In: Hans Fix, Annely Rothkegel, Erwin Stegentritt (Hrsg., 1982): Sprachen und Computer. Festschrift zum 75. Geburtstag von Hans Eggers. Dudweiler: AQ-Verlag, 61-80

HARALD H. ZIMMERMANN

# Maschinelle Verfahren zur Erschließung von Textkorpora*

Als in den 60er Jahren in der Bundesrepublik Deutschland die ersten Bände der Indices zur deutschen Literatur[1] erschienen, die mit Hilfe eines Computers erstellt waren, steckte die maschinelle Sprachdatenverarbeitung noch in den Kinderschuhen. Ähnliches galt (und gilt heute z.T. noch) für die Erstellung mehr oder minder repräsentativer Korpora2, wie sie z.B. das LIMAS-Korpus[3], das Lunder Zeitungs-Korpus[4] oder die Korpora des Instituts für Deutsche Sprache[5] darstellen.

Bei diesen ersten Textauswertungen mithilfe von Computern stehen die *reinen* Rechen- und Ordnungsfunktionen von EDV-Anlagen im Vordergrund, d.h., die Fähigkeit, Wörter (besser gesagt: Wortformen) zu *sortieren* (z.B. alphabetisch vor- und rückläufig), nach Häufigkeiten zu *ordnen* (Frequenzwörterbuch), gegebenenfalls auch noch die formale Funktion, gespeicherte Daten in einem vorgegebenen Kontext (Zeile, Satz) aufzuzeigen. Vielfach wurden (und werden) derartige Produkte einer Computeranalyse schon als Endergebnis angesehen und in dieser Rohform den möglichen Nutzern (vor allem den Sprach- und Literaturwissenschaftlern) präsentiert.

Natürlich gilt auch für *Text*wörterbücher, dass man nicht alles in einem derartigen korpusorientierten Wörterbuch tun soll, was man *mit seiner Hilfe* machen kann. Sicherlich lassen sich bei einer Textaufbereitung nicht alle Fragen vorhersehen, die zu einem Textelement oder -ausschnitt einmal gestellt werden können.

Andererseits ist es noch keine Leistung, den Computer als *Ersatz* für einen Karteikasten entdeckt zu haben: Kaeding[6] hat nahezu 11 Millionen laufende Wörter *ohne Computer* bearbeitet. Sein Material ist auch heute noch (man betrachte die Sprachstatistik von Meier[7] und die Untersuchungen des Goethe-Instituts anhand des KaedingMaterials[8]) für manche ergänzende (intellektuelle wie maschinelle) Auswertung gut.

Die Ansprüche an den Computer zur Erschließung von Sprachdatensammlungen sollten also durchaus etwas höher angesetzt werden. Hierzu bietet sich besonders der *morphosyntaktische* Bereich an. Bereits 1970 wurden daher bei einer derartigen Verarbeitung - es handelte sich um einen Index zu den Werken des österreichischen Dichters Georg Trakl - höhere Anforderungen an ein Produkt der computer*gestützten* Textverarbeitung gestellt, als dies sonst für die literarischen Indices üblich war[9]. Im Interesse einer besseren Nutzung (vor allem durch den Literaturwissenschaftler) sollten dabei folgende Aufgaben erfüllt werden:

(1) Morphologische »Lemmatisierung« der Wortformen, d.h. ihre Reduktion auf Grundformen
(2) Zerlegung zusammengesetzter Wortformen in sinnvolle Segmente
(3) Ermittlung von Derivationen (vor allem Suffixen)
(4) Klassifikation der Belege nach syntaktischen Kategorien (Wortklassen)
(5) Semantische Disambiguierung

*Zu 1 (Lemmatisierung)*

Unter Lemmatisierung im weitesten Sinne wird allgemein eine Zusammenfassung von Wortformen unter einem sie repräsentierenden Stichwort verstanden, die morphologisch, syntaktisch und semantisch auf gleiche Merkmale (z.B. ein Stammmorphem, eine Wortklasse und gleiche Bedeutungsmerkmale) zurückzuführen sind[10]. Diese Definition würde bei streng formaler Anwendung in einigen Fällen (z.B. bei WAR und BIST, also morphologisch divergenten Formen von SEIN) zu Problemen führen; andererseits lassen sich bislang im semantischen Bereich nicht immer Merkmale für eine ausreichende Bedeutungsdifferenzierung ermitteln, die jeder Überprüfung standhalten; schließlich sind selbst im syntaktischen Bereich einerseits Wortklassendifferenzierungen nicht immer ausreichend empirisch (z.B. über Distributionsanalysen) abzusichern, andererseits kann eine Zusammenfassung oder In-Beziehung-Setzen von Wortformen über eine (Haupt-)Wortklasse hinaus für die Anwendung von Nutzen sein (z.B.: SINGEN - GESANG; LIEBE - LIEBLING/LIEB). Die Differenzierung bzw. Zusammenfassung stellt daher in einer bestimmten Ausprägung - etwa einem »Lemmatisierten Index« in Buchform - immer eine von mehreren möglichen »Sichten« auf das zugrundeliegende Material dar, meist unter dem Aspekt einer möglichst günstigen (Be-)Nutzbarkeit.

*Zu 2 (Kompositumzerlegung)*

Im Deutschen stellen Komposita in Texten eine sehr häufige Form dar. Es handelt sich dabei vielfach um nichts weniger als eine bestimmte Schreibgewohnheit (d.h. der Zusammenschreibung), wie Alternativformen zu diesen sog. »Augenblickskomposita« (AMERIKAREISE - REISE NACH AMERIKA, MINISTERBESUCH - BESUCH DES MINISTERS) zeigen. Damit sollen nicht die unterschiedlichen Verknüpfungsmöglichkeiten im Kontext übersehen werden. Andererseits zeigen mehrwortige Ausdrücke wie JURISTISCHE PERSON, die analog als »Bedeutungseinheit« betrachtet werden können (vgl. die engl. Übersetzung mit LEGAL ENTITY), dass die Zusammenschreibung allein kein Indiz für eine Bedeutungsänderung ist und umgekehrt.

Insofern ist es im Interesse des Benutzers sinnvoll, zumindest in allen Fällen, in denen Kompositumelemente noch eine gewisse Selbständigkeit bewahrt haben - und da die Grenze in einer Reihe von Fällen fliessend ist: vielleicht bereits bei eher formalen Zerlegungsmöglichkeiten -, zusammengesetzte Wörter in relevant erscheinende Segmente zu zerlegen und dies in geeigneter Form für eine Darstellung in einem Index oder Register heranzuziehen.

