.[139]

– **Aktion 5:** Der Text enthält *Metainformationen*, die die Suche behindern würden. Es muss außerhalb reiner Suchläufe der Aufgabe nachgegangen werden, derartige Informationen auszuwerten.[140]

Hinweis auf spezifisches Sondergut von LXX.
Bei EPJER war bei der Zählung zu simulieren, dass der Text keine Kapitelangabe, sondern nur Verse aufweist. Die Überschrift ist ohne Zählung. Sie heißt in CoMOn: *EPJER000,000;*. Die Verse werden dann hochgezählt – bei gleichbleibendem 'Kapitel': 000.

135 Vgl. Codex Sinaiticus online: *http://www.codex-sinaiticus.net.*

136 *minuskeln* kennzeichnen erst mittelalterliche Handschriften.

137 Drei Punkte müssen bedacht werden: (1) Die Kapitel- und Versbezifferungen mussten aus dem CATTS-Text extrapoliert werden. Sie waren dort nur mit Platzhalter notiert (z.B. » ˜x« bedeutete: nächstes Kapitel, » ˜y«: nächster Vers). – (2) Es ist bekannt, dass die LXX bisweilen Textpassagen an anderer Stelle bietet als die hebräische Fassung. Das kann dazu führen, dass ein und dieselbe Kapitel-Versangabe in beiden Korpora unterschiedliche Texte addressiert. – (3) Unsere Stellenangaben folgen dem Prinzip: *BUCHZZZ,VVV;* An 169 Stellen begegnet die Modifikation: **BUCHZZZ,"VVV;** u. U. mit Kleinbuchstaben vor dem »;«. Es handelt sich um Textvarianten, die spezifisch für die LXX sind. Texte, die von Haus aus ohne Kapitel-Angabe stehen (z.B. Proömium zu Jesus Sirach), wurden mit einer *dummy*-Angabe ausgestattet: »000,« Ähnlich die ΕΠΙΣΤΟΛΗ ΙΕΡΕΜΙΟΘ, die zwar Verse bietet, aber keine Kapitelangabe. Auch hier gibt es ein nicht-nummeriertes Proömium. Bei uns: »000,000«.

138 Wie bei der hebräischen Fassung könnte *ein* anders gesetzter Akzent zudem den gesamten Treffer bei der Suche eliminieren. Man denke an den regelhaften Wechsel der Akzente am Wortende. – Ausgeschlossen werden also:
(: – Spiritus asper, steht nach dem zugehörigen Vokal/Rho.
) : – Spiritus lenis, steht nach dem zugehörigen Vokal.
+ : – Trema (nach dem zugehörigen Zeichen).
/ : – Akut (nach dem zugehörigen Zeichen).
= : – Zirkumflex (steht nach dem zugehörigen Zeichen).
\ : – Gravis (steht nach dem zugehörigen Zeichen).
| : – Jota subskribtum (nach dem zugehörigen Vokal).

139 . : – Punkt als griechisches Satzzeichen.
: : – Hochpunkt als griechisches Satzzeichen.
; : – Fragezeichen als griechisches Satzzeichen.

140 # : – steht für Apostrof in Jos 15,32 (B-Text, Zeichenfolge KQ#) und als Abschluss einzelner griechischer Grossbuchstaben in Buchüberschriften sowie als Abschluss der als Zahl zu lesenden Teile der Zwischenüberschriften in Ps 118.
* : – Nachfolgendes Zeichen ist ein Grossbuchstabe des griechischen Alphabets.
' : – Apostroph.
– : – Bindestrich zwischen Nummern zusammengefasster Verse und zwischen griechischen Wörtern.
[: – öffnende [-Klammer.
] : – schliessende]-Klammer.
a : – Zusätzliche Information zur Versnummerierung.
b : – Informationseinheit in mit ˜ beginnenden Zeilen.

– **Aktion 6:** Für den Suchlauf belassen wir die Großbuchstaben des CATTS-Projekts.[141] Für die Endausgabe der Ergebnisse (Button) benutzen wir die gewohnteren *minuskeln*. Im Griechischen – wie die letzte Anmerkung zeigt – ist nur *ein* Buchstabe dann zu verändern, wenn er am Wortende zu stehen kommt: ς.

Zwischenreflexion zum Verhältnis: Hebräisch – Griechisch:

Der automatische Suchlauf dauert bei einem griechischen Korpus länger als bei einem hebräischen. Dafür mögen im Einzelfall unterschiedliche Faktoren verantwortlich sein. Ein wesentlicher liegt in der unterschiedlichen Sprachstruktur. Während es im Hebräischen gängig ist, dass *eine* Wortform z.B. proklitische Präposition, Wortkern und enklitisches Personalpronomen enthält, so werden im Griechischen (und auch im Deutschen) diese Worttypen als 3 separate Wörter geboten. Im Vergleich der Sprachen bekommt die Suchbedingung: »Mindestlänge 3« einen deutlich anderen Charakter. Um im Griechischen auf einen ähnlichen Bedeutungsgehalt zu kommen wie im Hebräischen müsste man eher »Mindestlänge 5« oder gar »6« einstellen. Vom **Suchaufwand** her heißt das: belässt man die Einstellung »Mindestlänge 3«, so liefert das Programm mit hohem Aufwand sehr viele Befunde, die – auf das Hebräische übertragen – dort wegfallen würden, also die Suche entlasten würden, weil sie die Mindestlänge nicht erreichen (3 *tokens* im Griechischen sind im Hebräischen oft nur 1).[142] Wenn es die Fragestellung zulässt, sollte bei griechischen Korpora die Mindestlänge also angehoben, oder alternativ der Suchtext verkürzt werden.

---

c : – Informationseinheit in mit ˜ beginnenden Zeilen.
k : – Zusätzliche Information zur Versnummerierung.
p : – Informationseinheit in mit ˜ beginnenden Zeilen.
t : – Informationseinheit in mit ˜ beginnenden Zeilen.
w : – Zusätzliche Information zur Versnummerierung.
x : – Informationseinheit in mit ˜ beginnenden Zeilen und Zusatzangabe zur Versnummerierung.
y : – Informationseinheit in mit ˜ beginnenden Zeilen.
z : – Zusätzliche Information zur Versnummerierung.

[141]

		minuskeln
A : – Alpha.	A	α *U+03B1 (945)*
B : – Beta.	B	β *U+03B2 (946)*
C : – Xi.	Ξ	ξ *U+03BE (956)*
D : – Delta.	Δ	δ *U+03B4 (948)*
E : – Epsilon.	E	ε *U+03B5 (949)*
Z : – Zeta.	Z	ζ *U+03B6 (950)*
F : – Phi.	Φ	φ *U+03C6 (966)*
G : – Gamma.	Γ	γ *U+03B3 (947)*
H : – Eta.	H	η *U+03B7 (951)*
I : – Jota.	I	ι *U+03B9 (953)*
K : – Kappa.	K	κ *U+03BA (954)*
L : – Lamda.	Λ	λ *U+03BB (955)*
M: – My.	M	μ *U+03BC (956)*
N : – Ny.	N	ν *U+03BD (957)*
O : – Omikron.	O	ο *U+03BF (959)*
P : – Pi.	Π	π *U+03C0 (960)*
Q : – Theta.	Θ	θ *U+03B8 (952)*
R : – Rho.	P	ρ *U+03C1 (961)*
S : – Sigma.	Σ	σ *U+03C3 (963)* Schluss: ς: *U+03C2 (962)*
T : – Tau.	T	τ *U+03C4 (964)*
X : – Chi.	X	χ *U+03C7 (967)*
U : – Ypsilon.	Y	υ *U+03C5 (965)*
W: – Omega.	Ω	ω *U+03C9 (969)*
Y : – Psi.	Ψ	ψ *U+03C8 (968)*

[142] Die unterschiedliche Sprachstruktur *und* der größere Textumfang der Septuaginta – beides wirkt sich so aus, dass die Zahl der *tokens* (= Wortformen) in der Septuaginta mehr als doppelt so groß ist wie im Hebräischen (ca. 660.000 gegenüber ca. 305.000). Auch das erklärt den größeren Suchaufwand und Speicherbedarf. – Wenn diese *technischen* Voraussetzungen nicht gegeben sind, kommt es zum Systemabsturz. Am Programm selbst liegt es nicht.

3.5.2.3 Griechisches Neues Testament:

Für das griechische Neue Testament gelten codiertechnisch die gleichen Grundsätze wie für das griechische Alte Testament (siehe dort).

Das bedeutet für die *Buchkürzel*, dass sie – strenger als beim Alten Testament – in eine dreibuchstabige Form gebracht wurden. Das führt teilweise zu ungewohnten Lösungen, die aber eindeutig und unverwechselbar sind: **MAT, MAR, LUK, JOH, APG, ROM, 1KR, 2KR, GAL, EPH, PHP, KOL, 1TH, 2TH, 1TM, 2TM, TIT, PHM, HEB, JAK, 1PE, 2PE, 1JO, 2JO, 3JO, JUD, APK.**

Eine Besonderheit besteht darin, dass zum sekundären Markus-Schluss (MAR 16,9–20) auch eine Variante angeboten wird. Sie wird in der Behelfsnotation »MAR016,920;« mitgeführt.

Übernommen wurde der Text vom »Center for Computer Analyses of Texts (CCAT)« in Philadelphia. Im Fall des griechischen Neuen Testaments handelt es sich um die 2. Auflage (ALAND, MARTINI, METZGER, WIKGREN) von »The Greek New Testament« – für Forschungszwecke entsprechend verändert und um Apparate reduziert. Die Buchangaben wurden orientiert an der Einheitsübersetzung angepasst.[143]

3.5.2.3.1 Griechisches NT + AT:

Wenn schon Altes und Neues Testament *auf Griechisch* zur Verfügung stehen, so kann man beide zu *einem* Korpus verbinden, dabei das jüngere NT an den Anfang stellen, und das griechische AT (= *LXX, Septuaginta*) folgen lassen.

Der Sinn der Aktion ist, dass man auf diese Weise leicht AT-Zitate im NT überprüfen kann. Es kann sich um explizite und ausführliche Zitationen handeln.[144] Aber man kann nun auch – eine Ebene tiefer – gleichen Sprachgebrauch, gleiche Phraseme nachweisen. Die Septuaginta dürfte in viel umfassenderem Maße den Sprachgebrauch der griechisch schreibenden Autoren neutestamentlicher Texte geprägt haben. Entweder bei den von Anfang an griechisch schreibenden Autoren. Oder auch in den Fällen, wo eine ursprüngliche aramäische Fassung des Textes angenommen werden muss, die dann ins Griechische übersetzt worden war.

