Der General Inquirer III: ein Dinosaurier mit Zukunft

Züll, Cornelia; Mohler, Peter Ph.

Der General Inquirer III
Ein Dinosaurier mit Zukunft

von Cornelia Züll und Peter Ph. Mohler

1.0 Einleitung

Der General Inquirer wurde zu Anfang der sechziger Jahre von Philip J. Stone als inhaltsanalytisches Pendant zu Herbert Simons "General Problem Solver" entwickelt. Er entstand damals in engem Zusammenhang mit Entwicklungen zur maschinellen Übersetzung. Dementsprechend wurde zusammen mit den Programmen zugleich an der Entwicklung von Wörterbüchern (Diktionären) gearbeitet, mit deren Hilfe die gesellschaftliche Wirklichkeit vollautomatisch in sozialwissenschaftliche Begrifflichkeit übersetzt werden sollte. Auch wenn sich in der Folgezeit dieser hohe Anspruch nicht ganz hat einlösen lassen, so blieb der General Inquirer das einzige Programm für die computerunterstützte Inhaltsanalyse (cui), zu dem es große, auch in historischen Studien bewährte, Klassifikationssysteme gibt. In diesem Beitrag werden vor allem diese Klassifikationssysteme und die Programme erläutert. Daneben wird auch das Ziel verfolgt, dem Leser einen ersten Überblick über diesen komplexen textanalytischen Ansatz zu verschaffen.

2.0 Abriß der Entwicklung des General Inquirers

Eine erste größere Publikation zum General Inquirer erschien bereits 1966[1]. In dieser ersten, grundlegenden Publikation wurde der theoretische Hintergrund ausführlich dargestellt, insbesondere wurde der Begriff der Inhaltsanalyse schärfer als eine Methode der Inferenz von Texten auf soziale Wirklichkeiten definiert. Einen großen Teil des Buches nahmen aber schon Berichte zu Anwendungen des General Inquirers in der praktischen Forschung ein. Diese Anwendungen streuten über alle Gebiete der Humanwissenschaften, von der Psychologie über die Soziologie bis zur Geschichtswissenschaft.

Obwohl nach dieser ersten Publikation weiter am General Inquirer entwickelt wurde - es gab neben der Harvard Version noch eine Edinburgh Version, die Diktionäre wurden weiter validiert und ausgebaut - stagnierte die

[1] Stone, P. J. , D. D. Dunphy, M. S. Smith, and D. M. Ogilvie (eds.), 1966. The General Inquirer: A Computer Approach to Content Analysis. Cambridge: MIT Press.

Entwicklung, weil es nicht gelang, eine stabile, gut dokumentierte Standardversion der wissenschaftlichen Öffentlichkeit zur Verfügung zu stellen. Die Anwendung des General Inquirers erforderte zudem auch gute programmtechnische Kenntnisse, was seiner Verbreitung in den Geisteswissenschaften nicht förderlich war.

Um 1976 schien es dann einen neuen Aufschwung zu geben, als Kelly und Stone ihr Verfahren zur Disambiguierung von Homographen publizierten und damit eines der schwierigsten Probleme der cui wenigstens für die häufigsten amerikanisch-englischen Wörter lösten[2]. Aber auch danach gelang es aus verschiedensten Gründen nicht, eine Standardversion zu entwickeln.

Trotz dieser äußerst ungünstigen Lage, flaute die Nachfrage nach dem General Inquirer bei Philip Stone nicht ab, sie stieg eher im Laufe der achtziger Jahre an. ZUMA, das schon zuvor mit TEXTPACK das erste voll portable cui-Programm entwickelt hatte, erklärte sich auf Grund der bestehenden Nachfrage und des großen analytischen Potentials dann 1986 bereit, zusammen mit Robert Philip Weber (Harvard) eine Standardversion für IBM Computer zu entwickeln. Weber hatte zuvor den General Inquirer in einer umfangreichen historischen Untersuchung[3] eingesetzt und für solche Zwecke adaptiert. Die Entwicklung der Standardversion ist jetzt faktisch abgeschlossen, eine ausführliche Dokumentation liegt vor[4] und das Programm wird seit Anfang 1989 an interessierte wissenschaftliche Institute ausgeliefert.

2.1 Der theoretische Ansatz des General Inquirers

Wie schon erwähnt stand am Anfang der Entwicklung die Idee, Texte in eine Metasprache der Sozialwissenschaften zu übersetzen. Heute würde man eher sagen, daß mit Hilfe des General Inquirers semantische Strukturen, Bedeutungen, eindeutig abgebildet werden sollen. Die Abbildung ist nicht als spezifische, womöglich nur auf eine Dimension eingeschränkte Abbildung, gedacht, sondern als eine allgemeine und mehrdimensionale. Mit "allgemein" ist gemeint, daß unterschiedlichste Forschungsansätze und Fragestellungen ein und dasselbe Klassifikationsverfahren benutzen. Damit soll eine Basis für

[2] Kelly. E. F., and P. J. Stone, 1975. Computer Recognition of English Word Senses. Amsterdam: North Holland.