*Zu 3 (Derivationen)*

In vielen Sprachen liegen (meist morphologisch gekennzeichnete) Wortableitungen vor, die es sinnvoll erscheinen lassen, die entsprechenden Derivations*varianten* zu einem Kernelement (etwa einem Morphem) miteinander in Beziehung zu setzen. Umgekehrt kann man die Derivationselemente selbst (vor allem die Suffixe) mit den belegten Basiselementen verknüpfen, um Auswertungen über Häufigkeit, Typus usf. zu einem allgemeinsprachlichen Korpus oder einem autorbezogenen Textmaterial zu erleichtern.

Die in (2) und (3) dargestellten Zielsetzungen haben dazu geführt, dass in den Anfängen der Korpusforschung einfache *rückläufige* (Wortformen-)Listen erstellt wurden, die auf derartige Fragen mehr oder weniger vollständig Auskunft geben.

*Zu 4 (Wortklassenmarkierung)*

Die Markierung von Wortklassenangaben bei einer Textsammlung stellt einmal ein wichtiges Instrument zur Trennung der eher *funktionalen* Elemente einer Sprache (Funktionswörter, Partikel wie: Konjunktionen, Artikel, Präpositionen) von den *sinntragenden* Elementen (Substantive, Adjektive, Verben, Adverbien) dar. Dies kann es z.B. erlauben, sich bei Zeilen- oder Satzkonkordanzen auf diese »sinntragenden« Wörter zu beschränken, da diese in der Regel Gegenstand breiterer Forschungen sein werden.

Zugleich erlaubt eine stärker differenzierte Wortklassenmarkierung in manchen Fällen eine (partielle) *semantische* Disambiguierung von Wortformen, sofern sie einhergeht mit den unterschiedlichen Wortklassen (z.B. KREUZE (Verb) KREUZEN gegenüber KREUZE (Substantiv) KREUZ; WAGEN (Verb) gegenüber WAGEN (Substantiv); LAUTE (Adjektiv) LAUT gegenüber LAUTE (Verb) LAUTEN; LAUTEN oder LAUTE (Substantiv) (die) LAUTE/ (der) LAUT...).

Eine Wortklassenmarkierung kann schließlich die Grundlage bilden für weitergehende Auswertung des Textmaterials, etwa zur Häufigkeit des Gebrauchs des bestimmten Artikels gegenüber dem unbestimmten, zur Struktur der verwendeten Nominalgruppen usf.

*Zu 5 (Semantische Disambiguierung)*

Bei heterogenen Texten unterschiedlicher Themenstellung und bei größeren Textsammlungen ist eine Unterscheidung der Bedeutungen von Textwörtern bei auftretenden Mehrdeutigkeiten schon aus Gründen der praktischen Nutzbarkeit sinnvoll.

Dies führt zu zwei wesentlichen Problemen:

- Wie weit sollen »Bedeutungen« unterschieden werden? Meist gibt es noch ein Teilproblem hierzu, da neben (völlig?) unterschiedlichen Bedeutungen noch Bedeutungsfamilien und -hierarchien konstatiert werden können. Auf diese Weise können Textsammlungen - bei unterschiedlicher Bedeutungsdifferenzierung - leicht nicht mehr vergleichbar werden. Zumindest sollte in solchen Fällen ein allgemeiner Maßstab (oder einfach ein Bezugspunkt), z.B. ein Verweis auf ein bestimmtes Lexikon angegeben werden, der einem Benutzer die Differenzierung nachzuvollziehen erlaubt, und ein formaler Zugang über das undifferenzierte Stichwort möglich sein, wie es ja auch die traditionellen Lexika vorsehen.

- Wie weit lässt sich *für eine bestimmte Belegstelle* die »richtige« Bedeutung ermitteln? Oft reicht der Kontext (etwa eines Zeitungsartikels oder gar eines Gedichts) nicht aus, und leicht gerät man hier - wenn auch wohl in seltenen Fällen - ins Interpretieren (das man eigentlich dem Benutzer des Produkts überlassen wollte). Wenn also eine Bedeutungsdiffe-

renzierung vorgenommen wird, so müssten Zweifelsfälle möglichst als solche (zusätzlich) erkennbar bleiben.

Im folgenden soll zunächst am Beispiel der Erstellung des lemmatisierten Index zu dem Gesamtwerk des österreichischen Dichters Georg Trakl verdeutlicht werden, wie man bereits mit einfachster Computerunterstützung auf ökonomische Weise i.S. der vorgestellten Ziele zu ansprechenden Ergebnissen kommen kann:

Bei der Herstellung des lemmatisierten Index zu dem Werk von Georg Trakl[11] wurde der Computer - da zu diesem Zeitpunkt keine weitergehenden maschinellen Verfahren vorlagen - wie üblich als reines Sortier- und Speicherinstrument verwendet[12].

(1) Zunächst wurde der Text des Gesamtwerkes (über 5-Kanal-Lochstreifen) fortlaufend erfasst. Dabei wurde jede Seite der Kritischen Ausgabe[13] entsprechend gekennzeichnet und der Text selbst nach der Zeilenzählung der Herausgeber strukturiert. Hieraus ließen sich später beim Index als Verweis die Seiten- und Zeilenangaben erzeugen.

(2) *Nur* bei *Substantiven* wurde die *Wortklasse* (über eine Art Großschreibungsmerkmal) *explizit* bei der Erfassung festgehalten; trat eine *Mehrdeutigkeit* Funktionswort/Verb/Adjektiv auf, wurde ebenfalls bereits bei der Erfassung ein Differenzierungszeichen gesetzt.

(3) Nach der Erfassung wurde maschinell eine (alphabetisch sortierte) *Wortformenliste* erzeugt und auf Lochkarten ausgestanzt; die Belegstellen wurden auf Belegstellenkarten ausgegeben, die durch eine Identifikationsnummer mit der zugehörigen Wortformen(loch)karte (=Wortlautkarte) verknüpft waren.

(4) Über eine Sortiermaschine wurden *mechanisch*
- die Wortlautkarten von den reinen Belegstellenkarten getrennt
- die Substantive von den übrigen Wortformen separiert (hierzu war eine entsprechende Kennung vorhanden).

(5) Die Wortlautkarten wurden nun über einen automatischen Beschrifter »bedruckt«. Damit war der Inhalt »lesbar«. Diese Wortformenkarten - außer den Substantiven - wurden nun *intellektuell* nach 3 Gruppen getrennt: Partikel (»P«, d.h. Funktionswörter wie DER, EIN, ODER...), flektierte und unflektierte Adjektive (»A«) und Verben (»V«, auch die Partizipien umfassend). Anschließend wurden diese drei Karten*stapel* »automatisch« über einen Kartenlocher mit einer entsprechenden Wortklassenkennung versehen.