[143] Ein methodisches Problem kann bezüglich des Neuen Testaments nur erwähnt, nicht aber gelöst werden: die Codierung der hebräischen Bibel bezieht sich auf ein real existierendes Manuskript und gibt dieses – inklusive möglicher Abschreibefehler – getreu wieder. Vollkommen anders beim griechischen Neuen Testament: der Text beruht auf einem mehr oder weniger deutlichen Konsens der Forscher, die sich auf tausende von alten Handschriften stützen. Diese werden bewertet, gewichtet und im einen Fall für »original« gehalten, im anderen für »sekundär«. Etwas despektierlich klingend könnte man von einem *patchwork*-Text sprechen. Folglich geben die Editionen des griechischen NT einen Text wieder, den es in dieser Form als *einen* Text nie belegbar gegeben hat. Er beruht auf der Hoffnung / Annahme / Hypothese, dass die Vielzahl von Textzeugen dazu verhilft, eine Textfassung zu konstruieren, die dem »Urtext« möglichst nahe kommt. Die Kriterien für derartige Entscheidungen sind teilweise altehrwürdig (*lectio brevis, lectio difficilior* usw.), sind aber nicht so stabil, dass sie vor Kurzschlüssen bewahren würden. Auch mit ihnen ist es häufig möglich, den Text nach eigenen Vorlieben zurechtzuschneiden. – Eigentlich beansprucht die Forschergemeinde mit diesem textkritischen Ansatz, dem Endverbraucher das textkritische Urteil abzunehmen. Zwar wird Einblick in die jeweilige Quellenlage (vgl. wissenschaftliche Apparate) gegeben, so dass man sich auch abweichend entscheiden könnte. Aber kompetenter Sachverstand mit entsprechenden Entscheidungen floss bereits in die abgedruckte Textgestalt ein, so dass der Einzelne es wagen müsste, gegen diese Autoritäten aufzubegehren. – Suchtechnisch heißt das, dass dieser Ansatz die Ergebnisse verfälscht. Zwar ist auch das NT eine Schriftensammlung. Aber innerhalb einer einzelnen Schrift könnte damit die Stilistik eines Autors verfälscht worden sein, weil die Wortwahl inzwischen auf vielerlei Quellen gestützt uneinheitlich wurde. Die Wahrscheinlichkeit, es mit dem einheitlichen Schrifterzeugnis eines Original-Autors zu tun zu haben, ist eher gering.

[144] Da das NT seit langem ausführlich erforscht ist, weiß man, dass Bezugnahmen auf das AT häufig über die griechische Übersetzung liefen. Während man zum Nachweis mehrere Buchausgaben heranziehen musste, erledigt dies nun *ein* Programmaufruf.

Programmiertechnisch war die einzige Bedingung gewesen, dass bei der Kombination beider Korpora bei den *Buchbezeichnungen* keine Doppelungen vorliegen – das hätte zu Irritationen geführt. Da dies gewährleistet ist, kann *CoMOn* auch auf dieses Kombinationskorpus angewendet werden.

Abfragen dieser Art sind ein wahrer Härtetest für das Programm. Einerseits ist das zu durchsuchende Korpus (NT + AT, wobei letzteres in griechischer Fassung ohnehin doppelt so groß ist wie das hebräische AT) sehr umfangreich. Geht man – andererseits – bei der Definition des Suchtextes nah an die erlaubte Höchstlänge (6000 Zeichen) – und laufen womöglich noch andere Programme auf dem Rechner –, so ist die Gesamtanforderung an Rechnerleistung und Speicherverwaltung extrem hoch. Entsprechend wird Rechenzeit beansprucht.

Aber selbst unter solchen Extrembedingungen werden fantastische Ergebnisse erzielt, von denen frühere Forschergenerationen nicht einmal zu träumen wagten. Es stehen Ergebnisse in einer Fülle und Präzision zur Verfügung, die mit bisherigen *hand- und buch-gestützten* Verfahren nicht erreichbar waren.

Ein praktisches Beispiel: APG 2,1–47 mit einer Länge von 5103 Zeichen wurde als *searchtext* definiert. Gesuchte Mindestlänge der Treffer waren die voreingestellten »3«. Auf einem *dual core*-Rechner dauerte die Suche im Kombinationskorpus 'NT + AT auf Griechisch' ca. eine halbe Stunde. Gefunden wurden eine Übereinstimmung mit Länge 60 (!) mit Psalm 15, Übereinstimmung der Länge 34 mit Joel 3, Übereinstimmung der Länge 20 mit Ps 109, nochmals Joel 3 nun mit Länge 19, Übereinstimmung der Länge 18 mit Lukas 20, usw. Offenkundig liegt in den bisher genannten Fällen eine direkte literarische Abhängigkeit in Form eines »Zitates« vor. Mit »show context« kann man sich die jeweiligen einzelnen Korpus-Fundstellen anzeigen lassen.

Aktiviert man »generate conclusion« werden entlang dem Suchtext alle Verweisstellen aufgelistet; folglich wird zugleich erkennbar, zu welchen Suchtextpassagen es *keine* Treffer gegeben hatte. Die Ergebnisse können exportiert und entweder gedruckt oder weiterverarbeitet werden: Die Stellenangaben lassen sich sortieren. Dadurch kann man sichtbar machen, mit welchen Korpuspassagen der aktuelle Suchtext bevorzugt verbunden ist – noch effizienter ist es, wenn man zugleich auch die beteiligten *string*-Längen berücksichtigt, also aufaddiert.

Wer je schon ausführlich eigene Erfahrungen im *hand- und buch-gestützten* Verfahren gesammelt hat, und wer gleichzeitig die Interpretationsmöglichkeiten sieht, die eine solche korpuslinguistische Methode bietet, wird fassungslos und zugleich dankbar für diesen Quantensprung in der Datengewinnung sein.

3.5.2.4 Koran – deutsch:

Die von einem arabischen Studierenden weitergegebene deutsche Koranversion[145] wurde ebenfalls auf das Suchprogramm angepasst. Das heißt auch hier dreierlei:

– wie bei der arabischen Version (s.d.) wird in die Surenanfänge »Ordnung gebracht«. Das kann bedeuten, dass die insgesamt gesehen in der Überlieferungsgeschichte ohnehin nicht ganz systematische Verszählung in anderen Textausgaben gegenüber unserer Fassung variiert. Bei uns ist die deutsche Fassung strukturiert wie die arabische:

  Unter Vers »000« sind die diversen Vorspannangaben zusammengefasst und durch Kleinbuchstaben weiter differenziert: »000t« meint immer den **T**itel der Sure, »000r« bezieht sich auf den **R**egistereintrag[146] . »000o« nimmt den Hinweis auf, wo die Sure ge**o**ffenbart worden sei. – Erst nach diesen Vorspannangaben lassen wir mit dem eigentlichen Surentext V.1 beginnen.[147]

---

[145] Übersetzer: EDIN-HUSSEIN TOPCAGIC. Vgl. die Netzadresse: *http://www.druzeonline.com/modules.php?name=Quran&action=viewayat&surano=2*

[146] Dieser eröffnet in traditionellen Ausgaben öfters Vers 1 der jeweiligen Sure); »000b« steht für die **B***asmala*; auch hierbei kann sich eine Differenz in der Verszählung ergeben, weil die Basmala traditionell bisweilen zum Vorspann gezählt wird, in anderen Suren als Teil des V.1. Bei uns ist sie immer Teil des Vorspanns.

[147] Es ist möglich, dass bei Versangaben im Rahmen der Suchtreffer Unstimmigkeiten im Vergleich zur eigenen Ko-

– die Satzzeichen werden nicht eliminiert. Ebenso werden alle Großbuchstaben nicht in Kleinbuchstaben verwandelt.[148] Diese Maßnahmen dienten in früheren Versionen von CoMOn der Konsistenz der Suchergebnisse.[149]

– unsere deutsche Übersetzung enthält häufig Klammerbemerkungen – »(. . .)« – zum besseren Verständnis. Oder es wird der deutsche Surentitel damit ins Arabische transkribiert. – Derartige Erläuterungen sind lesefreundlich, stören aber die Suche empfindlich. Daher eliminieren wir die Klammerbemerkungen.

Alle Korpora integrieren wir in *UTF8* in das *CoMOn*-Tool, d.h. die Suche wird in einer konstanten *Zeichencodierung* durchgeführt. Folglich werden jetzt, beim Wechsel in die *deutsche Schrift*, die spezifischen Zeichen – »*ä ö ü ß*« – ebenfalls im Sinn von *UTF8* codiert. Wenn der Browser, mit dem das Programm aufgerufen wurde, *nicht* entsprechend eingestellt ist, kommt es zu verzerrten Darstellungen der über den *ascii*-Zeichensatz hinausgehenden Schriftzeichen.[150] Das Problem liegt von Haus aus somit nicht beim *CoMOn*-Applet, sondern muss lokal, beim Browser behoben werden. Aber wir versuchen via *CoMOn* eine Problemlösung anzubieten.

3.5.2.5 Quran – arabisch:

Grundsätzliches:

Es gibt genügend *websites*, auf denen der Koran kalligrafisch präsentiert wird. Bisweilen kann man sich auch die selbst gewählten Abschnitte vorlesen lassen. Es ist jedoch schwierig, den arabischen Koran als Rohtext (*plain text*) zu beziehen. Oder wenn man ihn findet, ihn so aufzubereiten, dass er für die Suche geeignet ist.

Die Aufgaben, die sich stellen, sind jedenfalls absehbar:

Zunächst besteht eine *weltanschauliche Hürde*: Ein elektronisch bereitgestellter arabischer Korantext kann unter der Bedingung benutzt werden, dass der Text selber in keiner Weise verändert wird. Wird diese Bedingung von islamischer Seite gestellt, kann unterstellt werden, dass die Gleichung: »Korantext = Gotteswort« wirksam ist.

Unsere Frage- und Problemstellung ist damit aber nicht erfasst: Analysen sollen am Korantext durchgeführt werden. Mehrere Aspekte gelten dabei:
– *literarische Analysen* am Koran sind für Teile islamischer Theologen außer Horizont, andere hingegen halten sie für möglich. Das Verbot, den Text zu ändern, scheint von der Seite zu kommen, die an literarische Analysen nicht denkt, sondern das Buch rein für religiös-theologische Ziele verwendet.
– indem wir Vokal- und Lesezeichen eliminieren, verändern wir die Codierung, nähern sie damit sogar der

---

ranausgabe entstehen. In solchen Fällen sollte man einen Blick auf den Vers davor oder danach werfen. Inkonsequenzen bei der Verszählung gehören zur Überlieferungsgeschichte des Koran. Unsere Praxis hat den Vorteil, in der deutschen wie in der arabischen Version die Vorspänne gleich zu behandeln und erst mit dem eigentlichen Text den V.1 beginnen zu lassen.

[148] In ersten Versionen des CoMOn-Programms arbeiteten wir beim deutschen Koran und auch bei den weiteren Texten moderner Sprachen mit der Eliminierung von Satzzeichen und grundsätzlicher Kleinschreibung.

[149] Ein normalerweise klein geschriebenes Wort, das am Satzanfang groß geschrieben wird, würde vom Suchalgorithmus nicht als identisch erkannt – sofern man nicht per Programm darauf reagiert. Inzwischen können wir die Texte belassen und trotz Satzzeichen und Groß-/Kleinschreibung Identitäten erkennen. – Die Lesbarkeit der Ergebnisse wird durch die Beibehaltung des gewohnten Schriftbildes erhöht.

[150] Unsere Erfahrung ist: Browser unter dem Betriebssystem *Linux* erkennen automatisch in *UTF8* codierte Daten, wogegen man dies den Browsern unter *Windows* erst eigens mitteilen muss. Wir arbeiten daran, dass *CoMOn* selbst schon für die richtige Decodierung sorgt.

Ursprungsfassung an, denn diese enthielt zunächst keine Vokalzeichen,
- wie in semitischen Sprachen üblich, beruhen die Bedeutungen bei *schriftlicher* Vermittlung auf der Basis der *Konsonanten*. Im Einklang mit der genannten Bedingung legen daher auch wir Wert darauf, dass es an dieser Basis keinerlei Veränderungen gibt.
- Was hier nicht geleistet werden kann, sind *textkritische* Diskussionen von Schriftvarianten. Kompetente Nutzer unseres *CoMOn*-Tools haben aber die Möglichkeit, ihre Erkenntnisse im »Suchtext« von Hand einzutragen.
- Auf der Ebene der reinen Codierung haben die Anbieter des elektronischen Korantextes selbst massive Änderungen zugelassen und praktiziert. Sie haben das Schriftmedium in die elektronische Speicherung umgesetzt. Das zeigt, dass ähnlich, wie wir es praktizieren, auf die Ebene geachtet werden muss und darf, auf der Änderungen nicht vorgenommen werden dürfen. Die reine Codierung ist jedenfalls unproblematisch.