[3] Namenwirth, Z. J., and R. P. Weber, 1987. The Dynamics of Culture. Boston: Allen & Unwin.

[4] Züll, C., R. P. Weber, and P. Ph. Mohler, 1989. Computer-aided Text Classification for the Social Science: The General Inquirer III. Mannheim: ZUMA.

interdisziplinäre, übergreifende und kumulative Forschung gelegt werden. Der General Inquirer ist von daher ganz bewußt auf vergleichende Forschung angelegt. Seine Klassifikationsschemata, die als Wörterbücher (Diktionäre) realisiert wurden, sind deshalb unter eher allgemeinen, denn spezifischen theoretischen Annahmen entwickelt worden (siehe 4.0). "Deduktiv entwickelt" heißt, daß die Einträge im Diktionär (Wörter) unabhängig von einer spezifischen Fragestellung allein aus theoretische Erwägungen jeweils einer oder mehreren Klassen (Kategorien) zugewiesen wurden. Dieser sogenannte a priori Ansatz der cui ist in der Literatur nicht unumstritten[5]. Trotz aller gegen den a priori Ansatz vorgebrachten Einwände muß man zugestehen, daß er, richtig eingesetzt, das Instrument der Wahl für groß angelegte vergleichende Studien ist.

2.2 Der General Inquirer in der historischen Sozialforschung

Von Anfang an wurde der General Inquirer zur Analyse historischer Texte und zur Beantwortung historischer Fragestellungen eingesetzt (vgl. diverse Beiträge in Stone et al. 1966). Die bislang bekannteste Untersuchung auf dem Gebiet der historischen Sozialforschung geht auf Zvi Namenwirth zurück, der mit einer systematischen Analyse der amerikanischen "Party Platforms" über 120 Jahre hinweg (1844-1964) zyklische Bewegungen der gesellschaftlichen Entwicklung nachweisen konnte. Robert P. Weber konnte die von Namenwirth postulierten Zyklen auch in Großbritannien an Hand von Thronreden (1689- 1974) nachweisen. In "Dynamics of Culture" (Namenwirth & Weber, 1987) haben beide ihre Ergebnisse nach theoretischen Grundlagen stringend und überzeugend zusammengefaßt.

Beide Studien weisen auf die Stärke des General Inquirers, der vorzüglich für die Analyse bekannter Klassen historischer Texte geeignet ist. Wenn erst einmal durch geeignete Vorstudien die Struktur einer Textklasse in bestimmten historischen Kontexten herausgearbeitet ist, kann mit dem a priori Ansatz des General Inquirers eine große Textmenge valide und zuverlässig klassifiziert werden. Zusätzlich erlaubt die allgemeine, mehrdimensionale Abbildung des General Inquirers einen Vergleich zwischen unterschiedlichen Textklassen. So könnte man sich einen Vergleich von Webers Analysen der Thronreden mit Parlamentsdebatten oder einen Vergleich von Testa-

[5] Mohler, Peter Ph. 1989. "Computergestützte Inhaltsanalyse: Überblick über die linguistischen Leistungen", in:Lenders, W. et al. (Hg.). Handbook of Computational Linguistics. Berlin: de Gruyter.

menten mit Gesetzestexten vorstellen. Der General Inquirer Ansatz könnte also wesentlich zur Systematisierung in der historischen Sozialforschung betragen.[6]

3.0 Wörterbücher im General Inquirer

Die Wörterbücher des General Inquirers enthalten neben den Wörtereinträgen und den dazugehörigen Kategorien zusätzlich Regeln zur Disambiguierung (d.h. Erkennung von Homographen) der Wörter und zum Bestimmen von Redewendungen (z.B. "ship of state" und "Secretary of State"). Im Moment stehen drei große Wörterbücher, die zusammen mit dem General Inquirer weitergegeben werden, zur Verfügung. Diese drei Diktionäre - Lasswell Value, Harvard IV-3 und Harvard IV-4 - werden im Abschnitt 4 beschrieben.

Das Erstellen und Validieren von inhaltsanalytischen Diktionären ist im allgemeinen ein sehr aufwendiges Unterfangen. Trotzdem wird es häufig erforderlich sein, die sehr allgemein gehaltenen Diktionäre zu modifizieren und vor allem, der eigenen Fragestellung angepaßt, zu erweitern. Um Diktionäre zu erstellen, zu modifizieren und zu erweitern steht innerhalb des General Inquirers eine eigene Sprache zur Verfügung, in der alle Einträge und Regeln abgefaßt sein müssen.