(6) Die morphologische Lemmatisierung (Grundformenzuordnung) sollte wortklassenbezogen erfolgen (nur die »Partikel« blieben als Wortformen erhalten). Zu diesem Zweck wurden alle Wortformen, bei denen der Wortlaut gleich der Grundform war, manuell herausgezogen und anschließend automatisch mit der (identischen) Grundform in dem (zuvor freigelassenen) Grundformenfeld ausgestattet. (Dies diente später zu Sortier- und Kontrollzwecken.) Die übrigen wurden »per Hand« mit der entsprechenden Grundform versehen (dies war für ca. 2.000 verschiedene Wörter je einmal durchzuführen).

(7) Die so ermittelten *unterschiedlichen Grundformen* (Lemmata) wurden über den Rechner rückläufig sortiert. Intellektuell wurden nun die Kompositumteile und Derivationselemente für die durch einfaches Vergleichen hierbei festgestellten (und sinnvoll erscheinenden) Einträge auf spezifischen *Verweiselementkarten* (Simplex-Kompositum) nachgetragen; ähnlich wurde bei Suffixen verfahren.

(8) Die Ausgangsdaten wurden nun in der Ordnung: Grundform-Wortform-Belegstellen über eine Sortiermaschine zusammengemischt und-sortiert, anschließend wurden die Verweiskarten intellektuell/manuell hinzugeordnet.

(9) Mithilfe eines Kontroll- und (Seiten-)Umbruch*programms* wurden anschließend das äußere Druckbild (Seitenform) für den *Index* erstellt und zugleich die Steuerzeichen erzeugt, die es erlaubten, den auf Magnetband zwischengespeicherten »Output« per Lichtsatz auf einen Film für die Buchvorlage zu bringen (vgl. Abb. 1).

(10) Über ein weiteres maschinelles (Sortier-)Programm wurde, aufbauend auf den Gesamtdaten, ein *Frequenzwörterbuch* erzeugt, dessen Grundlage die *Grundformen* bildeten. Die Berücksichtigung der Wortklassenangaben im Druckbild veranschaulichte - sinnbildlich sich in den häufigsten Adjektiven äußernd: - die »Welt« von Georg Trakl und seiner Dichtungsepoche: dunkel, blau, schwarz, leise, still... (vgl. Abb. 2).

Abgesehen von der Texterfassung und -kontrolle sowie der Programmerstellung wurden die Phasen (3) bis (10) für die gesamten Daten mit einem Zeitaufwand von ca. 1 Mannmonat bewältigt - einschließlich des diesbezüglichen Erfassungsaufwands. Damit wurde deutlich, dass ein entsprechendes Arbeitskonzept auch bei einfachsten Randbedingungen zu ökonomisch vertretbaren und zugleich ansprechenden Ergebnissen führen kann.

Seit dieser Arbeit sind inzwischen über 10 Jahre vergangen. An die Stelle der unhandlichen Lochstreifen und Lochkarten mit ihrem stark eingeschränkten Zeichenvorrat sind Terminals mit Groß- und Kleinschreibung getreten, an einigen Forschungsinstituten - so an der Universität des Saarlandes - stehen auch zumindest im Modell anwendbare *elektronische* Verfahren zur Verfügung, die es ermöglichen, den Computer nicht mehr nur als *Sortier*instrument einzusetzen, sondern auch seine »Intelligenz« zur Lösung der Fragen zu benutzen, die oben vorgestellt wurden. Ein solch »intelligentes« Basissystem stellt das »Saarbrücker Verfahren zur automatischen Textanalyse« dar. Es ist u.a. funktional darauf ausgerichtet, zu (nahezu) beliebigen deutschsprachigen Texten (besser: Sätzen) insbesondere morphologisch, syntaktisch und semantisch differenzierte Wörter (Lemmata) zu ermitteln[14]. Bereits in den 60er Jahren waren unter der Leitung von Hans Eggers hierzu verschiedene Forschungen durchgeführt worden[15].

Inzwischen liegt eine erste praktisch brauchbare Version des Systems vor, die zu anwendungsorientierten Untersuchungen genutzt werden kann. Allerdings wurde es nicht für sinnvoll gehalten, diese entwickelten Funktionen gleich an dem »Extremfall« der Belletristik oder Dichtung zu erproben. Günstiger erschien es, hierzu Texte der Gebrauchsprosa zugrundezulegen. Die diesbezüglichen Arbeiten können dabei auch auf einen größeren *praktischen* Nutzen ausgerichtet werden, als dies bei der Aufbereitung eines Dichterwerkes normalerweise der Fall ist.

Ein solcher praktischer Nutzen ergibt sich vor allem für den Bereich der (Fach-) Information und Dokumentation (IuD). Dabei geht es - grob gesagt - darum, fachlich relevante *Dokumente* (z.B. Zeitschriftenartikel, Urteilstexte, Verordnungen, Protokolle, Patentbeschreibungen) zumindest so weit zu erschließen, dass sie über darin vorkommende Begriffe »wiedergefunden« werden können. Es handelt sich also um eine Art der »Automatischen Indexierung«.

Da davon auszugehen ist, dass die gespeicherten Datenmengen sehr umfangreich sein und gegebenenfalls in einer Daten- oder Informationsbank auch heterogene Texte/Dokumente gespeichert

sein werden, ist letztlich das Problem der *Bedeutungsdifferenzierung* mitzubehandeln. (Dies konnte bei der Auswertung der Werke von G. Trakl noch ausgeklammert werden.)

Das hierzu aufbauend auf dem Saarbrücker System entwickelte Verfahren zur automatischen Indexierung (das in seiner modellhaften Ausrichtung auf *ju*ristische *Do*kumentation als Forschungsprojekt den Namen »JUDO« trägt) umfasst daher *alle* obengenannten Teilaspekte, also:

- Morphosyntaktische Lemmatisierung
- Dekomposition
- Derivation
- Wortklassenermittlung
- Semantische Disambiguierung.

In Teilbereichen geht es dabei über die oben genannten Fragestellungen noch hinaus. So werden z.B. auch *mehrwortige* Begriffe (wie z.B. JURISTISCHE PERSON) identifiziert; zwischen den vereindeutigten bzw. lexikalisch eindeutigen Elementen werden zudem eine Reihe von semantischen Relationen (Synonyme, Ober/Unterbegriff usf.) hergestellt, die vor allem bei dem späteren Retrieval, also dem Auffinden von Belegstellen, nützlich sind; schließlich werden auch im Text auftretende *syntaktische* Relationen (wie Adjektiv-Substantiv, Substantiv und angereihtes Substantiv) in unmittelbarer Verknüpfung (als sog. Komplexe Deskriptoren) für das Retrieval verfügbar gemacht.