Mit diesen Überlegungen halten wir fest: bei uns gibt es Änderungen des Korantextes auf der Ebene der Codierungen, in keiner Weise jedoch am Konsonantenbestand. Daher wird unsere Schriftversion um die Vokale reduziert sein, jedoch weiterhin für die genau gleichen Bedeutungen stehen wie die Vollversion.

Transkription für die Suche:

Wenn die Suchergebnisse abschließend präsentiert werden (»Generate conclusion«), lässt sich zwar der Grundbestand arabischer Schriftzeichen wieder herstellen. Aber wie beim Hebräischen fehlen dann weiterhin die Vokal- und Lesezeichen. Und man muss sehen, inwiefern wir dem Spezifikum der arabischen Schrift nachkommen können, das meist eine vierfache Form der Konsonantbuchstaben kennt (alleinstehend, rechts-, links-, beidseitig-verbunden), muss man sehen.[151]

Anders als im Hebräischen (in der masoretischen Phase), wo Doppelkonsonanz auch schriftlich zum Ausdruck gebracht wird (*Dageš forte*), ist im Arabischen Doppelkonsonanz ein rein lautliches Problem, ohne Konsequenzen für die schriftliche Realisierung auf Buchstabenebene (jedoch kann bei Assimilation das Zeichen für Vokallosigkeit entfallen; und das für Verdoppelung [*Tašdīd*] hinzutreten – beide werden bei uns als *Zusatz*zeichen eliminiert). Der Wechsel in unserer Transkription zwischen Groß- und Kleinbuchstaben hat somit nichts mit Doppelkonsonanz zu tun (wie im Fall der hebräischen Transkription), sondern bezeichnet unterschiedliche Konsonanten (z. B. »d« ist ein anderer Buchstabe als »D«, ebenso »r« im Gegensatz zu »R«).

Vergleichbar mit dem Hebräischen ist die Behandlung des Alif | Aleph: Die schriftliche Wiedergabe steht einmal für seinen konsonantischen Wert (»ʾ«), in anderen Fällen ist es eine *mater lectionis*, d.h. der scheinbare Konsonant steht für einen Vokal (»ā«). Um diese Doppeldeutigkeit für die Lektüre aufzulösen, setzt das Arabische das *Hamza* als Zusatzzeichen ein. Wo es steht, hat der *konsonantische* Wert des Alif zu gelten (Kehlkopfverschluss *ʾa* bzw. *ʾi*). – Wir sind an diesen Wegweisern für die richtige *lautliche* Wiedergabe des Konsonantenbestandes nicht interessiert, werden also das Hamza ebenso streichen wie die weiteren Vokal- und Lesezeichen.

Rekonstruktion der arabischen Klarschrift:

---

[151] Bei den Fragen der Rekonstruktion des arabischen Textes konnte auf Studierende aus arabisch sprechenden Ländern zurückgegriffen werden. Ihnen sei an dieser Stelle gedankt.

Einige Feinheiten der arabischen Schrift übergehen wir für die Rekonstruktion des Suchtextes in Klarschrift (wie schon im Fall des Hebräischen). Angestrebt werden soll, dass der Suchtext gut in arabischer Klarschrift gelesen werden kann, auch wenn einige Hilfskonstruktionen benutzt werden müssen.

Bei der Buchstabengruppe: *a, d, ḏ, r, z. w* gibt es nicht alle 4 Schreibpositionen, sondern jeweils nur zwei: nach rechts verbundene Endbuchstaben ([Kleinbuchstabe] – gemessen an Schreibrichtung: links ← rechts), unverbunden = alleinstehend ([Großbuchstabe]). Eine Unterstützung für *Ligaturen* gibt es in Unicode nur für *lām-Alif. Fā-Yā* bzw. *Lām-Yā* bleiben bei uns daher unberücksichtigt.

Für die Extraktion des transkribierten Korantextes aus einer *online* verfügbaren arabischen Version leistete TUSTEP (Tübinger System von Textverarbeitungs-Programmen) in beiden Aspekten wertvolle Dienste: (a) als Software, vor allem mit dem Kommando UMWANDLE, (b) konzeptionell als präzise und kompakte Beschreibung der Codierungen fürs Arabische.[152]

Die Zurückverwandlung des transkribierten Suchtextes in die arabische Klarschrift muss die Buchstaben der Transkription übersetzen in das entsprechende Unicode-Zeichen und zusätzlich versuchen, die Veränderung der Einzelzeichen je nach Position im Wort zu berücksichtigen. Da die Suche am reinen Konsonantentext vollzogen wurde, werden nun Umgebungsbedingungen definiert und im Umcodierungsprozess dazwischengeschaltet, mit denen die richtige grafische Gestalt des einzelnen Buchstabens (meist 4 mögliche Varianten) ausgewählt werden kann; initial | medial | final | isoliert.[153]

---

[152] Legt man dann noch C. BROCKELMANN, Arabische Grammatik. Paradigmen, Literatur, Übungsstücke und Glossar. Leipzig [14]1965 daneben, so besteht eine gute Chance, die Geheimnisse der arabischen Schrift zu durchleuchten.

[153]

Tustep-Transkription	Unicode-Wert	gemeint:	
^a	*U+FE8E (65166)*	ʾ/a	final
^A	*U+FE8F (65167)*	ʾ/a	isoliert
^b	*U+FE90 (65168)*	b	final
b	*U+FE92 (65170)*	b	medial
B	*U+FE91 (65169)*	b	initial
^B	*U+FE8F (65167)*	b	isoliert
^t	*U+FE94 (65174)*	t	final
t	*U+FE98 (65176)*	t	medial
T	*U+FE97 (65175)*	t	initial
^T	*U+FE95 (65173)*	t	isoliert
^o	*U+FE9A (65178)*	ṯ	final
o	*U+FE9C (65180)*	ṯ	medial
O	*U+FE9B (65179)*	ṯ	initial
^O	*U+FE99 (65177)*	ṯ	isoliert
^j	*U+FE9E (65182)*	ǧ	final
j	*U+FEA0 (65184)*	ǧ	medial
J	*U+FE9F (65183)*	ǧ	initial
^J	*U+FE9D (65181)*	ǧ	isoliert
^h	*U+FEA2 (65186)*	ḥ	final
h	*U+FEA4 (65188)*	ḥ	medial
H	*U+FEA3 (65187)*	ḥ	initial
^H	*U+FEA5 (65189)*	ḥ	isoliert
^x	*U+FEA6 (65190)*	ḫ	final
x	*U+FEA8 (65192)*	ḫ	medial
X	*U+FEA7 (65191)*	ḫ	initial
^X	*U+FEA5 (65189)*	ḫ	isoliert
^d	*U+FEAA (65194)*	ḏ	final
d	*U+FEA9 (65193)*	d	isoliert

Die 4 möglichen **Umgebungsbedingungen** lauten:

ˆD	*U+FEAC (65196)*	d̲	final
D	*U+FEAB (65195)*	d̲	isoliert
ˆr	*U+FEAE (65198)*	r	final
r	*U+FEAD (65197)*	r	isoliert
ˆR	*U+FEB0 (65200)*	z	final
R	*U+FEAF (65199)*	z	isoliert
ˆs	*U+FEB2 (65202)*	s	final
s	*U+FEB4 (65204)*	s	medial
S	*U+FEB3 (65203)*	s	initial
ˆS	*U+FEB1 (65201)*	s	isoliert
ˆw	*U+FEB6 (65206)*	š	final
w	*U+FEB8 (65208)*	š	medial
W	*U+FEB7 (65207)*	š	initial
ˆW	*U+FEB5 (65205)*	š	isoliert
ˆc	*U+FEBA (65210)*	ṣ	final
c	*U+FEBC (65212)*	ṣ	medial
C	*U+FEBB (65211)*	ṣ	initial
ˆC	*U+FEB9 (65209)*	ṣ	isoliert
ˆg	*U+FEBE (65214)*	ḍ	final
g	*U+FEC0 (65216)*	ḍ	medial
G	*U+FEBF (65215)*	ḍ	initial
ˆG	*U+FEBD (65213)*	ḍ	isoliert
ˆp	*U+FEC2 (65218)*	ṭ	final
p	*U+FEC4 (65220)*	ṭ	medial
P	*U+FEC3 (65219)*	ṭ	initial
ˆP	*U+FEC1 (65217)*	ṭ	isoliert
ˆz	*U+FEC6 (65222)*	ẓ	final
z	*U+FEC8 (65224)*	ẓ	medial
Z	*U+FEC7 (65223)*	ẓ	initial
ˆZ	*U+FEC5 (65221)*	ẓ	isoliert
ˆy	*U+FECA (65226)*	‘	final
y	*U+FECC (65228)*	‘	medial
Y	*U+FECB (65227)*	‘	initial
ˆY	*U+FEC9 (65225)*	‘	isoliert
ˆv	*U+FECE (65230)*	ġ	final
v	*U+FE8D0 (65232)*	ġ	medial
V	*U+FE8CF (65231)*	ġ	initial
ˆV	*U+FE8CD (65229)*	ġ	isoliert
ˆf	*U+FED2 (65234)*	f	final
f	*U+FED4 (65236)*	f	medial
F	*U+FED3 (65235)*	f	initial
ˆF	*U+FED1 (65233)*	f	isoliert
ˆq	*U+FED6 (65238)*	q	final
q	*U+FED8 (65240)*	q	medial
Q	*U+FED7 (65239)*	q	initial
ˆQ	*U+FED5 (65237)*	q	isoliert
ˆk	*U+FEDA (65242)*	k	final
k	*U+FEDC (65244)*	k	medial
K	*U+FEDB (65243)*	k	initial
ˆK	*U+FED9 (65241)*	k	isoliert
ˆl	*U+FEDE (65246)*	l	final
l	*U+FEE0 (65248)*	l	medial
L	*U+FEDF (65247)*	l	initial
ˆL	*U+FEDD (65245)*	l	isoliert
ˆm	*U+FEE2 (65250)*	m	final
m	*U+FEE4 (65252)*	m	medial
M	*U+FEE3 (65251)*	m	initial
ˆM	*U+FEE1 (65249)*	m	isoliert
ˆn	*U+FEE6 (65254)*	n	final
n	*U+FEE8 (65256)*	n	medial
N	*U+FEE7 (65255)*	n	initial
ˆN	*U+FEE5 (65253)*	n	isoliert
ˆe	*U+FEEA (65258)*	h	final
e	*U+FEEC (65260)*	h	medial

blank-[Buchstabe]–*Buchstabe*	initial für [Buchstabe]
Buchstabe-[Buchstabe]–*Buchstabe*	medial für [Buchstabe]
Buchstabe-[Buchstabe]–*blank*	final für [Buchstabe]

Der Faktor *blank* kann auch *im* Wort eintreten, wenn der vorausgehende Buchstabe keine linksverbindende Form hat (in arabischer Schreibrichtung gedacht), oder auch, wenn der nachfolgende Buchstabe keine rechtsverbindende Form aufweist. Letzte Möglichkeit: die umgebenden Buchstaben hätten sehr wohl verbindende Formen, aber nicht der aktuell zur Debatte stehende Buchstabe.

Tritt also mit echten *blanks* und/oder mit fehlenden Buchstabenverbindungen die Isolierung eines Buchstabens ein, gilt:

blank-[Buchstabe]–*blank*	isoliert.

Die Buchstaben, die alle 4 Realisierungsformen aufweisen, sind in der Transkription links wiedergegeben mit (Gross-/Kleinschreibung irrelevant): *b, t, o, j, h, x, s, w, c, g, p, z, y, v, f, q, k, l, m, n, e, i.* **G1** bezeichnet diese Gruppe.

Nur »final« oder »isoliert« begegnen: *a, d, D, r, R, u.* **G2** bezeichnet diese Gruppe.