Die Disambiguierung und Vercodung erfolgt auf Satzebene in drei Stufen: zunächst wird das Wort selbst überprüft, danach werden im selben Satz andere, spezifische Wörter gesucht und als letztes werden Kategorien überprüft, die anderen Wörtern im Satz bereits zugewiesen wurden. Danach kann die Bedeutung des Worts in den meisten Fällen bestimmt werden. Im Folgenden soll kurz auf den Aufbau der Wörterbücher und der Regeln im Wörterbuch eingegangen und die wesentlichen Merkmale aufgezeigt werden. Tabelle 3.0.1 zeigt am Beispiel des Wortes "LAST" den Aufbau eines Wörterbucheintrags.

[6] Die General Inquirer Diktionäre liegen als vollentwickeltes Klassifikationssystem zur Zeit nur für englische Texte vor. Manuel Eisner versucht im Moment eine deutschsprachige Version des Lasswell Value Diktionärs (LVD) zu entwickeln. Der a priori Ansatz ist zudem nicht an den General Inquirer gebunden, er kann auch mit anderen Systemen, z. B. der LVD-Lib oder TEXTPACK V realisiert werden.

Tabelle 3.0.1: Wörterbucheintrag zu "LAST"

```
LAST=TAGS: ( 1) DET.NUMB.ORD.MODIF.TIMESP.'ADJ "PREVIOUS,
               PAST, PRIOR, MOST RECENT '
           ( 2) DET.NUMB.ORD.POS.MODIF.TIMESP.'ADJ-ADV "FINAL,
               FINALLY '
           ( 3) LY.TIMESP.'IDIOM-ADV ""AT . . .," "AT LONG LAST"--
               FINALLY'
           ( 4) SUPV.ENDS.'VERB "ENDURE, REMAIN '
           ( 5) MODIF.UNDEF*.'ADJ ""LASTING"--ENDURING,
               REMAINING'

RULES:     ( 7) TOR(K+0,K+0,,27,ROOT.)
           ( 8) WOR(K+1,K+1,,12,TIME.FALL.)
           ( 9) WOR(K-I,K-I,,APLY(I),THE.)
           (10) TOR(K-2,K-2,APLY(2),APLY(1),BE.)
           (12) TOR(K-1,K-1,,22,DET.)
           (13) TOR(K+1,K+3,,17,T1ME.)
           (14) WOR(K+1,K+1,APLY(2),,DAY.MINUTE.MOMENT.
               SECOND.)
           (15) TOR(K-1,K-1,APLY(2),APLY(1),GEN.)
           (17) TORK(K-5,K+5,APLY(1),,POLIT.)
           (18) TORK(K-6,K+6,,APLY(2),CARD.)
           (19) WOR(K+1,K+1,APLY(2),APLY(1),ONE.)
           (22) WOR(K-1,K-1,DEL1D(3),,AT.LONG.)
           (23) TORK(K-I,K+I,APLY(I),,TIME.)
           (24) TOR(K-1,K-1,APLY(4),,TO.MOD.DO.LY.INDEF.)
           (25) TOR(K+1,K+1,APLY(4),APLY(2),DET.PREP.)
           (27) TOR(K+0,K+0,APLY(2),,LY.)
           (28) TOR(K+0,K+0,,APLY(4),ING.)
           (29) TOR(K+1,K+1,APLY(4),APLY(5),DET.PREP.)
END;
```

Jeder Eintrag im Wörterbuch besteht aus zwei Teilen: einer Liste aller möglichen Bedeutungen des Worts (TAGS, Bedeutungscodes), gefolgt von Regeln (RULES, formale Disambiguierungsregeln), mit denen das Wort der richtigen Kategorie zugeordnet werden kann.

Im ersten Teil werden jeweils alle möglichen Bedeutungen eines Worts aufgelistet und mit Bedeutungscodes versehen. Im obigen Beispiel hat das Wort LAST als Verb z.B. den Code 4. Nach dem jeweiligen Code werden alle Kategorien (syntaktische und semantische) aufgeführt, die dem Wort zugewiesen werden sollen, wenn es in dieser Bedeutung vorkommt.

Sind alle Bedeutungen definiert, folgen die Regeln für die Zuordnung der richtigen Bedeutung zu einem Wort. Ist das Wort immer eindeutig, entfällt dieser zweite Teil des Wörterbucheintrags. Die Regeln werden, wie vorher schon die verschiedenen Bedeutungen, in Form einer Liste angegeben.