An dieser Stelle soll nur kurz auf die hierbei zugrundegelegte Verfahrensweise eingegangen werden[16]:

(1) Ein (Fach-)Text wird nach dem Einlesen über eine normierte Input-Schnittstelle mit einem morphosyntaktischen Wörterbuch abgeglichen. Dabei werden einer Wortform verschiedene, für die weitere (syntaktische) Analyse nötige Informationen zugeordnet, zugleich wird die (mögliche) Grundform ermittelt. Bei Wortzusammensetzungen und Derivationen, die nicht in dem allgemeinen morphosyntaktischen Lexikon belegt sind, wird eine Dekompositions- bzw. Derivationsanalyse durchgeführt. Auf diese Weise lassen sich zugleich alle formalen Schreibfehler ermitteln: Da das zugrundegelegte Wörterbuch zum Deutschen inzwischen mehr als 90.000 Grundformen repräsentiert, ist mittlerweile die Wahrscheinlichkeit groß, dass es sich bei nicht-identifizierbaren Elementen um Schreibfehler handelt (vgl. Abb 3a und 3b).

(2) Der Lexikonabgleich aus (1) erbringt die *potentiellen* Wortklassen und Grundformen. Aufgrund einer (satz-)kontext-orientierten maschinellen *syntaktischen* Analyse werden nunmehr in verschiedenen Schritten die relevanten Funktionen ermittelt. Dieses Teilverfahren erfüllt die Funktion der Wortklassenermittlung sowie der morphologischen Lemmatisierung, insofern als die *potentiellen* Angaben aus dem Lexikon auf die im Kontext *aktuellen* reduziert werden (vgl. Abb. 4).

(3) In einem weiteren Schritt werden auf der Grundlage eines *semantischen* Lexikons, das u.a. Regeln zur Ermittlung mehrwortiger (flektierter) Begriffe sowie Merkmale und Regeln zur Disambiguierung enthält, derartige Mehrwortausdrücke identifiziert sowie semantische Vereindeutigungen vorgenommen, soweit dies aufgrund des Kontexts (auf Satzebene) möglich ist (vgl. Abb. 5).

(4) In einem vorläufig letzten Schritt wird versucht, aufgrund von statistischen Angaben (Wahrscheinlichkeiten für das Auftreten einer Bedeutung in einem engeren Fachgebiet) unter Heranziehung des *satzübergreifenden* Kontexts und unter Nutzung der in einem Fachlexikon aufgeführten *semantischen* Relationen zwischen Fachbegriffen die Restmehrdeutigkeiten zu vereindeutigen.

(5) Den Abschluss bildet ein Verfahren, das die Informationen so aufbereitet, dass sie als Deskriptoren in eine Datenbank oder als Einträge in ein Register übernommen werden können (zur Deskriptorenerstellung vgl. Abb. 6; ein Beispiel für einen Datenbankzugriff bringt Abb. 7).

Auch wenn dieses System erst in Modellform entwickelt wurde, zeigte doch die Laboranwendung an einem fachgebietsorientierten Textmaterial von über 100.000 laufenden Textwörtern (im Bereich *Datenschutz*), dass die automatische Texterschließung praktikabel wird. Einfachere Verfahren wie die Erstellung von Wort*formen*indices, entsprechende Häufigkeitslisten oder KWIC / KWOC-Konkordanzen erscheinen den heutigen Möglichkeiten nicht mehr angemessen. Die Verfahren zur Bearbeitung von Textkorpora sind somit in eine Phase getreten, die man durchaus als die einer »2. Generation« der Textkorpuserschließung ansprechen kann. Bei jeder Korpus-Arbeit sollte man sich daher - sofern nicht ad-hoc eine Verarbeitung kurzfristig nötig erscheint - zunächst darauf besinnen, das »Werkzeug« Computer durch die Entwicklung von Lexika und Regeln zur Sprachdatenverarbeitung so weit zu verbessern, dass ähnliche Verfahren verwendet werden können.

Anmerkungen mit Literatur

* Erweiterte Fassung eines Vortrags auf dem Symposium »Computercorpus der serbokroatischen Sprache«, Belgrad, 14. - 18. 12. 1981.

1. Reihe »Indices zur deutschen Literatur«, Hrsg. v. H. SCHWERTE und H. SCHANZE. Athenäum, Frankfurt.

2. Zum Begriff des Korpus (oder Corpus, hier variiert die Schreibweise selbst bei einem Autor) vgl. die verschiedenen Beiträge in dem Sammelband: Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. Hrsg.: H. BERGENHOLTZ, B. SCHAEDER. Königstein/Ts., 1979. Darin auch der Artikel von B. RIEGER: Repräsentativität. Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. (S. 52-70).

3. Microfiche-Ausgabe: MCS-Verlag, Nürnberg 1979. Reihe: Regensburger Materialien auf Microfiche (RMM). Zum Aufbau des LIMAS-Korpus vgl. R. GLAS: Das LIMAS-Korpus, ein Textkorpus für die deutsche Gegenwartssprache. In: Linguistische Berichte 40 (1975) S. 63-66.

4. Vgl. I. ROSENGREN: Ein Frequenzwörterbuch der modernen Zeitungssprache - wie und wozu? In: Beiträge zur Linguistik und Informationsverarbeitung 14 (1968) S. 7-21.

5. Zu den Korpora des Instituts für Deutsche Sprache vgl. B. SCHAEDER: Das Bonner Zeitungskorpus: Eine maschinelle Dokumentation von Tageszeitungen der BRD und der DDR. Mimeo. Bonn 1978, und U. ENGEL: Das Mannheimer Corpus. In: Forschungsberichte des Instituts für Deutsche Sprache 2, Mannheim 1969, S. 75-84.

6. F.W. KAEDING (Hrsg.): Häufigkeitswörterbuch der Deutschen Sprache. Festgestellt durch einen Arbeitsausschuss der deutschen Stenographiesysteme. Steglitz b. Berlin 1898. Eine aus-

führliche Beschreibung findet sich in W.D. ORTMANN (Hrsg.): Hochfrequente deutsche Wortformen I, München 1975, S. 5-26.

7. H. MEIER: Deutsche Sprachstatistik. Hildesheim 1964. Untersuchungen dieses »idealistischen Einzelgängers« (so ORTMANN, S. 27) fußen auf dem Kaeding-Material, wobei in mehr als 40 Jahren Freizeit-Arbeit das Kaeding-Material manuell aufbereitet und modifiziert wurde.

8. Die maschinellen Auswertungen des Kaeding-Materials durch die Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts unter der Leitung von W.D. ORTMANN haben inzwischen zu einer Serie von Veröffentlichungen geführt. Die Arbeiten dienen vor allem phonologischen Studien.