Es können somit – zunächst grob unterschieden – folgende Figurationen auftreten (in späterer Klarschrift: rechts → links) – dabei werden die Umgebungsbedingungen wie folgt unterschieden (alles in späterer Klarschrift gedacht, also Schreibrichtung rechts → links):

(»-v« = **keine** Verbindung vom rechten Zeichen her möglich: rechts vom fraglichen Zeichen steht entweder ein blank oder ein Element der Zeichengruppe G2. In dieser Konstellation muss der aktuelle Buchstabe entweder »initial« oder »isoliert« sein.[154]

»+v« = Verbindung vom rechten Zeichen her möglich: d.h. rechts steht aus G1 ein Buchstabe, der entweder beidseitig verbunden ist, oder zumindest nach links zum nächsten = aktuellen.[155]

»v+« = Verbindung zum linken Zeichen möglich: der aktuelle Buchstabe ist (1) offen für den Folgebuchstaben (gehört also G2 an) und – (2) – der Folgebuchstabe ist entweder Endbuchstabe oder beidseitig verbunden.[156]

E	*U+FEEB (65259)*	h	initial
ˆE	*U+FEE9 (65257)*	h	isoliert
ˆu	*U+FEEE (65262)*	w/u	final
ˆU	*U+FEED (65261)*	w/u	isoliert
ˆi	*U+FEF1 (65266)*	y/i	final
i	*U+FE8F4 (65268)*	y/i	medial
I	*U+FEF3 (65267)*	y/i	initial
ˆI	*U+FEF1 (65265)*	y/i	isoliert
Lām-Alif	*U+FEFB (65275)*	Ligatur	isoliert
Lām-Alif	*U+FEFC (65276)*	Ligatur	final

Wie angedeutet: wir arbeiten am transkribierten Korantext *ohne* die Vierer- bzw. Zweierdifferenzierung der Konsonanten. Letztere wird erst für die Schlussauswertung per Programm wieder hergestellt.

[154] Die Entscheidung hängt davon ab, welches Zeichen auf den aktuellen Buchstaben folgt *und* davon, ob der aktuelle Buchstabe zu G1 oder G2 gehört. Nur in letzteren Fall kann er die Verbindung aufnehmen.

[155] D.h. es kann sich auch um einen Anfangsbuchstaben handeln, der nach links offen ist.

[156] In beiden Fällen bekäme der aktuelle Buchstabe eine Andockmöglichkeit.

»v-« = **keine** Verbindung zum linken Zeichen: ausgehend vom aktuellen Zeichen ist das folgende entweder ein blank. Dann muss das aktuelle das Wort abschließen durch die Endgestalt. Oder das Wort ist noch nicht zu Ende, aber es folgt aus G2 ein Buchstabe in »isoliert«-Form. Dann kann das aktuelle Zeichen – vorausgesetzt, es gehört zu G1 – auch nicht andocken.

Es müssen also *Kompatibilitäten* integriert werden.

G1 hat alle 4 Positionen.

G2 hat nur final bzw. isoliert als Positionen.

Umgebung1 + Buchstaben + *Umgebung2*: Das ist das Muster, mit dem die richtige Form für den einzelnen Buchstaben bestimmt werden kann. Beispiele – nun in Schreibrichtung der Transkription, also von links nach rechts:

final G2 *G2* ergibt: »isoliert« für das erste G2, da G2 nach dem vorausgehenden »final«-Signal im Wort nur neu beginnen kann. Das zweite G2 könnte zwar die Verbindung zum ersten aufgreifen. Von dort gibt es aber keine Andockmöglichkeit.

initial G2 *G1* ergibt: »final« für G2, da G2 das vorausgehende »initial«-Signal aufgreifen, aber es hat keine »medial«-Form, kann also nicht überleiten zum des folgenden G1.

medial G1 ergibt: »final« für G1, da G1 das vorausgehende »medial«-Signal aufgreifen kann. Es folgt dann das Wortende.

blank G1 *G2* ergibt: »isoliert« für G1, da G1 am Wortanfang steht und kein »medial« bietet (ansonsten könnte eine Verbindung zum folgenden G2 geschaffen werden).

Zur weiteren Erläuterung kann gesagt werden: die Rede von *Anfangs*position (initial) meint nicht nur den Wortanfang. Sie ist auch innerhalb eines Wortes erwartbar, wenn nämlich der vorausgehende Buchstabe keine Linksverbindung anbietet. – Analog *End*position (final): auch sie ist nicht auf das Wortende beschränkt, sondern wird *im* Wort realisiert, wenn der nachfolgende Buchstabe keine Andockmöglichkeit bietet (*medial* oder *final*). – Letzte Variante: *isoliert* heißt nicht zwingend, ein Buchstabe sei von *blanks* umgeben. Er kann auch von Buchstaben umgeben sein, die jedoch im Fall des vorausgehenden über keine Linksverbindung verfügen, im Fall des nachfolgenden über keine Rechtsverbindung.

»Lām-Alif« repräsentieren die Konstellation: »G1 + G2«. Wenn das *Lām* in Initial- oder Medialposition auftritt: in beiden Fällen werden die zwei Buchstaben durch eine Ligatur verbunden. Weitere an sich mögliche Ligaturen sieht Unicode nicht vor.

Diese Gruppenbildung (G1, G2) zusammen mit den Positionsmustern erlauben die Bildung von Regeln, nach denen jeder Transkriptionsbuchstabe in seine korrekte Schriftform (Unicode) umgewandelt werden kann. Ist diese Umcodierung geleistet, gilt es nur noch die Schreibrichtung umzudrehen; links ← rechts.

Aber offenkundig haben die Java-Herausgeber erkannt, dass die akzeptable Darstellung arabischer Schriftzeichen nicht jedem einzelnen Nutzer aufgebürdet werden sollte. Daher bieten sie ein Programm an, das die korrekte Umsetzung der Buchstaben durchführt. Grundlage sind die »General Unicode«-Codierungen. Die Umsetzung in die richtige Kontextform des Buchstabens besorgt das Programm.[157]

[157] Hätten wir davon früher gewusst, hätte eine Reihe von Experimenten, die in diesem Punkt sich niedergeschlagen haben, ausfallen können . . . Aber das gedankliche Eindringen in das arabische Schreibsystem wäre dann auch entfallen.

Unsere Aufgabe beschränkt sich darauf, die für die Suche viel günstigeren TUSTEP-Codierungen in den »General Unicode« zu konvertieren.[158]

Suren und Verse:

Grundsätzlich kann beim Koran die gleiche Datenstruktur angewendet werden wie bei der Bibel. Der Unterschied ist ein semantischer: die vorangestellten Buchstaben (»QURAN«) bezeichnen immer das ganze Buch, wogegen in der Bibel Einzelschriften an dieser Stelle genannt werden.

Es folgt dann dreistellig die Bezifferung der »Sure«. Sie entspricht einem biblischen »Kapitel«. Nach Komma folgt dreistellig die »Vers«-Angabe – wie im Falle der Bibel. Abgeschlossen wird dieser Komplex durch »; «

Der Beginn der Suren weist Unregelmäßigkeiten auf bzw. dem eigentlichen Text vorangestellte Elemente. Immer rechnen kann man mit der »Überschrift«, dem »Titel«. Wir verwenden dafür die Versangabe: »ZZZ,000t;«

Darauf folgt meist, aber nicht immer, die *basmala* (»Im Namen Allahs des Allerbarmers«). In Sure 9 fehlt diese Formel. Wir verwenden für die *basmala* die Versangabe: »ZZZ,000b;« In der Koranüberlieferung ist es nun unklar, ob die *Basmala* als eigener Vers gezählt werden soll – so dass die Zählung aller folgenden Verse um eins erhöht ist, oder eben nicht. Und es gibt die Zwischenstufe, wonach das erste Auftreten der *Basmala* in Sure 1 gezählt wird, die späteren Wiederholungen jedoch nicht.

---

[158]

TUSTEP (für Suche)	**General Unicode**
a	0627
b	0628
t	062A
o	062B
j	062C
h	062D
x	062E
d	062F
D	0630
r	0631
R	0632
s	0633
w	0634
c	0635
g	0636
p	0637
z	0638
y	0639
v	063A
f	0641
q	0642
k	0643
l	0644
m	0645
n	0646
e	0647
u	0648
i	064A

Wir benötigen ein transparentes Verfahren, ohne deswegen zum Schiedsrichter für die unterschiedlichen Praktiken zu werden: Wir zählen die *Basmala* nie, geben ihr – wie beschrieben – immer die Verszahl »,000b; «. Damit folgen wir der Mehrheitsmeinung, allerdings mit dem Unterschied, dass in der kurzen Sure 1 (wo viele die *Basmala* zählen), die Verszählung divergieren könnte.

Wer demnach den Eindruck hat, die Stellenangabe unserer Treffer passe nicht – verglichen mit seiner gewohnten Koranausgabe –, *sollte die Verszahl des Treffers um eins variieren (d.h. meist: erhöhen)***. Damit kann man in das andere Zählsystem umsteigen. Andere Differenzen in der Zählung werden nicht auftreten.**

Ein weiteres Element am Surenbeginn, kann als ein *Registraturhinweis* angesehen werden. Das Element kann auch fehlen. Wenn es steht, sind es Abkürzungen, die am Beginn des ersten Verses genannt sind. Anschließend beginnt dann der eigentliche Surentext. Wir verfahren so, dass auch dieser Registraturhinweis – wo er denn steht – unter Ziff. »,000r;« geführt wird, so dass bei uns der jeweilige V.001 erst mit dem eigentlichen Surentext beginnt, ohne die organisatorischen Zusatzblöcke.

3.5.2.6 Günter Grass, »Die Blechtrommel«

Die »Blechtrommel« in elektronischer Form diente bei uns unterschiedlichen Forschungszwecken.[159] Eine weitere Möglichkeit der Auswertung besteht darin, den Text auch in das CoMOn-Tool zu intergrieren. Damit hat man die Möglichkeit, für einen gewählten Ausschnitt (Suchtext) nachforschen zu lassen, inwiefern seine Wortverbindungen im übrigen Teil des Romans vorkommen.

In punkto *Zählung* kann die von der Bibel her bekannte Struktur im Prinzip übernommen werden (was die Integration des Textes in das Programm erleichtert), auch wenn die faktische Anwendung etwas variiert:

> Der Roman ist unterteilt in 3 Bücher.[160] Diesen könnte man die Etiketten vergeben: *1BUCH, 2BUCH, 3BUCH*. Aber wir halten die Orientierung an den Kapiteln für ergiebiger.
>
> Die Kapitel werden im Original nicht eigens gezählt. Wir führen zusätzlich die Zählung ein. Also folgt auch bei uns auf den Buchnamen »BLECHTR« die Kapitel-Ziffer. [161]
>
> Innerhalb der Kapitel finden sich *vom Autor gewollte Paragrafen.* Eine solche *optisch vorgegebene* Gliederung sollte berücksichtigt werden: wir zählen innerhalb eines Kapitels die Paragrafen durch, so dass die Themenangabe = Kapitelüberschrift immer die Nr. »001« hat. Die nach dem Komma folgende Zahl gibt somit an, wie viele Paragrafen das jeweilige Kapitel aufweist.[162]

---

[159] Der Verlag ist entsprechend informiert. – Besonders ergiebig war das Aufspüren von *Alliterationen* gewesen, vgl. *http://www-ct.informatik.uni-tuebingen.de/daten/allit1.pdf.*

[160] Eine Vereinfachung, die wir vornehmen, besteht darin, die drei Großbereiche »Erstes Buch«, »Zweites Buch«, »Drittes Buch« zu übergehen.