Es gibt drei verschiedene Hilfen zum Testen: die GOTO Funktion, die SUPV-Funktion, die auf syntaktischer Ebene testet, und den normalen Kontext-Test, wie er unten beschrieben wird. Mit GOTO können Regeln und Tests übersprungen, bzw. Regeln gezielt angesteuert werden, mit SUPV kann festgestellt werden, ob es sich um eine Verb-Form handelt oder nicht und in Abhängigkeit davon zur nächsten Regel übergegangen werden. Die Kontext-Tests sind immer in der Form

test name(r-start,r-end,b-success,b-fail,items)

aufgebaut. Dabei definiert das erste Feld (test name) die Art der Suche. Es gibt die verschiedensten spezifischen Funktionen zu suchen, auf die hier nicht alle eingegangen werden kann. Die wichtigsten sind WOR, TOR, TSAME, WORK und TORK. WOR und WORK testen, ob die unter "items" als letztem Parameter angegebenen Wörter im definierten Bereich vorkommen oder nicht. Bei WORK ist das Schlüsselwort selbst vom Test ausgenommen, bei WOR dagegen wird es in den Test einbezogen. Bei TOR und TORK werden anstelle des Worts, die in der Itemliste angegebenen Codes mit den für das Wort (oder die Wörter) im angegebenen Bereich vergebenen Codes verglichen. TSAME prüft, ob dem Wort selbst schon ein bestimmter Code zugewiesen wurde.

Der Bereich, der überprüft wird, wird durch r-start und r-end in der oben beschriebenen Regel festgelegt. Die Angaben beziehen sich auf die Wortpositionen, die geprüft werden sollen. Ist r-start,r-end z.B. mit K+1,K+1 angegeben, so wird das Wort, das auf das gesuchte Wort folgt, getestet; ist K+5,K+5 angegeben, so werden die 5 Wörter vor und 5 Wörter nach dem gesuchten Wort in die Überprüfung miteinbezogen.

b-success legt fest, was geschehen soll, wenn die Bedingung erfüllt ist, b-fail entsprechend, was bei nicht-erfüllten Bedingungen geschehen soll. Dabei können als Ergebnis mit der Funktion APLY die gefundenen Codes zugewiesen werden. Es kann aber auch zu einer weiteren Regel gesprungen werden oder, wenn nichts angegeben ist, mit der nächsten Regel weitergetestet werden. Eine letzte Möglichkeit - die Funktion DELID - erlaubt die Behandlung von Idiomen oder idiomatischen-Redewendungen, d.h. es wird geprüft, ob ein Wort Teil einer Redewendung ist. Wenn ja, werden alle bereits zugewiesenen Codes wieder gelöscht und stattdessen der besondere Code für das Idiom zugewiesen. Im obigen Beispiel wird dies in der Regel 22 gemacht. Geprüft wird, ob es sich um AT LONG LAST handelt. In diesem Fall erhält das Wort den Code 3. Ist das nicht zutreffend, wird mit der nächsten Regel weitergetestet.

Im einzelnen läuft die Prüfung der Regeln für das Wort LAST folgendermaßen ab: zunächst wird geprüft, ob es sich bei dem Wort um den Wortstamm oder um eine andere Form (z.B. LASTING) handelt. Wenn nein, wird zur Regel 27 gesprungen, wenn ja, wird mit Hilfe der nächsten Regel getestet. Dort wird nun geprüft, ob auf LAST die Wörter TIME oder FALL unmittelbar folgen (K+1,K+1; "last time" oder "last fall"). Wenn ja, geht es weiter mit Regel 12, wenn nein, wird die nächste Regel geprüft. Nun wird geprüft, ob vor LAST das Wort THE steht (K-1,K-1; "the last"). Wenn nein, wird der Code 1 zugewiesen, das Wort als Adjektiv eingestuft, und es erhält folgende syntaktische und semantische Codes: DET, NUMB, ORD, MODIF, TIMESP. Wenn diese Bedingung nicht erfüllt ist, wird die nächste Regel angewandt. So werden alle Regeln nacheinander abgearbeitet, bis die Bedeutung des Worts feststeht.

4.0 Die verfügbaren Diktionäre zum General Inquirer: Harvard IV-3 und Harvard IV-4 Diktionär, Lasswell Value Diktionär

Der General Inquirer wird zusammen mit drei Diktionären, dem Harvard (Version IV-3 und IV-4) sowie dem Lasswell Value Diktionär ausgeliefert. Im folgenden soll kurz auf die theoretischen Ansätze und die allgemeine Struktur dieser Diktionäre eingegangen werden. Für eine ausführlichere Erörterung wird auf Züll et al. verwiesen.