9. W. KLEIN, H. ZIMMERMANN: Index zu Georg Trakl. Dichtungen. Frankfurt 1971.

10. Zur formalen Definition von »Lemma« in diesem erweiterten Sinne vgl. R. DIETRICH: Automatische Textwörterbücher. Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen. (= Linguistische Arbeiten 2, Hrsg. v. H.E. BREKLE et al.) Tübingen 1973, bes. S. 1f. Dietrich fußt dabei auf der Definition von H.D. MAAS: Homographie und maschinelle Sprachübersetzung. In: Linguistische Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universität des Saarlandes, Nr. 8, Saarbrücken 1969.

11. Vgl. Anm. 9. Als technische Instrumente wurden herangezogen: Ein Rechner vom Typ Philips Electrologica X1; ein Rechner CDC 2200 und Lochkarten-Sortiermaschinen.

12. Da dieses Verfahren - wenn auch hardwaretechnisch inzwischen einige Veränderungen eingetreten sind - auch heute noch arbeitsökonomisch interessant erscheint, v.a. in Fällen (bzw. bei Sprachen), in denen (noch) keine tiefergehenden automatischen Analyseverfahren vorliegen, werden die einzelnen Schritte kurz vorgestellt.

13. Georg Trakl. Dichtungen und Briefe. Historisch-kritische Ausgabe. Hrsg. W. KILLY und H. SZKLENAR. Salzburg 1969 (2 Bde.). In den Index eingegangen sind daraus nur die Dichtungen, allerdings die Varianten eingeschlossen. Dies hat v.a. die Häufigkeitszählung beeinflusst, da unterschieden wurde zwischen Häufigkeiten einschließlich und ausschließlich der Varianten. Vgl. dazu das Vorwort zum Index.

14. Zur Beschreibung des Verfahrens vgl.: SALEM. Ein Verfahren zur automatischen Lemmatisierung deutscher Texte. Hrsg.: Sonderforschungsbereich 100 »Elektronische Sprachforschung«, Projektbereich A. Tübingen 1980.

15. Als wichtigstes Ergebnis dieser frühen Forschungen ist wohl zu nennen: H. EGGERS et al.: Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Tübingen 1969.

16. Das Verfahren ist beschrieben in: H. H. ZIMMERMANN: Ansätze einer realistischen automatischen Indexierung unter Verwendung linguistischer Verfahren. In: R. KUHLEN (Hrsg.): Datenbasen, Datenbanken, Netzwerke. Bd. 1 München 1979, S. 311-338 sowie: ders.: Bürgernahe Informationsvermittlung am Beispiel des Modellsystems »Juristische Dokumentanalyse im Bereich Datenschutz« (JUDO-DS). In: Österreichische Gesellschaft für Informatik (Hrsg.): Informationssysteme für die 80-er Jahre (Fachtagung 1980), Bd. 1, Linz 1980, S. 143-168.

**Abbildungen**

Abbildung 1: Alphabetischer Index zu Georg Trakls Dichtungen (Beispielseite 90):

lieb (9,8) A
liebe (3,2) 22,12 *350,66* *497,60*
Lieben (2,2) 107,12 342,19
lieber (2,2) 442,10 *444,18*
Lieber (1,1) 443,12
Liebes (1,1) 456,56
Liebe (26,15)
Liebe (26,15) 30,13 37,10
50,24 64,25 88,10 89,35 114,08
125,09 162,12 176,15 178,17 226,93
248,07 309,13 *333,06* *334,06* *337,09*
352,10 *363,22* *363,09* *392,15* *395,15*
*407,36* *410,49* *413,09* *413,13*
lieben (68,48) V
geliebt (1,1) A 253,06
geliebte (1,1) A 278,05
Geliebte (7,7) 194,35 194,45
195,16 195,17 195,21 198,09 255,11
geliebten (1,1) A 439,26
Geliebten (2,2) 193,27 196,58
Geliebtes (1,0) *346,10*
lieb (3,2) 49,16 *362,14* 444,19
lieben (2,2) 196,53 197,90
liebend (2,2) A 53,16 144,14
Liebende (9,6) 18,09 62,11
92,04 139,11 275,17 310,10 *368,11*
*404,53* *409,11*
liebenden (1,1) A 72,76
Liebenden (23,12) 45,24 58,35
57,21 80,02 85,06 199,17 119,21
143,04 159,07 288,48 313,12 329,05
*364,32* *368,16* *370,15* *372,05* *374,26*
*385,15* *388,10* *394,64* *398,26* *404,26*
*421,20*
liebender (1,1) A 137,12
Liebender (5,3) 29,03 34,11
*305,16* 306,05 *359,03*
Liebendes (4,2) 299,19 *343,04*
*344,04* 394,14
liebt (1,1) 311,01
liebte (3,3) 95,03 146,35 148,59
liebten (1,1) 447,05
*belieben*
*verlieben*
*vielgeliebt*
Liebesgeflüster (1,1)
Liebesgeflüster (1,1) 189,11
Liebeslallen (1,1)
Liebeslallen (1,1) 270,11
Liebesmär (1,1)
Liebesmär (1,1) 265,10
Liebesmahl (1,0)
Liebesmahl (1,0) *415,04*
Liebesnot (1,1)
Liebesnot (1,1) 248,05
Liebkosung (1,1)
Liebkosungen (1,1) 196,42
lieblich (2,2) A
liebliche (1,1) 194,44
lieblicher (1,1) 149,68
Lied (28,24)
Lied (19,15) 49,08 72,78 81,15
136,08 147,17 150,15 *204,44* 225,36
225,67 227,04 234,02 270,13 289,15
303,01 *325,05* *399,24* *402,26* *404,45*
442,03
Lieder (8,8) 54,06 *204,35* 235,83
235,04 *235,08* *235,09* 280,11 321,09
Liedern (1,1) 217,97
*Hirtenlied*
*Totenlied*
*Wiegenlied*
Liedlein (1,1)
Liedlein (1,1) 442,05
liegen (25,20) V
gelegen (1,1) A 191,71
lag (12,12) 88,10 88,18 147,13
148,39 148,47 168,18 169,51 170,54
266,07 267,03 267,12 272,07
lagen (2,1) 273,04 *367,29*
lagst (2,1) 322,04 *328,03*
liegen (1,1) 56,31
liegst (1,0) *328,06*
liegt (6,4) 12,18 14,73 51,02
189,27 *334,03* *368,17*
*abgelegen*
*Entlegenheit*
*erliegen*
*Gelage*
*gelegen*
Lilie (4,4)
Lilien (4,4) 66,14 269,45 316,17
445,13
*Wasserlilie*
lind (7,5) A
lind (5,4) 31,10 49,10 67,25
*283,09* 295,08 *363,09*
linden (1,1) 28,13
Linde (5,5)
Linde (1,1) 456,43
Linden (4,4) 190,54 190,60 271,80
274,05
Lindenbaum (2,2)
Lindenbäume (1,1) 189,07
Lindenbaum (1,1) 443,02
*-ling*
*Fremdling*
*Fremdlingin*
*Frühling*
*Jüngling*