[161] Anfangs wirkte sich ein kleiner Fehler bei der Texterfassung aus: Die Kapitel 28–30 waren zu *einem* Kapitel »28« zusammengefasst, so dass die nachfolgenden Kapitel um »2« zu niedrig angezeigt wurden. Bitte dies bei der Auswertung von Abfragen, die **vor dem 11. Februar 2011** durchgeführt wurden, beachten! Seit diesem Datum ist der Fehler behoben.

[162] Anders und mehr zähltechnisch gesagt: die *Paragrafen* der Blechtrommel entsprechen den *Versen* bei biblischen Texten. Das Programm erzwingt, dass das Datenformat der Zählung gleich bleibt. Neben allen inhaltlichen Differenzen in der Interpretation der Zählung gibt es auch quantitative: Häufig übertreffen die *Paragrafen* bei weitem, was

Man könnte noch eine Ebene tiefer steigen und z.B. pro Paragraf/Absatz Sätze unterscheiden (anhand der Satzzeichen). Das wäre zwar kein Problem, würde aber beim Thema »*string*-Suche« keine Vorteile bringen. Im Gegenteil: Unsere *string*-Suche richtet sich gerade nicht nach Satzgrenzen (das wären bedeutungsbasierte Einheiten), sondern kann auch Befunde über Satzgrenzen hinweg liefern.[163]

Es gelten bezüglich Satzzeichen und Groß-/Kleinschreibung die gleichen Ausführungen wie oben bei: »Koran – deutsch«. Inzwischen können wir diese Merkmale belassen – ohne deswegen die Suchergebnisse zu verfälschen. [164]

Das Suchtool erlaubt es, auf der Basis eines gewählten = definierten Suchtextes (z. B. Teil eines Kapitels [solange die erlaubte Maximalgröße von 6000 Zeichen nicht überschritten wird]) zu sehen, wie der Text mit dem Rest des Buches verknüpft ist, ob typische Sprechweisen / Eigentümlichkeiten des Poeten zu erkennen sind.[165]

3.5.2.7 Zeitungskorpus: Frankfurter Rundschau (NEGRA)

Die Computerlinguistik an der Universität des Saarlandes hat aus Jahrgängen der »Frankfurter Rundschau« Beispielsätze extrahiert und sie »annotiert«, d.h. zu jedem Wort in einem Satz wurde eine differenzierte Bestimmung gestellt (Wortart, Funktion im Satz). Zwei Ausgaben gibt es derzeit: NEGRA1 (10.000 Sätze) und NEGRA2 (20.000 Sätze, wobei darin NEGRA1 eingeschlossen ist). [166]

Wir sind am Sprachmaterial interessiert, also an den Beispielsätzen, nicht jedoch – auf der aktuellen Ebene (Ausdruckssyntax) – an den Annotationen.[167] Also extrahieren wir die Beispielsätze, nehmen die umfangreichere Version NEGRA2 in *ein* Korpus und können das Material mit CoMOn durchsuchen.

Da es sich um eine Ansammlung von Einzelsätzen handelt, die zudem von sehr unterschiedlichen Schreibern stammen (Journalisten, Redakteure), variiert der Gebrauch von CoMOn zwangsläufig: Was man finden wird, repräsentiert breiter gestreute Sprachgewohnheiten, kann aber nicht im Sinn von *literarischen* Auffälligkeiten *eines* individuellen Werkes (und eines Abschnittes = Suchtext darin) gedeutet werden.[168]

---

im Rahmen eines biblischen Durchschnitts-*Verses* erwartet werden kann. Folglich wird die Alternative bei der Bestimmung des Suchtextes faktisch außer Kraft gesetzt: Man kann nicht mehr *entweder* bis zu 70 Verse wählen, *oder* sich an der Zeichenanzahl von 6000 ausrichten. Die Zahl 6000 wird in der Regel schon bei deutlich weniger als 70 Versen = Paragrafen erreicht.

163 Es ist dann ab der Bedeutungsanalyse notwendig, mit feinerer Segmentierung zu arbeiten: dann reichen die »starken« Satzzeichen nicht mehr, sondern es werden »Äusserungseinheiten« auch auf der Basis von Bedeutungsverstehen gesucht und gebraucht. Vgl. in diesem Papier Ziff. 1.2.5.

164 Man kann sich bei dieser Gelegenheit aber bewusst machen, dass beide »Objektivierungen« des Textes einen großen grafischen Beitrag leisten zum richtigen Verständnis *der Bedeutungen*. Für die reine Suche an der Buchstabenfolge wären sie unwichtig; aber das jeweils richtige Verständnis anhand *dieser* Textgestalt (Kleinschreibung, keine Satzzeichen) herauszubekommen, löst häufig Irritationen aus.

165 Erste Tests ergaben signifikante Einsichten: (a) Die Dichte der Treffer ist erstaunlich niedrig. (b) Über die Mindestlänge »3« hinaus gehende längere Wortketten sind selten. (c) Die Zahl der Belege pro Treffer ist niedrig. – Schon unter Ausblendung von Erkenntnissen zur *Bedeutung* wird bereits auf *Ausdrucksebene* sichtbar, was man unter *kreativer* Sprache zu verstehen hat: Auffallend wenige Anleihen bei der Art, wie »man« Wörter zu verketten pflegt.

166 Vgl. *http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html.*

167 Für sie ist denkbar, im Bereich Semantik Äquivalenzen zu unserer Terminologie herzustellen und so als Fundus in unserer Datenbank bereit zu halten. Vgl. Ziff. 5.1 und Unterpunkte.

168 Das gibt Anlass daran zu erinnern, dass man mit CoMOn zwar via Zählung einen Suchtext auswählen kann. Aber der lässt sich dann von Hand ändern, bis zum Extrem, dass man den Suchtext von Hand ganz frei als die Wortkette eingibt, an der man im Moment besonders interessiert ist.

Die Besonderheit dieses Korpus veranlasst uns, die schon gegebene Möglichkeit der *freien Eingabe* in CoMOn durch eigenen Button besser sichtbar zu machen: Ergänzend zur Suchtextbestimmung via BUCH|KAPITEL|VERS gibt es einen Button, der die Maske für die Bestimmung von Anfangs- und Endposition *deaktiviert* und direkt auf das Suchtextfenster zur freien Eingabe dessen verweist, was man gesucht haben möchte. Die Funktion der »Freien Eingabe« kann aber auch bei den anderen Korpora benutzt werden. Sie ist immer dann relevant, wenn nicht ein Gesamttext (=Suchtext) der Suche zugrundeliegen soll, sondern einzelne Phrasen – das typische Szenario grammatisch-linguistischer Detailarbeit. In der Praxis wird es immer wieder eine Zweistufigkeit geben: (1) Man sucht via *freie Eingabe* nach Belegen für eine einzelne Phrase. Sobald man auf Belegsätze gestoßen ist, kann man diese – (2) – via Zählung zum Suchkriterium erheben und damit auf weiterführende Befunde treffen.

Die beiden zusammengefassten Korpora werden bei uns als Buch mit der Bezeichnung »2NEGRA« geführt. Während in den Quelldaten die Beispielsätze lediglich bis zum 5-stelligen Bereich hochgezählt werden, haben wir die selben Ziffern zur Wahrung des auch sonst gebräuchlichen Formats grundsätzlich durch führende Nullen in eine sechsstellige Ziffer verwandelt. Nach den drei ersten Ziffern setzten wir ein Komma – auch das, um das bisherige Format zu wahren. Wer genau wissen möchte, um welche Satznummer es sich handelt, denkt sich das Komma weg. Satz Nr. »1« in 2NEGRA hat also bei uns das Aussehen: »000,001«.

Die Chance CoMOn auch bei NEGRA *einzeltextbezogen* zu verwenden ist durchaus gegeben. Häufig kann man als Einzelsätze Äusserungen identifizieren, die offenkundig Zeitungsüberschriften darstellen; andererseits findet man Namenskürzel von Redakteuren. Ein Einzeltext ist demnach das, was sich zwischen Überschrift und Namenskürzel findet.[169] Mehrere Artikel des selben Redakteurs müssten somit stilistisch einheitliche Merkmale zeigen, die sich von einer Textgruppe aus anderer Hand unterscheiden.

3.5.2.8 Mark Twain, »Tom Sawyer« und »Huckleberry Finn«

Vom »Projekt Gutenberg« ließen sich beide Texte herunterladen. Sie wurden ähnlich aufbereitet wie die »Blechtrommel«. Beide Romane wurden in *ein* Korpus »Twain« gepackt, das folglich die Bücher »Tom« und »Huck« enthält. Die Zählung wie im Fall der »Blechtrommel«. Dadurch, dass Twain oft den Südstaaten-Slang wiedergibt, wo vieles verschluckt wird, ergeben sich entsprechend viele einzeln stehende, durch Apostrof abgetrennte Buchstaben. Das die Kürzung anzeigende Apostrof wurde – anders als die übrigen Satzzeichen – *nicht* für die Suche neutralisiert.[170] Damit steht für den Sprachgebrauch von Twain – zwar geschrieben, aber stark an der Phonetik orientiert – ein beachtlich umfangreiches Korpus zur Verfügung.

---

[169] Durch weite Suchtexteingaben, kann man sich längere Passagen anzeigen lassen, um so die Grenzen von Einzeltexten zu erkennen.

[170] Man könnte dies so rechtfertigen: Normale Satzzeichen dienen dazu, Bedeutungseinheiten innerhalb des Text*strings* sichtbar zu machen. Dagegen vertritt das Apostrof einen ausgefallenen Wortteil, trägt also dazu bei, das Geschriebene an das Gesprochene anzunähern. Das Apostrof hat damit eine prinzipiell andere Funktion.

3.5.2.8.1 Marcel Proust, »À la recherche du temps perdu«

Der Roman ist – fast – komplett über *wikisource*[171] zugänglich,[172] Um den umfangreichen Text in CoMOn zu nutzen waren wie bei den anderen Texten umfangreichere Vorarbeiten notwendig.

> Zunächst galt es, das *Zählsystem* zu integrieren. Die einzelnen Bände schlagen sich nieder in der Buchbenennung: *vol.1* ⇒ *PROUST1* usw.
>
> Die erste dreistellige Zahl benennt also bei der ersten Ziffer die Bandzahl, durch die Ziffer unmittelbar vor dem Komma die *Kapitel* – bisweilen im Ursprungstext auch so bezeichnet, aber nicht konsequent. Die zweite dreistellige Zahl (nach dem Komma) beschreibt wieder – wie schon bei der »Blechtrommel« die optisch wahrnehmbaren Einschnitte innerhalb eines Kapitels (Paragrafen), auch in dem Fall dass nur Rede und Widerrede durch neue Zeile getrennt werden.

Die Texte werden im Quelltext elektronisch nicht nur mit Metadaten im Vor- und Nachspann geboten – die es also zu eliminieren galt. Vielmehr sind diverse Verweise in den Fließtext selber eingebaut, markiert z.B. durch[173] »[...]« oder »n#[2013]« oder »[#[2009] ... #[2009]]« oder »⟨ ... ⟩«.[174] Hochstellungen – wie bei Abkürzungen für »Madame« oder »Mademoiselle« oder Ordinalzahlen – werden rückgängig gemacht.

Damit steht Interessenten am Französischen ein umfangreiches Sprachmaterial für gepflegtes Französisch zur Recherche zur Verfügung. Angesichts des Korpusumfangs und falls man bei der Suchtextbestimmung nah an die Schwelle von 6000 Zeichen geht, bekommt auch der modernste Rechner reichlich zu tun, benötigt also Zeit. Aber – die Suche wird abgearbeitet, aufbereitet, und man bekommt Ergebnisse zu Gesicht, die so bislang nicht zugänglich waren.[175]

3.5.2.8.2 Lew N. Tolstoi, »Anna Karenina«

Der russische Text des Romans kann bezogen werden über:
http://az.lib.ru/t/tolstoj_lew_nikolaewich/text_text_0080.shtml
Wie im Fall der anderen Korpora auch, musste er erst noch in das Format gebracht werden, das CoMOn verlangt. Die Arbeit wird erleichtert durch eine offenbar deutlich bessere elektronische Speicherung des Textes (im Vergleich zum PROUST-Text).