4.1 Harvard IV-3 und IV-4

Die Harvard Diktionäre gehen auf eine Idee von Philip J. Stone zurück, der in einem umfassenden Klassifikationssystem psychologische und soziologische Ansätze zusammenfügte. Dies spiegelt sich auch im vollen Namen dieser Diktionäre wieder, der "Harvard Psycho-Sociological Dictionary" lautet. Dabei fanden aus der Psychologie insbesondere Vorstellungen von Freud und aus der Soziologie solche von Parsons Eingang in die Kategorienkonstruktion. Version IV-3 ist Stones Weiterentwicklung der im General Inquirer von 1966 beschriebenen Version III (Stone et al. 1966). Insbesondere hat er in der neuen Version versucht, vermehrt syntaktisch "reine" Kategorien zu bilden, indem er systematisch zwischen Nomen, Verben und Modalen unterscheidet. Dadurch erhält man zusätzlich zur Bedeutung auch noch Hinweise auf die Statik (Nomen) oder Prozesshaftigkeit (Verben) in den untersuchten Texten. Weiterhin wurde versucht, die Kategorien als Meßindikatoren mittels verschiedener Maßnahmen robuster als früher zu machen. So konnte es früher der Fall sein, daß eine Kategorie nur dann valide maß, wenn mehrere ihrer Wörter in einem Text auftraten. Trat dagegen von einer Kategorie in einem Text immer nur ein Wort auf, konnten erhebliche Verzerrungen entstehen (z.B. in der Kategorie ACTION NORM, wenn in einem Text nur über BREAKFAST gesprochen wurde, dann war dies eher ein Indikator für "ORAL" denn für ACTION NORM).

Entsprechend seiner allgemeinen Zielsetzung spannen die Kategorien ein weites Feld zwischen "Animal Noun", über "Hostile" bis zu "Work" auf. Zusätzlich zu der syntaktischen Trennung in Nomen, Verben und Modale, den semantischen Hauptkategorien enthält der Harvard IV-3 eine Reihe von Hilfskategorien (Markerkategorien), wie z.B. "CONJ" für Konjunktionen, "QUOTE" für Anführungszeichen usw., die u.a. auch für die Disambiguierung benötigt werden. Die Version IV-4 ist eine vollständig revidierte Fassung, die von Dexter C. Dunphy, Cederik G. Bullard und Elinor R.M. Crossing (University of New South Wales/Australien) entwickelt wurde. Dunphy et al.[7] überprüften zuerst die hierarchische Ordnung des Klassifikationssystems und validierten dann in einem aufwendigen Verfahren die einzelnen Kategorien. Ihre Version des Harvard Diktionärs enthält 65 Markerkategorien als Hilfskategorien für die Disambiguierung, 79 "First Order" Kategorien, die den semantischen Kern ausmachen, und 25 weitere "Second Order" Kategorien, die zusätzliche Dimensionierungen der Analyse erlauben.

[7] Dunphy, D. C., C. G. Bullard, and E. E. M. Crossing. 1974. Validation of the General Inquirer Harvard IV Dictionary. Paper presented at 1974 Pisa Conference on Content Analysis.

4.2 Lasswell Value Diktionär

Der Lasswell Value Diktionär (LVD) geht auf eine Idee von Zvi Namenwirth zurück, der damit eine Umsetzung der Theorien von Lasswell und Kaplan[8] in ein inhaltsanalytisches Klassifikationsschema anstrebte. In neueren Veröffentlichungen wird dieser Anspruch eher zugunsten einer Definition als allgemeines Klassifikationsschema denn als direkte Operationalisierung bestimmter theoretischer Überlegungen zurückgenommen (Namenwirth & Weber, 1987). Dennoch zeigt sich in der Organisation auch weiterhin der Bezug zu Lasswell und Kaplans politisch-soziologischem Ansatz. So wird von acht Hauptkategorien ausgegangen, die zu zwei Gruppen zusammengefaßt werden können: die "Deference Values" (POWER, RECTITUDE, RESPECT, AFFECTION), die als Indikatoren für Interaktionen zwischen Handelnden genommen werden, und die "Welfare Values" (WEALTH, WELL BEING, ENLIGHTENMENT, SKILL), die als Indikatoren für die Aktivitätserhaltung (notwendige Ressourcen der Individuen) genommen werden. Zu diesen acht Hauptkategorien treten, ähnlich den Second Order Kategorien des Harvard IV-4, weitere Nebenkategorien, wie die Gruppe der "general value transaction indicators" (z.B. transaction gains, transaction losses), einer Kategorie "ANOMIE", Indikatoren für Gefühle (Affect), und Indikatoren für Raum und Zeit. Dazu kommen noch die Markerkategorien für die Disambiguierung.

5.0 Programme des General Inquirers

Die Programme des General Inquirer lassen sich in drei Hauptgruppen unterteilen:

1. Programme, die die Texte aufbereiten und bearbeiten (Disambiguieren und Vercoden);
2. Programme, die auf Grund der vorher erstellten Klassifikationen Häufigkeiten erstellen, die Vercodungen im Satzzusammenhang darstellen und die Vercodungen für die weitere Analyse aufbereiten;
3. Programme, die als Hilfsmittel zur Wörterbucherstellung und deren Überprüfung dienen.

Zur ersten Gruppe gehören die Programme TEXTREAD und TAGGER, zur zweiten TALLY, RETRIEVE und CONNECT, und die dritte setzt sich aus den Programmen PARSER, DICTMERG, PEEL und DOCUMENT zusammen.

[8] Lasswell, H. D., and A. Kaplan. 1963 (1950). Power and Society: A Framework for Political Inquiry. New Haven: Yale University Press.