Abbildung 2: Häufigkeitsindex zu Georg Trakls Dichtungen (Beispielseite 169):

| Nr. | Rang | Häufigkeit rel. | Häufigkeit abs. | Wortkl. | Lemma |
|---|---|---|---|---|---|
| 1 | 1 | 3,789 | 1240 | P | und |
| 2 | 2 | 3,673 | 1202 | P | in |
| 3 | 3 | 3,493 | 1143 | P | die |
| 4 | 4 | 3,056 | 1000 | P | der |
| 5 | 5 | 2,698 | 883 | P | ein |
| 6 | 6 | 1,295 | 424 | P | das |
| 7 | 7 | 1,248 | 408 | V | sein |
| 8 | 8 | 1,243 | 407 | P | an |
| 9 | 9 | 1,234 | 404 | P | des |
| 10 | 10 | 1,057 | 346 | P | den |
| 11 | 11 | 0,950 | 311 | P | von |
| 12 | 12 | 0,812 | 266 | P | ich |
| 13 | 13 | 0,696 | 228 | A | dunkel |
| 14 | 14 | 0,693 | 227 | P | sich |
| 15 | 15 | 0,684 | 224 | P | es |
| 16 | 16 | 0,678 | 222 | P | ihr |
| 17 | 17 | 0,620 | 203 | P | auf |
| 18 | 18 | 0,602 | 197 | P | o |
| 19 | 19 | 0,574 | 188 | P | zu |
| 20 | 20 | 0,540 | 177 | P | aus |
| 21 | | | 177 | P | du |
| 22 | 21 | 0,528 | 173 | P | sie |
| 23 | 22 | 0,516 | 169 | P | sein |
| 24 | 23 | 0,504 | 165 | P | mein |
| 25 | 24 | 0,501 | 164 | P | mit |
| 26 | 25 | 0,495 | 162 | S | Nacht |
| 27 | 26 | 0,492 | 161 | A | blau |
| 28 | 27 | 0,479 | 157 | A | schwarz |
| 29 | 28 | 0,476 | 156 | P | wie |
| 30 | 29 | 0,440 | 144 | P | dem |
| 31 | 30 | 0,430 | 141 | P | da |
| 32 | | | 141 | P | über |
| 33 | 31 | 0,424 | 139 | A | leise |
| 34 | 32 | 0,421 | 138 | S | Schatten |
| 35 | 33 | 0,382 | 125 | P | dein |
| 36 | 34 | 0,369 | 121 | P | durch |
| 37 | 35 | 0,330 | 108 | V | gehen |
| 38 | 36 | 0,317 | 104 | P | mich |
| 39 | | | 104 | V | schweigen |
| 40 | 37 | 0,314 | 103 | P | er |
| 41 | 38 | 0,305 | 100 | V | sehen |
| 42 | 39 | 0,293 | 96 | S | Abend |
| 43 | 40 | 0,290 | 95 | A | still |
| 44 | 41 | 0,275 | 90 | P | vor |
| 45 | 42 | 0,271 | 89 | A | alt |
| 46 | | | 89 | A | golden |
| 47 | 43 | 0,262 | 86 | A | weiss |
| 48 | 44 | 0,250 | 82 | S | Auge |
| 49 | | | 82 | A | rot |
| 50 | | | 82 | A | sanft |
| 51 | 45 | 0,244 | 80 | P | dies |
| 52 | 46 | 0,241 | 79 | P | mir |
| 59 | 49 | 0,229 | 75 | S | Hand |
| 60 | | | 75 | S | Herz |
| 61 | 50 | 0,223 | 73 | P | nicht |
| 62 | 51 | 0,213 | 70 | S | Gott |
| 63 | 52 | 0,210 | 69 | A | purpurn |
| 64 | 53 | 0,207 | 68 | A | grün |
| 65 | | | 68 | V | werden |
| 66 | 54 | 0,204 | 67 | A | braun |
| 67 | | | 67 | V | verfallen |
| 68 | | | 67 | P | wenn |
| 69 | 55 | 0,195 | 64 | V | singen |
| 70 | 56 | 0,192 | 63 | P | nach |
| 71 | 57 | 0,186 | 61 | A | silbern |
| 72 | | | 61 | S | Stirn |
| 73 | 58 | 0,183 | 60 | S | Antlitz |
| 74 | | | 60 | S | Baum |
| 75 | | | 60 | S | Blut |
| 76 | | | 60 | S | Garten |
| 77 | 59 | 0,180 | 59 | S | Stille |
| 78 | | | 59 | V | tönen |
| 79 | | | 59 | S | Wein |
| 80 | | | 59 | P | wir |
| 81 | 60 | 0,177 | 58 | V | sinken |
| 82 | | | 58 | V | treten |
| 83 | 61 | 0,171 | 56 | A | kühl |
| 84 | 62 | 0,168 | 55 | A | einsam |
| 85 | | | 55 | V | fallen |
| 86 | | | 55 | A | wild |
| 87 | 63 | 0,165 | 54 | A | tot |
| 88 | 64 | 0,161 | 53 | S | Fenster |
| 89 | | | 53 | P | noch |
| 90 | | | 53 | V | stehen |
| 91 | | | 53 | P | voll |
| 92 | | | 53 | S | Wind |
| 93 | 65 | 0,158 | 52 | P | um |
| 94 | | | 52 | S | Wolke |
| 95 | 66 | 0,155 | 51 | P | dich |
| 96 | | | 51 | V | kommen |
| 97 | 67 | 0,152 | 50 | V | dämmern |
| 98 | | | 50 | P | jen |
| 99 | | | 50 | S | Leben |
| 100 | 68 | 0,149 | 49 | P | aber |
| 101 | | | 49 | P | als |
| 102 | | | 49 | S | Haupt |
| 103 | | | 49 | A | tief |
| 104 | 69 | 0,146 | 48 | S | Engel |
| 105 | | | 48 | A | lang |
| 106 | | | 48 | V | lieben |
| 107 | | | 48 | V | sterben |
| 108 | | | 48 | P | unser |
| 109 | 70 | 0,143 | 47 | A | fern |
| 110 | | | 47 | A | schön |

Abbildung 3a: Beispieltext (Original) zum Bundesdatenschutzgesetz (BDSG §33):

akademische Grade, die Anschrift sowie auf eine Angabe über die Zugehörigkeit des Betroffenen zu dieser Personengruppe beschränkt und kein Grund zur Annahme besteht, daß dadurch schutzwürdige Belange des Betroffenen beeinträchtigt werden.