---

171 *http://fr.wikisource.org/wiki/*

172 Es fehlen aus dem 5. Band die Abschnitte 9–12. – Es sei angemerkt, dass die Aufbereitung des Textes recht schwierig war, also die Entfernung der Metadaten und Integrierung der Zählung. Nicht auszuschließen sind vereinzelte Textverluste. Zudem ist der elektronisch beziehbare Text hie und da unvollständig. – Trotz all der Einschränkungen steht damit eine beachtliche homogene Textmenge für unsere Untersuchungen zur Verfügung. Jedenfalls übertrifft das Textvolumen sogar das Kombinationskorpus aus Altem und Neuem Testament (griechisch).

173 So zumindest wurden sie nach Umwandlung in unser Analyseprogramm TUSTEP sichtbar.

174 Die Konventionen dabei sind nicht immer konsistent – z.B. an welchen Positionen *blanks* zu erwarten sind – , so dass man mit Variationen rechnen muss.

175 So dauerte die Suche (»min. length«: 3) bei knapp 6000 Zeichen Länge bei einem Test mit Dual core-Rechner etwa eine halbe Stunde, fand aber auch 3769 Treffer. »chemin de fer« kommt 31× im Korpus vor. Die Kette »mon coeur refroidi ne vous entend« nur ein einziges Mal. »que je ne« 398×. Keine längere Passage (aus dem Schlussteil) ist *ohne* Anklang im restlichen Korpus – was für die stilistische Homogenität (= einheitliche Autorschaft) des gewählten Suchtextes spricht. (Anderer Befund bei Sure 1 im arabischen Koran),

Die *Zählung* muss wieder das vom biblischen Befund abgeleitete System ansteuern, um das darauf geeichte *Datenformat in CoMOn* benutzen zu können. Neben diesem technischen Erfordernis sei aber festgehalten: Durch das, was bei uns – auch in den anderen Korpora – als **Vers** geführt wird, erfassen wir **vom Autor gewollte optische Gliederungen** seines Textes: durch absichtliche Zeilenschaltung oder zusätzliche Absatzbildung/Einrückung. Dies sichtbar zu machen ist mehr als ein informatisches Bedürfnis, weil damit die Ausdrucksseite genau dieses Textes, die Art, wie er materiell strukturiert den Leser erreicht, erfasst wird. Folglich kann man sicher sein: Unsere »Vers«-Zählungen beruhen nicht auf Willkür, sondern bilden ab, was vom Autor vorgegeben worden war.[176]

Zu den *Zählelementen* im Detail:

1. Die Werkbezeichnung ist durchgehend: **ANNAK**.
2. Das Werk ist gegliedert in **8 Teile = Bücher**. Wieder ist die **1. Ziffer der Kapitelangabe** für die Buchzählung reserviert. »723,012« meint also eine Abschnittsangabe in Buch/Teil 7.
3. Die nächsten zwei Ziffern vor dem Komma sind der **Kapitelzählung** vorbehalten. In der elektronischen Fassung ist die Kapitelzählung mit römischen Ziffern durchgeführt.[177] Sie mussten nun nur zusätzlich in arabische Ziffern umgesetzt und in unser Zählsystem integriert werden.
4. Nach dem Komma – wie erwähnt – die **Verse**. Mit jedem Kapitelbeginn startet die Zählung neu.[178]

3.5.2.8.3 The Lancaster Corpus of Mandarin Chinese

Über die Homepage des »Oxford Text Archive (OTA)« ist die genannte Sammlung vieler Einzeltexte (über 400) zugänglich.[179] Sie umfasst ca. 15 verschiedene Texttypen. Darunter Nachrichten, literarische Texte, akademische Prosa, offizielle Dokumente usw. – alle Anfang der 1990er Jahre in China publiziert.

Anstelle der üblichen »Buch«-Bezeichnung setzen wir ein Kunstwort ein, weil es nur aus Buchstaben bestehen sollte:

Oxford Text Archive 2 4 7 4

wird bei uns zu

O T A Z V S V

---

[176] Es ist eine kleine Ironie, dass im Herkunftsbereich dieses Zählsystems, in der Bibel, die Verse genau diese Aussagekraft *nicht* haben, denn es liegen uns keine handschriftlichen Originale der Ursprungstexte vor, an denen man die optische Gliederung ablesen könnte. Weitgehend repräsentiert das viel später eingefügte Zählsystem Bedürfnisse, die nichts mit der unmittelbaren literarischen Textgestalt zu tun haben, z.B. gleichförmige Textportionierungen; zusammen mit Akzenten Uniformierung gottesdienstlicher Lesungen.

[177] In der deutschen Übersetzung in arabischen Ziffern: LEW TOLSTOJ, Anna Karenina. Hg. von Gisela Drohla. insel taschenbuch 3484. Berlin 2010. [da »herausgeben« im üblichen Sprachgebrauch nicht das selbe wie »übersetzen« ist, erfährt man merkwürdigerweise nicht, wer den Text übersetzt hat]

[178] Man kann folglich zählen, wie oft es einen Vers »,001« gibt. Antwort: 239×, d.h. soviele Kapitel enthält das umfangreiche Buch. Die Zahl der Verse (in unserem Verständnis), also der optisch vorgegebenen und separierten Einheiten beträgt 7655.

[179] Editoren: A. M. MCENERY; R. XIAO. Zur Verfügung gestellt wurde das Korpus im Jahr 2004.

Z = »zwei«
V = »vier«
S = »sieben«,
V = »vier«

An der »Kapitel«-Position werden die Einzeltexte durchgezählt. Innerhalb der Einzeltexte – gezählt nach dem Komma – kann es Abschnitte/Paragrafen geben.[180] An etwa 10 Stellen enthält der Text *englische* Wortketten. Sie haben wir gelöscht – auch wenn der jeweilige Einzeltext dadurch beschnitten wird. Aber unser Interesse ist es, einen sprachlich einheitlichen Text zu bieten, an dem Wortketten ausschließlich des Chinesischen untersucht werden. Inhaltliche Lücken stören bei dem insgesamt doch recht umfangreichen Korpus nicht. Ein möglichst umfangreiches chinesisches Textmaterial interessiert.

Browser: Wir machten die Erfahrung, dass auf aktuellen Browsern der in UTF8 codierte Text gut dargestellt wird. Bisweilen, wenn dies nicht der Fall war, musste die Einstellung verändert werden:

»Zeichenkodierung« aufrufen, darin:
»Chinesisch vereinfacht (GB18030)« aktivieren

Eine zweite Aufgabe bestand darin, dem JAVA-Programm, auf dem CoMOn basiert, den Umgang mit dem Chinesischen beizubringen. Das war in unserer Version noch nicht gewährleistet.

3.5.2.8.4 Franz Kafka, »Der Prozeß«

Der Roman – so bedeutend er ist – leidet editorisch an den bekannten Problemen. Der Text ist nicht original vom Autor überliefert, korrigiert und abgesegnet. Die Manuskriptblätter hätten ja sogar von MAX BROD verbrannt werden sollen. Dies ist glücklicherweise nicht geschehen – ein »Jahrhundertroman« wäre verloren gewesen. Aber es blieben einige Unklarheiten. Am stärksten die der korrekten Reihenfolge der Kapitel.

Es gibt inzwischen *kritische Editionen*. In diese Debatte konnten wir uns jedoch nicht einschalten. Vielmehr übernahmen wir den Text, so, wie er in »gutenberg.de« zur Verfügung gestellt wird.

Aber auch diese Textgestalt war erst noch verschiedentlich zu bearbeiten. Davon ist näher die Rede in:

http://www.alternativ-grammatik.de

speziell: *Modul 0.12* (erreichbar via Inhaltsverzeichnis). Zu editorischen Fragen wird es dort in Kürze noch weitere Ausführungen geben. **In der »Alternativ-Grammatik«, im angegebenen Modul, sind alle 10 Kapitel in der Textfassung als PDF-files nachlesbar, die auch in *CoMOn* verwendet wurde.**

Hier sind für *CoMOn* zunächst folgende Hinweise wichtig:

---

[180] Die Aufbereitung für CoMOn besorgte 2011 LEI HUANG, ein chinesischer Student der Informatik in Tübingen.

(1) In *CoMOn* zählt als »Book« (aus der Anwendung auf die Bibel übernommen) »ein einzelnes Kapitel« des Romans. Demnach besteht für uns »Der Prozeß« technisch-terminologisch aus »10 Büchern«. Das ist eine informatische, keine literaturwissenschaftliche Konvention. – Dem entsprechen auf der angegebenen Web-Seite der Alternativ-Grammatik 10 PDF-Files mit dem vollen Text des jeweiligen Kapitels.

(2) *CoMOn* sieht bei der Textbestimmung dann die Rubrik »chapter« vor. Sie nützen wir, um *einzelne Abschnitte im Roman insgesamt durchzuzählen.* Konkret: der Roman – so betrachtet – besteht aus 158 Absätzen. Auch zu den Absätzen bietet die Webseite der Alternativ-Grammatik in Ziff. 1.2 einige statistische Befunde.

(3) Unter »Vers« werden einzelne *»Sätze« innerhalb eines Abschnitts* verstanden. Mit deren Zählung (nach dem Komma) wird bei Abschnittsbeginn jeweils neu begonnen.

(4) Hatte man bei der Suchtextbestimmung *abschnittsübergreifend* den Suchtext festgelegt, so wird man mehrere Versangaben wiederholt angezeigt bekommen: z.B. Vers »,001«. Aber es handelt sich um Verse in *unterschiedlichen Absätzen*. Eine Verwechslungsmöglichkeit gibt es letztlich nicht.

(5) Die selben Abschnitts-/Vers-Zählungen sind in die 10 PDF-Files integriert, so dass man auch dort nachschauen und sich orientieren kann.

3.5.2.9 Heatmap

Seit Ende 2012 ist in die Ergebnisausgabe (Drücken des Buttons »Generate Conclusion«) die Option **»Heatmap«** eingebaut (Autor: OMAR EL GHARBI via DA). Dadurch wird dem Benutzer ein Werkzeug angeboten, mit dem er/sie die u.U. erschlagende Fülle von Treffern besser durchschauen kann. Die Treffer können weiterhin lokal gespeichert werden und durch eigens darauf abgestimmte statistische Auswertungen bearbeitet werden.

Die Heatmap liefert jedoch ein erstes, schnelles Bild über die Verteilung der Suchwortketten im gesamten Korpus. In vielen Fällen wird diese auch quantitativ gewichtete Übersicht zur Auswertung bereits genügen oder zumindest klar anzeigen, in welche Richtung die weitere Recherche vorangehen sollte. Einige praktische Hinweise:

Zur Erstellung der Heatmap wird die gewohnte Trefferliste vom Programm herangezogen.

Nach Kapiteln unterschieden (vgl. linke Spalte) werden die Treffer dargestellt: Ganz unten an der Heatmap wird angezeigt, dass die Treffer nach Länge differenziert werden: Hatte man mit »Mindestlänge = 2« die Suche durchgeführt, so führt die erste Farbspalte die Belege für Zweierketten aus dem Suchtext auf, die sich im betreffenden Kapitel ebenfalls finden. Als zweite Spalte werden die 3er-Ketten erfasst, usw.

Die Farbintensität reicht von fast nicht mehr sichtbarem GRÜN bis hin zum tief-sattem ROT. Damit sind die Extreme markiert: von Einzelbelegen bis hin zu ausgesprochen häufigem Vorkommen.