5.1 Programm TEXTREAD

TEXTREAD liest den Eingabetext des Anwenders und erstellt daraus einen internen "Systemfile", der für die Disambiguierung und Vercodung mit TAGGER erforderlich ist. TEXTREAD hat dabei zwei wesentliche Aufgaben. Zum einen zerlegt es den Text abhängig von Satzzeichen in einzelne Sätze. Die Sätze bilden dann die Basis für die weitere Vercodung, d.h. der General Inquirer arbeitet dabei immer auf Satzebene. Zum anderen werden von TEXTREAD bereits Kategorien für Satzzeichen und Zahlen vergeben und Abkürzungen überprüft (z.B. wird Mr. als Abkürzung erkannt, gekennzeichnet und entsprechend vercodet).

Der General Inquirer kann alle (und nur) englischsprachige Texte verarbeiten. Bei der Verschriftung bzw. Textvorbereitung müssen dabei nur einige Grundregeln beachtet werden:

- Die Texte können in Groß-/Kleinschreibung verschriftet sein. Zur Vercodung werden sie jedoch immer in Großbuchstaben umgesetzt, da alle Regeln in den zur Verfügung stehenden Wörterbüchern nur in Großschreibung gespeichert sind. Diese Einschränkung stellt bei englischsprachigen Texten (im Gegensatz zu deutschen) keinen Informationsverlust dar.
- Sätze dürfen maximal 2000 Zeichen lang sein, Wörter 20 Zeichen. Längere Sätze werden vom Programm automatisch in kürzere zerlegt, längere Wörter werden abgeschnitten. Das Programm druckt in solchen Fällen eine entsprechende Warnung.
- Der Text kann durch Identifikatoren in logische Blöcke unterteilt werden, die aber auf die Analyse keinen Einfluß haben, sondern nur der Dokumentation dienen. Dasselbe gilt für Kommentare, die in den Text eingefügt werden können. Während die Identifikatoren als Dokumentation auch für TAGGER und die nachfolgenden Programme erhalten bleiben, werden die Kommentare im Text von TEXTREAD ausgedruckt, danach aber aus dem Text entfernt.

5.2 Programm TAGGER

Nachdem TEXTREAD den Eingabetext aufbereitet hat, übernimmt TAGGER die entgültige Disambiguierung und Vercodung. TAGGER vercodet auf Grund der Regel eines Wörterbuchs (siehe oben). Wie diese Regeln aufgebaut sind und wie die Vercodung erfolgt, wurde im Kapitel 3.0 ausführlicher beschrieben. Jedem Wort werden ein oder mehrere syntaktische und

semantische Codes zugewiesen. Bei nicht eindeutigen Wörtern wird versucht mit Hilfe des Satzzusammenhangs, in dem sie stehen, festzustellen, welche Bedeutung sie an dieser Stelle haben (Disambiguierung). Diese Wörter werden in der Ausgabe in der Form Wortstamm, Bedeutungscode und gegebenenfalls Endung (z.B. ARMS wird ARM1S) gespeichert. Die nachfolgende Tabelle zeigt anhand eines Satzes die Vercodung durch TAGGER. Sie ist ein Auszug aus der Druckerausgabe von TAGGER. In der letzten Spalte sind alle zugewiesenen syntaktischen und semantischen Kategorien aufgeführt.

Tabelle 5.2.1: Vercodung durch TAGGER zu dem Satz "Systems of thought have been jarred, ways of life have been uprooted, institutions are under siege."

```
****TRACE****SENTENCE 10 ******DOCUMENT 1 ******IDENTIFICATION AD1968
1: SYSTEM      S   0  1  NOUN S B ABS ABS* THINK KNOW ORDER
2: OF              0  1  PREP ROOT
3: THOUGHT  4      4  1  NOUN ED ABS ABS* THINK KNOW ACTVI
4: HAVE     3      4  1  SUPV ROOT VERB HAVE
5: BEEN     3      2  1  SUPV ROOT VERB BE ED PASSIVE
6: JAR      2 RED  4  1  SUPV ED MOVE EXERT NGTV) ACTV2 STRNG2
7: ,               0  1  PUNC COMMA
8: WAY      1 S    3  1  NOUN S THINK EVALU MEANS PFREQ
9: OF              0  1  PREP ROOT
10: LIFE           0  1  NOUN ROOT NATURE
11: HAVE    3      4  1  SUPV ROOT VERB HAVE
12: BEEN    3      2  1  SUPV ROOT VERB BE ED PASSIVE
13: UPROOT     ED  0  1  ED X
14: ,              0  1  PUNC COMMA
15: INSTITUTION S  0  1  NOUN S PLACE SOCIAL POLIT ECON STRNG2
16: ARE     1      2  1  SUPV ROOT VERB BE
17: UNDER   1      1  1  PREP ROOT POWER -
18: SIEGE          0  1  E X ROOT
19: .              0  1  PUNC PER
```

Alle Wörter, die nicht im Wörterbuch definiert sind und nicht vercodet werden können, werden in einer Leftover-Liste gedruckt und in einer Datei (als Wortstamm) abgelegt. Sie müssen, sofern sie für die Analyse von Bedeutung sind, nachträglich ins Wörterbuch aufgenommen oder anderweitig nachvercodet werden.