§ 33

**Datenveränderung**

Das Verändern personenbezogener Daten ist zulässig, soweit dadurch schutzwürdige Belange des Betroffenen nicht beeinträchtigt werden.

§ 34

**Auskunft an den Betroffenen**

(1) Werden erstmals zur Person des Betroffenen

Abbildung 3b: Ergebnis der morphosyntaktischen Analyse:

| SNR | WNR | TEXTWORTFORM | WKL | LEMMANAME | STW |
|---|---|---|---|---|---|
| 2 | 1 | Das | REL | D- | FWK |
| 2 | 1 | | ARTB | D- (ARTB) | FWK |
| 2 | 1 | | PER | D- | FWK |
| 2 | 2 | Veraendern | SBI | VERAENDERN | VRB |
| 2 | 3 | personenbezogener | ADJ | PERSONENBEZOGEN | ADJ |
| 2 | 4 | Daten | SUB | DATUM | SUB |
| 2 | 5 | ist | FIV | SEIN (VRB) | VRB |
| 2 | 6 | zulaessig | ADV | ZULAESSIG | ADJ |
| 2 | 7 | , | | , | |
| 2 | 8 | soweit | UKO | SOWEIT | FWK |
| 2 | 9 | dadurch | ADV | DURCH D- | FWK |
| 2 | 10 | schutzwuerdige | ADJ | SCHUTZWUERDIG | ADJ |
| 2 | 11 | Belange | SUB | BELANG | SUB |
| 2 | 12 | des | ARTB | D- (ARTB) | FWK |
| 2 | 13 | Betroffenen | SUB | BETROFFENE | SUB |
| 2 | 13 | | SUB | BETROFFENER | SUB |
| 2 | 13 | | SBA | BETREFFEN | VRB |
| 2 | 13 | | SBA | BETROFFEN | ADJ |
| 2 | 14 | nicht | ADV | NICHT | FWK |
| 2 | 15 | beeintraechtigt | ADP | BEEINTRAECHTIGEN | VRB |
| 2 | 15 | | PTZ2 | BEEINTRAECHTIGEN | VRB |
| 2 | 15 | | FIV | BEEINTRAECHTIGEN | VRB |
| 2 | 16 | werden | INF | WERDEN | VRB |
| 2 | 16 | | FIV | WERDEN | VRB |

Abbildung 4: Wortklassenbestimmung durch syntaktische Analyse:

| SNR | WNR | TEXTWORTFORM | WKL | LEMMANAME | STW FS BEDEUTU |
|---|---|---|---|---|---|
| 2 | 1 | Das | ARTB | D- (ARTB) | FWK |
| 2 | 2 | Veraendern | SBI | VERAENDERN | VRB |
| 2 | 3 | personenbezogener | ADJ | PERSONENBEZOGEN | ADJ |
| 2 | 4 | Daten | SUB | DATUM | SUB |
| 2 | 5 | ist | FIV | SEIN (VRB) | VRB |
| 2 | 6 | zulaessig | ADV | ZULAESSIG | ADJ |
| 2 | 7 | , | | , | |
| 2 | 8 | soweit | UKO | SOWEIT | FWK |
| 2 | 9 | dadurch | ADV | DURCH D- | FWK |
| 2 | 10 | schutzwuerdige | ADJ | SCHUTZWUERDIG | ADJ |
| 2 | 11 | Belange | SUB | BELANG | SUB |
| 2 | 12 | des | ARTB | D- (ARTB) | FWK |
| 2 | 13 | Betroffenen | SUB | BETROFFENER | SUB |
| 2 | 14 | nicht | ADV | NICHT | FWK |
| 2 | 15 | beeintraechtigt | PTZ2 | BEEINTRAECHTIGEN | VRB |
| 2 | 16 | werden | FIV | WERDEN | VRB |
| 2 | 17 | * | | * | |

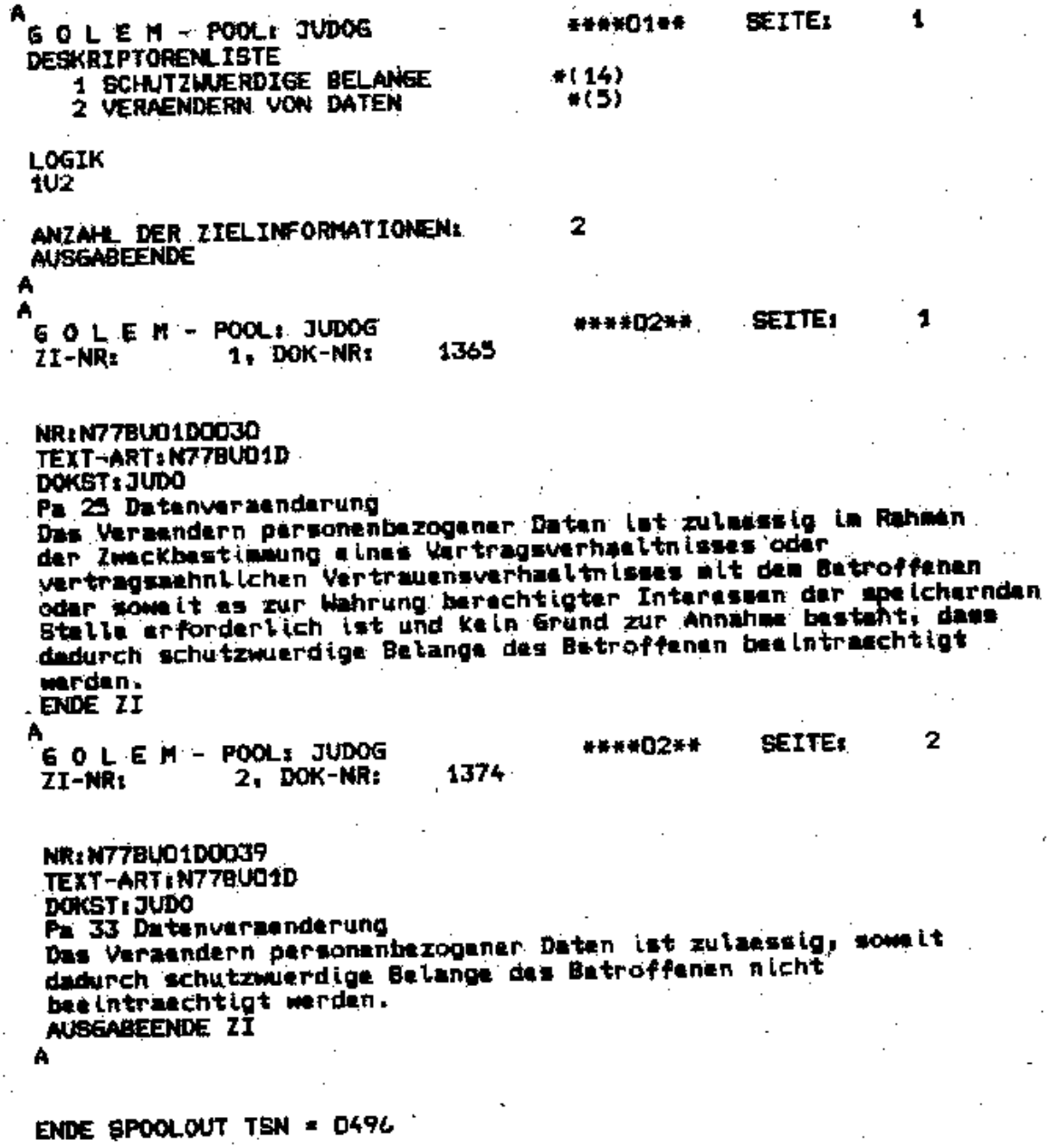
Abbildung 5: Semantische Analyse: Disambiguierung, Ermittlung von Mehrwortausdrücken:

| SNR | WNR | TEXTWORTFORM | WKL | LEMMANAME | STW | FS | BEDEUT |
|---|---|---|---|---|---|---|---|
| 2 | 1 | Das | ARTB | D- (ARTB) | FWK | | |
| 2 | 2 | Veraendern | SBI | VERAENDERN | VRB | FS | |
| 2 | 2 | | SBI | VERAENDERN PERSONENBEZOGENER DATEN | VRB | FS | |
| 2 | 2 | | SBI | PERSONENBEZOGENE DATEN | VRB | FS | |
| 2 | 2 | | SBI | VERAENDERN VON DATEN | VRB | FS | |
| 2 | 3 | personenbezogener | ADJ | PERSONENBEZOGEN | ADJ | FS | |
| 2 | 4 | Daten | SUB | DATUM | SUB | FS | |
| 2 | 5 | ist | FIV | SEIN (VRB) | VRB | | |
| 2 | 6 | zulaessig | ADV | ZULAESSIG | ADJ | | |
| 2 | 7 | , | | , | | | |
| 2 | 8 | soweit | UKO | SOWEIT | FWK | | |
| 2 | 9 | dadurch | ADV | DURCH D- | FWK | | |
| 2 | 10 | schutzwuerdige | ADJ | SCHUTZWUERDIG | ADJ | FS | |
| 2 | 11 | Belange | SUB | BELANG | SUB | FS | |
| 2 | 11 | | SUB | SCHUTZWUERDIGE BELANGE DES BETROFFEN | SUB | FS | |
| 2 | 11 | | SUB | BELANGE DES BETROFFENEN | SUB | FS | |
| 2 | 11 | | SUB | SCHUTZWUERDIGE BELANGE | SUB | FS | |
| 2 | 12 | des | ARTB | D- (ARTB) | FWK | | |
| 2 | 13 | Betroffenen | SUB | BETROFFENER | SUB | FS | |
| 2 | 14 | nicht | ADV | NICHT | FWK | | |
| 2 | 15 | beeintraechtigt | PTZ2 | BEEINTRAECHTIGEN | VRB | | |
| 2 | 16 | werden. | FIV | WERDEN | VRB | | |
| 2 | 17 | * | | * | | | |

Abbildung 6: Deskriptorenliste zum Beispieltext

```
SATZ      1
Datenveraenderung (

SATZ      2
Das Veraendern personenbezogener Daten ist zulaessig , soweit
dadurch schutzwuerdige Belange des Betroffenen nicht
beeintraechtigt werden *
STOP

ENDE DTVTEXT (7909.26)  0.18

START DESKRIPTOREN (8202.24)


  DESKRIPTOREN ZU SATZ      1
-------------------------------------------------------------
        DATENVERAENDERUNG
  TEIL: DATUM1
  TEIL: VERAENDERUNG1


  DESKRIPTOREN ZU SATZ      2
-------------------------------------------------------------
        BEEINTRAECHTIGEN
        BELANG
        BELANG1
        BELANG BEEINTRAECHTIGEN
        BELANG G BETROFFENER
        BELANGE DES BETROFFENEN
        BETROFFENER
  TEIL: BEZOGEN
        DATUM
        DATUM1
  TEIL: PERSON2
        PERSONENBEZOGEN
        PERSONENBEZOGENE DATEN
        PERSONENBEZOGENES DATUM
  TEIL: SCHUTZ1
        SCHUTZWUERDIG
        SCHUTZWUERDIGE BELANGE
        SCHUTZWUERDIGE BELANGE DES BETROFFEN
        SCHUTZWUERDIGER BELANG
        VERAENDERN
        VERAENDERN1
        VERAENDERN2
        VERAENDERN G DATUM
        VERAENDERN PERSONENBEZOGENER DATEN
        VERAENDERN VON DATEN
  TEIL: WUERDIG2
        ZULAESSIG
```

Abbildung 7: Beispiel für eine Recherche in der Datenbank:



```
A
 G O L E M - POOL: JUDOG                    ****01**    SEITE:     1
 DESKRIPTORENLISTE
    1 SCHUTZWUERDIGE BELANGE              #(14)
    2 VERAENDERN VON DATEN                 #(5)

 LOGIK
 1U2

 ANZAHL DER ZIELINFORMATIONEN:             2
 AUSGABEENDE
A
A
 G O L E M - POOL: JUDOG                    ****02**    SEITE:     1
 ZI-NR:       1, DOK-NR:     1365


 NR:N77BU01D0030
 TEXT-ART:N77BU01D
 DOKST:JUDO
 Pa 25 Datenveraenderung
 Das Veraendern personenbezogener Daten ist zulaessig im Rahmen
 der Zweckbestimmung eines Vertragsverhaeltnisses oder
 vertragsaehnlichen Vertrauensverhaeltnisses mit dem Betroffenen
 oder soweit es zur Wahrung berechtigter Interessen der speichernden
 Stelle erforderlich ist und kein Grund zur Annahme besteht, dass
 dadurch schutzwuerdige Belange des Betroffenen beeintraechtigt
 werden.
 ENDE ZI
A
 G O L E M - POOL: JUDOG                    ****02**    SEITE:     2
 ZI-NR:       2, DOK-NR:     1374


 NR:N77BU01D0039
 TEXT-ART:N77BU01D
 DOKST:JUDO
 Pa 33 Datenveraenderung
 Das Veraendern personenbezogener Daten ist zulaessig, soweit
 dadurch schutzwuerdige Belange des Betroffenen nicht
 beeintraechtigt werden.
 AUSGABEENDE ZI
A


 ENDE SPOOLOUT TSN = 0496
```