Man kann also ablesen, in welchen externen Kapiteln der aktuelle Suchtext besonders viele bzw. wenige Entsprechungen auf der Ebene der Wortketten aufweist. Wenn belegt, sollten nicht nur die Farbwerte etwa der ersten (linken) Spalten berücksichtigt werden. Sondern zusätzlich etwaige Vorkommen längerer Ketten. Letztere sind naturgemäß seltener, farblich also blasser. Aber weil länger, eben auch von Gewicht. –

Der Hinweis soll verhindern, dass man bei der Auswertung sich nur von den intensiveren Farben gefangen nehmen läßt und die Kettenlänge zu schwach oder gar nicht berücksichtigt.

In der Ergebnisdarstellung folgen die Kapitel nicht in der Reihenfolge des Korpus, sondern nach Intensitätsgesichtspunkten: Am Anfang die Kapitel mit den meisten Bezügen, am Schluss diejenigen mit den geringsten.

Es wird noch angestrebt, die Sortierung nach der ursprünglichen Reihenfolge des Korpus durchzuführen. – Derzeit ist schon eine Sortierung nach »Büchern« möglich: die dazugehörigen Kapitel sind zusammensortiert, so dass man überprüfen kann, wie gleichmäßig bzw. selektiv einzelne Bücher auf das umgebende Korpus verweisen.

Die grafische Ergebnisausgabe kann auch exportiert und in andere Textzusammenhänge integriert werden.

Bei diesem Typ Suche und dieser Auswertung können sehr viele Daten anfallen, was die Arbeitszeit des Rechners verlängert. Es ist daher daran zu erinnern, dass die *Suchtextlänge auf 6000 Zeichen* in CoMOn beschränkt ist. Aus Gründen der Laufzeit ist es aber besser, diese Maximallänge nicht auszureizen, sondern lieber von vornherein kleinere Text'portionen' als Suchtext zu wählen, die Prozedur also mehrfach durchzuführen.

3.5.2.9.1 Ausblicke:

Seit das *tool* in der beschriebenen Form so benutzbar ist, dass es für die Öffentlichkeit freigegeben werden konnte (Herbst 2009, damals beginnend mit dem hebräischen Alten Testament), können weitere Funktionen und weitere Korpora in Angriff genommen werden:

– Es werden weitere, für die Forschung freigegebene Korpora in das Suchtool integriert. Wir nehmen Anregungen (und Korpora) von anderen auf. Wenn das *copyright* unbedenklich ist, die Codierung in UTF8 erfolgte, gilt es nur noch die interne Struktur/Zählung (biblisches Zählsystem bzw. analog) zu überprüfen ggf. anzupassen. Dann kann das neue Korpus geladen und zur Verfügung gestellt werden.

– Button *Tolerance* aktivieren: Die Suche nach kompletter Entsprechung ist der Ausgangspunkt. Liegt im Hebräischen aber der Wechsel von Plene- und Defektivschreibung vor, werden alle Treffer, die sich nur in dieser Hinsicht unterscheiden, übergangen, also nicht erkannt. Folglich soll bald mit *Tolerance* experimentiert werden können. Das heißt, bis zu einem gewissen Maß sind abweichende Wortformen akzeptiert und werden ebenfalls ausgegeben.[181]

– Button *Permutation* aktivieren: Es wird nach Treffern gesucht, deren Wortkette Umstellungen aufweist. Der Befund an Wortformen ist der gleiche, die Reihenfolge ist jedoch anders.

Sobald die beiden Buttons aktiviert sind, werden sie dazu verleiten, in Kombination eingesetzt zu werden. Man sollte sich vor Augen halten, dass dann der Suchaufwand exponentiell steigt. Die »guten«, d.h. interpretierbaren Treffer werden zunehmend unter einem Berg von eher willkürlich erscheinenden verschwinden. Auch wird man die Länge des Suchtextes (aktuell 6000 Zeichen bzw. 70 Verse) deutlich verringern müssen, weil sonst der Speicher im Rechner nicht ausreicht und das Programm abstürzt. Den größten Nutzen wird man haben, wenn man die Parameter nur in geringem Maß verändert bzw. auf deren Kombination verzichtet.

– Als Servicefunktion wird in Kürze die Möglichkeit bestehen, bei der Trefferliste (nach Betätigen von »Generate conclusion«) jede einzelne Stellenangabe anzuklicken und sich dann in einem separaten Fenster den dort geltenden Kontext anzeigen zu lassen. Dann wird man nicht lediglich mit u.U. sehr vielen Stellenangaben konfrontiert, sondern kann im Einzelnen nachschauen, wie der jeweilige Kontext

[181] Es müssen praktische Erfahrungen gesammelt werden, die zeigen, bis zu welchem Ähnlichkeitsmaß die Ergebnisse vorwiegend akzeptabel sind, ab wann dagegen die Treffer eher nach Willkür ausschauen.

des Treffers beschaffen ist – was die Auswertung des intertextuellen Bezugs erheblich erleichtert und beschleunigt.

Ein Ausblick der besonderen Art besteht darin, dass *CoMOn* in überschaubarer Zeit verbunden werden wird mit einigen wortstatisischen Analysemöglichkeiten. Die Datengrundlagen (Korpora) sind vorhanden, auch der jeweils zu definierende Einzeltext. Damit ist es möglich, neben der schon bereit gestellten *Konkordanzfunktion* weitere Analysen durchführen zu lassen, so dass auf Ausdrucksebene der Einzeltext schneller und klarer mit seinem Profil vor Augen steht (z.B. Clusteranzeigen – nicht nur Einzelwortformen, sondern auch Ketten umfassend). *CoMOn* bekäme dadurch ein zweites Standbein neben der bisher schon implementierten Funktion.

Wer vergleichen will, wie unser Konkordanz-Programm sich zu dem verhält, das in die *Stuttgart Electronic Study Bible (SESB)* integriert ist, wird signifikante Unterschiede feststellen:

– die SESB-Konkordanz legt es darauf an, *syntaktisch* gleiche Strukturen zu finden. Der Begriff ist im Sinn der traditionellen Grammatik verstanden: Satzbau, Satzglieder, verlangt also Bedeutungsverstehen. Bei uns, im aktuellen Kapitel 3 *(Ausdrucks-)SYNTAX* liegt ein komplett anderes Verständnis vor: es interessiert die *syn-tax = zusammen-stellung* der Ausdrücke, Wortformen allein. Die daran haftenden Bedeutungen werden komplett übergangen. Die mehrfache Differenz schlägt sich in unterschiedlicher Benutzbarkeit der jeweiligen Tools nieder. Das Programm in SESB verlangt nicht nur, dass man gut Hebräisch beherrscht, sondern auch, dass man mit der nötigen grammatischen Terminologie vertraut ist. Das ist für wissenschaftliche Zwecke ein gutes *tool*. Der Benutzerkreis wird durch die mehrfachen Wissensbedingungen jedoch sehr eingeschränkt. – In unserem Fall sind *keine* Bedeutungsfunktionen integriert. Daher ist es denkbar, dass auch jemand, der des Hebräischen nicht kundig ist, Querverbindungen zu anderen Texten sich auf der Basis des Hebräischen ausgeben lässt, allein mit Hilfe der Stellenangaben.[182]

– die SESB-Konkordanz übergeht die Ebene, die zunächst jeden Leser erreicht und anspricht: die der äusseren Wortformen. Insofern verlangt diese Konkordanz Expertenwissen. Bei unserem Werkzeug wird die Ebene der sinnenhaften Ebene (optisch, akustisch) in ihrem Eigenwert gewürdigt. Man kann leicht beobachten und testen, dass die Orientierung von Lesern sehr schnell auf die Bedeutungsebene zielt. Das ist verständlich, hat doch Sprachgebrauch weitgehend seinen Zweck auf der inhaltlichen Ebene. Es wird aber standardmäßig bei diesem vorschnellen Schritt die erste Ebene vernachlässigt: der Beitrag der Ausdruckselemente – das optisch oder akustisch Gebotene – zur Textrezeption wird übersehen, wird als vernachlässigbar betrachtet. Das ist – um ein großes Wort zu wählen – eine Form von Leibverachtung, linguistischer Manichäismus.

– Es ist in SESB nicht vorgesehen, dass ein ganzer Einzeltext als Suchtext, -kriterium, definiert und dann im gesamten Korpus nach *strings* gesucht wird, wobei die Ketten von Wortformen, für die Treffer gefunden werden, eine unterschiedliche Länge haben können. Lediglich die Mindestlänge wird bei uns definiert. Nach oben muss sich bei uns der Benutzer nicht festlegen. Das erlaubt, dass der Benutzer sich überraschen lassen kann. Er muss nicht vorab definieren, also im Grund schon kennen, wonach gesucht werden soll.

– Damit keine falschen Oppositionen entstehen: das Konkordanz-Programm in SESB kann sinnvolle Dienste erweisen. Es liegt aber auf der Ebene von »Satzbauplänen«, gleichen Inhaltsfunktionen in gleicher Abfolge. Um das Programm zu benutzen muss man den betreffenden Satz verstanden haben.[183]

---

[182] Mehr konkretisiert: Angestoßen durch eine Fragestellung in seiner deutschen Übersetzung will jemand den hebräischen Korpusbefund verifizieren. Konkordanzen zu einzelnen deutschen Übersetzungen gibt es, sie liefern jedoch vom hebräischen Befund abweichende Ergebnisse, da keine Übersetzung den Quelltext eins-zu-eins wiedergibt. Mit *CoMOn* kann man sich Querverbindungen auf der Basis des Hebräischen ausgeben lassen, auch wenn man die Sprache nicht versteht. Die Stellenangaben genügen.

In *wikipedia* wurde der *link* auf das CoMOn-Programm bei mehreren Stichwörtern untergebracht: »Bibelprogramm«, »Altes Testament«, »Neues Testament«, »Koran«. – Eine Erweiterung des Beitrags »Konkordanz« löste Widerstände und – zunächst – die Löschung der Erweiterung aus. Einwand: es werde für eine *software* geworben.[184] In der dazugehörigen Rubrik »Diskussion« wurde von uns darauf in 9 Punkten eingegangen und kann dort nachgelesen werden. Ein *wikipedia*-Problem könnte sich damit gezeigt haben: so *demokratisch* und verdienstvoll das Portal ist, so hat es genau dadurch (auch durch die Supervisoren im Hintergrund) die Gefahr, dass nur aufgenommen wird, was breiter Standard ist. Innovationen haben tendenziell keine Chance. Es wird nivelliert. Im aktuellen Fall kommt die auch sonst breit belegbare Phobie der Geisteswissenschaftler vor der Informatik hinzu. Statt erfreut die Kluft zu überwinden, zu sehen, dass ein mächtiges und wertvolles Werkzeug angeboten wird (das zudem nichts kostet), wird vor der fremden Disziplin gewarnt. Die Phobie bezieht sich auch darauf, Wortketten in den Blick zu nehmen. Stattdessen zieht man sich weiterhin auf das Einzelwort zurück. Die Einzelwortorientierung war verständlich im Zeitalter gedruckter Konkordanzen (wie sollte man komplexere Befunde in Buchform darstellen?). Im Zeitalter des Computers kann man sie aufgeben und die Suche dynamisieren und ausweiten. Schließlich ist die *Phraseologie* ein unstrittig wichtiger Bestandteil der Literaturwissenschaft; und dass »Text« die letztlich entscheidende Größe in der Sprachwissenschaft sein sollte, diese Erkenntnis ist seit Beginn der 1960er Jahre dabei sich durchzusetzen, ab Mitte der 1970er Jahre als »Pragmatik«.