Am Ende der Vercodung erhält der Anwender einige Statistiken zum eingesetzten Wörterbuch und zum Text wie z.B. durchschnittliche Länge der Regeln im Wörterbuch, Zahl der Sätze im Text, Anzahl aller Wörter und Zahl der Satzzeichen.

5.3 Programm RETRIEVE

Nachdem die Vercodung beendet ist, stehen dem Anwender einige Möglichkeiten der Überprüfung der Vercodung und der Weiterverarbeitung zur Verfügung. Eines dieser Programme ist RETRIEVE, das dem Anwender die Möglichkeit gibt, Sätze auf Basis ausgewählter Codes auszudrucken. Eingabe in RETRIEVE ist die von TAGGER erstellte Ausgabedatei. Tritt ein vorgegebener Code in einem Satz auf, wird der ganze Satz gedruckt. Dabei sind alle dieser Kategorie zugeordneten Wörter unterstrichen. Bei großen Texten kann eine Satzauswahl der Form "jeder n-te Satz wird gedruckt" getroffen werden.

RETRIEVE erlaubt drei verschiedene Auswahlmöglichkeiten: Nach der ersten werden alle Sätze gedruckt, in denen die vom Anwender vorgegebene Kategorie auftritt (z.B. alle Sätze, in denen Kategorie POLIT vergeben wurde).

Eine zweite Möglichkeit der Auswahl ist die Definition zweier Kategorien. Das Auswahlkriterium heißt, daß ein Satz dann gedruckt wird, wenn in ihm Kategorie A und B zusammen auftreten (AND Verknüpfung) (z.B. alle Sätze, in denen die Kategorien POLIT und POWCON (power conflict) auftreten).

Ein drittes Auswahlkriterium bezieht sich auf Mehrfachvercodungen eines Worts (z.B. Sätze, in denen für ein Wort sowohl NOUN als auch POLIT vergeben wurden).

5.4 Programm TALLY

TALLY erlaubt einfache Häufigkeitsauszählungen aller syntaktischen und semantischen Kategorien. Als Grundlage dient dazu die TAGGER Ausgabedatei. Die Häufigkeiten werden pro Dokument berechnet und gespeichert, d.h. immer wenn die Identifikation wechselt, werden neue Häufigkeiten errechnet.

Da es verschiedene Möglichkeiten der Prozentuierung gibt, erstellt TALLY nur einfache Häufigkeiten. Ein Beispiel für verschiedene Vorgehensweisen bei der Prozentuierung ist die Kategorie der N-Type-Wörter im

Lasswell Value Diktionär (siehe dort). Da diese Wörter im allgemeinen keine semantische Bedeutung haben, sollen sie nicht immer in die Prozentuierung eingehen. Nutzer der Harvard IV Diktionäre dagegen möchten die Summe aller Wörter zur Prozentuierung verwenden. TALLY erstellt immer eine Datei mit allen Häufigkeiten, die dann zu weiteren Analysen mit Statistiksoftware-Paketen verwendet werden können.

5.5 Programm CONNECT

CONNECT ermöglicht den Übergang von General Inquirer zu anderen Textanalyse und Statistikprogrammen (z.B. zu TEXTPACK V oder zu SAS, SPSS usw.). Um mit den Vercodungen oder dem disambiguierten Text weitere Analysen machen zu können, stellt CONNECT geeignete Dateien zur Verfügung. Alle Ausgabesätze können mit Identifikatoren und Satznummern versehen werden. Dabei sind drei Formen der Ausgabe möglich.

Die erste Form ist die Ausgabe des disambiguierten Texts. Dieser Text kann so formatiert ausgegeben werden, daß er mit anderen Textanalysesystemen wie z.B. TEXTPACK V[9] weiterverarbeitet werden kann. Alle nicht eindeutigen Wörter, für die Disambiguierungsregel im Diktionär vorhanden waren, werden in der Textdatei so gespeichert, daß nach dem Wortstamm der Bedeutungscode folgt (siehe Disambiguierung).

Eine weitere Möglichkeit ist die Ausgabe einer numerischen Datei, die pro Satz für jede Kategorie die Kennzeichnung vergeben/nicht vergeben enthält (Dummy Variablen).

Die dritte Form der Ausgabe ist ebenfalls eine numerische Datei, die pro Kategorie die Häufigkeit des Auftretens in diesem Satz enthält. Die beiden numerischen Dateien können mit Statistiksoftware-Paketen weiterverarbeitet werden.

5.6 Programm PARSER

PARSER ist das zentrale Programm zum Bearbeiten von Diktionären. Das Programm liest die Vercodungs- und Disambiguierungsregeln des Benutzers, die in einer eigenen Sprache abgefaßt sind (siehe Wörterbücher, Kapitel 3.0). Die Regeln werden auf syntaktische Fehler überprüft und in ein internes Format umgesetzt, so daß TAGGER damit seine Disambiguierung

[9] TEXTPACK ist ein von ZUMA entwickeltes und betreutes Programmsystem zur computerunterstützten Inhaltsanalyse . Es ist für Mainframe-Rechner und PCs unter MS-DOS erhältlich.

und Vercodung durchführen kann. Das Programm kann verwendet werden, um neue Regeln aufzubauen oder bestehende Regeln zu ändern. Soll ein bestehendes Wörterbuch verändert bzw. erweitert werden, so werden alle neuen bzw. geänderten Regeln zunächst mit PARSER umgesetzt und anschließend mit dem im folgenden beschriebenen Programm DICTMERG an das bestehende Wörterbuch, bereits mit PARSER bearbeitete Wörterbuch angefügt.

5.7 Programm DICTMERG

Mit DICTMERG können neue oder geänderte Regeln an ein bestehendes Wörterbuch angefügt werden. Sowohl das bereits bestehende Wörterbuch wie auch die neu dazuzufügenden Regeln müssen vorher mit PARSER geprüft und umgesetzt worden sein. DICTMERG ersetzt dann geänderte Regeln, löscht entsprechend markierte Regeln und fügt alle neuen Regeln in das Diktionär ein.

5.8 Programm PEEL

PEEL stellt ein Hilfsmittel dar, um aus einem Diktionär im sogenannten Source-Format für einzelne Wörter Regeln auszuwählen. Will man Regeln verändern, ist der einfachste Weg, die bestehende Form mit Hilfe eines Editors oder eines Textverarbeitungssystems (z.B. WORDMARC, WORD, NORTON EDITOR) zu editieren. Da die Regeln im Source-Format eine Satzlänge von bis zu 3600 Zeichen haben, kann es mit einigen Textverarbeitungsprogrammen Probleme beim Bearbeiten dieser langen Sätze geben. Mit PARSER bereits bearbeitete Diktionäre können in keinem Texteditor verändert werden. PEEL unterteilt deshalb alle Regeln in Zeilen mit nicht mehr als 80 Zeichen, die dann einfach zu bearbeiten sind. Diese Regeln können dann wieder mit PARSER und DICTMERG oder mit DOCUMENT an ein bestehendes Wörterbuch angefügt werden oder die alte Regel im Wörterbuch ersetzen.

5.9 Programm DOCUMENT

Um bestehende Diktionäre zu drucken oder zu dokumentieren, kann DOCUMENT verwendet werden. DOCUMENT druckt die Regeln eines Wörterbuchs oder Listen aller im Wörterbuch vorhandenen Einträge. Optional können entweder alle Einträge mit semantischer Bedeutung oder alle

Einträge, die nur syntaktische Bedeutung haben (z.B. Artikel) oder auch alle Einträge (semantische und syntaktische) aufgelistet werden. Alle Listen werden in drei weitere syntaktische Gruppen unterteilt: Substantive, Verben und anderes.

Eine zusätzliche Möglichkeit von DOCUMENT ist das Ändern und Erweitern von Source-Format Diktionären. Der Anwender gibt seine Regeln ein. Diese Regeln werden dann an das bestehende Wörterbuch angefügt oder ersetzen dort eine bestehende Regel.

6.0 Ausblick

Der General Inquirer mag wie ein Dinosaurier im weiten Feld der heutigen Textmanipulationsprogramme wirken. Aber, trotz aller Fortschritte in der Hard- und Software und vor allem in der Benutzeroberfläche solcher Programme bleibt die Leistung des General Inquirers, große und größte Textmengen, zu disambiguieren und zuverlässig zu klassifizieren, unerreicht. Dies liegt vor allem an den Disambiguierungsregeln und den umfassenden Klassifikationsschemata des Harvard IV sowie des Lasswell Value Diktionärs, denen nichts vergleichbares im englischen Sprachraum entgegensteht. Für die historische Textanalyse eröffnet der General Inquirer, wie wir anzudeuten versuchten, neue Dimensionen der Massendatenanalyse. Für den deutschen Sprachraum empfiehlt sich TEXTPACK[10] mit seiner Menüführung, sofern man keine Disambiguierung benötigt. Es ist zu hoffen, daß mit der allgemeinen Verfügbarkeit des General Inquirers, der Klassifikationsschemata und von TEXTPACK solche Untersuchungen nicht mehr nur wenigen Spezialisten vorbehalten bleiben.

[10] Züll, C., P. Ph. Mohler, A. Geis, 1991. Computerunterstützte Inhaltsanalyse mit TEXTPACK PC. Stuttgart: Gustav Fischer.