---

[183] Ein Zusatzproblem liegt darin – das soll nicht breit entfaltet werden –, dass zu den Inhaltsfunktionen auch eine Koppelung an bestimmte Ausdrucksmuster verlangt wird. Das macht die Suche noch komplexer und unübersichtlicher. Aber diese Orientierung ist veranlasst durch die traditionelle Grammatikauffassung, die von der Koppelung von Ausdruck und Bedeutung nicht loskommt (*Mophem* ist die kleinste, bedeutungstragende Ausdruckseinheit – wie gebetsmühlenhaft betont wird). Wir trennen durch das ganze SLANG-Projekt hindurch klar zwischen *Ausdruck* und *Bedeutung*, was im Fall des Konkordanzprogramms eine viel höhere Effizienz, Transparenz und viele, sonst meist übersehene Resultate einbringt.

[184] Es wird anschließend um konzeptionelle Aspekte gehen. Dennoch die Rückfrage: Wo liegt der Sündenfall, wenn werbend auf eine *software*-Lösung hingewiesen wird? – Ein solches ’Argument’ ist nichts anderes als ein Ressentiment. In der Welt der Sekundärliteratur ist es gang und gäbe, empfehlend oder warnend mit anderen Publikationen umzugehen. Warum im elektronischen Zeitalter nicht auch mit *software*?

4. Wechsel zur Bedeutungsanalyse (SLANG II)

Die Zäsur ist klar und eindeutig. Entsprechend groß sind die Auswirkungen auf Analysemethoden, Verwendung des Computers und Beteiligung des Benutzers. Mit der Bedeutungsanalyse wird ein qualitativ anderes Feld beschritten. Das heißt zunächst, dass die Einheiten, unter denen der Text bislang betrachtet worden war (Buchstaben, Wortformen, Wortketten), so nicht mehr brauchbar sind. Zwar wird dieses Wissen nicht weggeschoben und ignoriert – die Referenz auf den materiellen Text muss weiterhin erhalten bleiben. Aber *die Bedeutungsanalyse benötigt ihre eigenen Einheiten*. Diese gilt es zunächst zu definieren. Als erstes die kleinsten Bedeutungseinheiten; darauf aufbauend gelangen wir zu immer größeren.

> Die zeichentheoretische Begründung für die Zäsur muss noch etwas differenziert werden: »Zeichen« ist laut DE SAUSSURE eine *mentale* Realität. D. h. der Sprachbenutzer hat – (1) – das Repertoire an Ausdrücken, die in der Einzelsprache gelten, gelernt. Aufgrund dieses Wissens ist er in der Lage, ein Gekritzel mit Buchstaben und Wortformen zu identifizieren. Der Sprachbenutzer hat weiterhin – (2) – inhaltliche Vorstellungen / Konzepte ausgebildet. Er kann die Bedeutung ⟨⟨GEHEN⟩⟩ imaginieren, auch wenn aktuell niemand zu sehen ist, der geht. Und – (3) – er hat gelernt, Elemente des Ausdrucksrepertoires *richtig*, d. h. so, wie es in jener Sprachgemeinschaft *üblich* ist, mit den Bedeutungsvorstellungen zu verbinden.
>
> Ein dreifaches Lernen ist somit die Voraussetzung und wird aktiviert, wenn jemand das Konzept ⟨⟨MANN⟩⟩ mit dem von ⟨⟨GEHEN⟩⟩ verbinden will, und zwar im Französischen: *il va*.
>
> Wer diesen Zusammenhang *informatisch* modellieren will, sollte sich vor Augen halten, dass ein solches »Zeichen«-Verständnis sich völlig von dem unterscheidet, was informatisch eingeschliffen ist. Dort wird oft als »Zeichen« etikettiert, was auf Papier oder am Bildschirm zu sehen ist, also z. B. ein Buchstabe oder eine Ziffer. Oder das, was material realisiert ist, wird als »Symbol« benannt, etwa in logischen Zusammenhängen.
>
> Genau genommen müssten die Begriffe also sorgfältiger verwendet werden: Was in irgendeinem Medium geschrieben ist, ist weder »Zeichen« noch »Symbol«. So spricht man allenfalls abgekürzt. Vielmehr bildet sich das Wissen, es liege ein »Zeichen« oder »Symbol« vor, erst im Geist des Sprachbenutzers. Dort liegt zudem das Wissen bereit, welche Bedeutung mit dem wiedererkannten Ausdruckselement zu verbinden ist.[185]

Der theoretische und methodische Einschnitt beim Übergang von der Ausdrucksstruktur zum Feld der Bedeutungen wird auch über die Art erkennbar, wie der Text im jeweiligen Feld wahrgenommen wird. Wer einen Text liest, folgt der *linearen Reihung* der Ausdrücke (Buchstaben, Wortformen). In dieser Hinsicht ist der Text streng sequenziell gebaut und muss auch so wahrgenommen werden. Diese zeitliche Differenzierung der Textwahrnehmung bleibt auch beim Entschlüsseln der Bedeutungen erhalten. Allerdings erfordert die Konzentration auf die Bedeutungen eine *hierarchische Strukturierung*.[186] Der Leser muss die nur sequenziell entschlüsselbaren Bedeutungsfunktionen[187] auf verschiedenen Ebenen und unterschiedlichen Bezügen / Funktionen hierarchisch richtig zuordnen. In natürlichen Sprachen *kann* die Hierarchie der Bedeutungen innerhalb der Linearität der Ausdrücke überhaupt nicht abgebildet werden.[188] Letztlich ist der gesamte Text hinsichtlich der Be-

---

[185] Das würde die weitere Debattenebene öffnen, wie nämlich natürlich und formale Sprachen zu verstehen seien? Ob letzteren nicht die Bedeutungen ausgetrieben seien – deswegen heißen sie ja »formale« Sprachen, deswegen operieren sie mit Abstraktion, auf Quantitäten hin orientiert – alles Anschauliche, Sinnenhafte ist eliminiert? – Hier nur soviel: auch Abstraktionen und Formalisierungen sind Bedeutungen, wenn auch blasse. *Variable, Objekt, Rekursion, Gleichung, Modalität usw.* sind gedankliche, inhaltliche Konzepte. Wären sie es nicht, könnte formale Logik nur die Distribution von Ausdrücken untersuchen, aber nicht »Schlüsse«, Logiken, Argumentationen usw.

[186] Siehe unten *Appendix 1 – 3*, die alle die Hierarchie eines Textes *auf Bedeutungsebene* abbilden.

[187] Weil die Wortformen als *Träger* nur sequenziell wahrgenommen werden können.

deutungen eine hierarchische Struktur.[189] – Also auch die Tatsache, dass Sprachbenutzer qualitativ grundlegend verschiedene Geistesfunktionen aktivieren müssen, um den sprachlichen Äusserungen gerecht zu werden, spricht dafür, Ausdrucksanalyse und Bedeutungsanalyse klar zu trennen.

Implizit war diese Doppelnatur in der Sprachwissenschaft immer bekannt. So hat es den Charakter eines Glaubenssatzes, wenn *Morphem* von jedem Angehörigen der Zunft definiert wird als »kleinste bedeutungstragende Einheit«. Das meint ja: ein Trägerelement (Laut- oder Schriftzeichensequenz) ist mit einem Bedeutungsmerkmal verbunden. Die gleiche Doppelnatur wird im CHOMSKY-Kontext (und Derivaten) durch *hybride* Baumstrukturen dargestellt: Sie münden in die lineare Kette der Wortformen eines gegebenen Satzes; darüber wird hierarchisch sichtbar gemacht, was wie zusammenhängt, unter- bzw. übergeordnet ist.

Die *methodische* Frage angesichts dieses Wissens ist, ob weiterhin *beide*, theoretisch und praktisch so klar unterscheidbaren Aspekte *verbunden* analysiert werden sollen?[190] Oder ob aus der klaren Theorie nicht endlich auch methodisch die nötigen Folgerungen zu ziehen wären? – Für uns ist die Frage eine rhetorische, da die Weichen bereits gestellt sind: Das vorige Kapitel hat schon den ersten Teil der Weichenstellung durchgeführt. Nun geht es um den zweiten, also die eigenständige Analyse der Bedeutungsstruktur eines Textes.

Der Benutzer eines Analyseprogramms muss also nun, wenn der Bedeutungsbereich hinzugenommen wird, in unvergleichlich höherem Maß Entscheidungen treffen. Es ist sein Sprachwissen, -verstehen, das benötigt wird. Und der Benutzer muss sich seinerseits vergewissern, ob er sich in Einzelfällen ein Sonderverstehen zurecht gelegt hat, oder ob er noch im allgemeinen Konsens handelt. Er muss also immer wieder externes Wissen, Lexikon- und Grammatikwissen, beiziehen um sein eigenes Sprachverstehen mit dem der anderen abzugleichen.

Das Nicht-Beachten der scharfen Zäsur zwischen Ausdruck und Bedeutung hat in der Vergangenheit häufig zu einem *informatischen Scheitern* geführt. Die Daten, die der Rechner benutzen kann, sind die, die digital codiert, also »geschrieben« sind. Die digitale Codierung ist auch eine Form von Schrift, von Ausdruckselementen. Damit kann die Maschine virtuos umgehen. Wogegen sie »Gedanken, Bedeutungen, Argumente« nicht bearbeiten kann: sie haben immateriellen Charakter.

Dies ist ein Dilemma: nun hat man seit einem halben Jahrhundert leistungsfähige Maschinen zur Verfügung. Sie sind jedoch in einem Bereich leistungsfähig, der Benutzer meist nicht so sehr interessiert, in der Ausdrucksebene. Viel stärker erwünscht wäre die Effizienz des Computers im Feld der Bedeutungen.

Die Strategie, um dem Dilemma zu entkommen, bestand nun häufig darin, dass man hoffte, über raffinierte Analyse der Ausdrucksebene (= Stärke des Computers) Aufschlüsse über die dahinter liegende Bedeutungsstruktur zu erhalten. Man wollte sozusagen über die Hintertür die Leistungsfähigkeit des Computers auch für die Bedeutungsebene nutzen.[191]

---

[188] Formale Sprachen könnten sich verschiedener Stufen von Klammerung bedienen, um die Hierarchie doch in die Linearität zu zwingen.

[189] Ausgehend von Einzelbedeutungen, bei denen zwischen unselbstständig (*stop words*: »dass, für«) und selbstständig (»Gericht, kaufen«) unterschieden wird, über Sememgruppen (»Bienenhonig«), Satz (»Die Olympiade ist eröffnet«), Makro- Satz, Textgrammatische (TGE) bzw. Textlinguistische Einheit (TLE) – die letzten drei Begriffe werden später erklärt. ? ? ?

[190] Trotzig untermauert durch den Hinweis, der ebenfalls einem Glaubenssatz gleicht: bei natürlicher Sprache könne man »Form und Inhalt nicht trennen«.

[191] Vgl. H. KLEIN, Computerunterstützte Inhaltsanalyse mit INTEXT. Münster 1996. Es wird mit verschiedenen *Kategorien* gearbeitet, z. B. »Ausmaß eines Ereignisses«, »Konsequenzen eines Ereignisses«, »Nähe (Ort und Wichtigkeit«, »Prominenz« usw. Die *Methode* ist die, dass nach typischen Stichwörtern gesucht wird (aufgrund von Inhaltswissen). Bei manchen Kategorien heißt es, sie seien »leicht zu operationalisieren« (zu »Prominenz« kann man die Stichwörter: »Politiker, Sportler, Kulturschaffende« bzw. auch deren Eigennamen vorgeben). Andere, z. B. »Humor / Spaß« sind für »computerunterstützte Inhaltsanalyse in der heutigen Form nicht geeignet« (166ff). – Dem muss man hinzufügen: Vorgabe von Suchwörtern ist denn auch eine allzu schlichte Modellierung.