Qualitative Inhaltsanalyse durch Computer - ein uneinlösbarer Anspruch?
-
Untersuchungen zur algorithmischen Textinhaltserschließung am Beispiel der referentiellen Interpretation

Thesenpapier zum Disputationsvortrag

Roland Stuckardt

7. Januar 2000

1 Ausgangspunkt: Quantitative vs. Qualitative Inhaltsanalyse

Gegenstand der Dissertation ist die *Inhaltsanalyse* von Texten als zentrale Methode der empirischen Sozialforschung. Allgemein wird unterschieden zwischen Quantitativer und Qualitativer Inhaltsanalyse, zwei Herangehensweisen mit grundsätzlich unterschiedlichem Skopus. Die *Quantitative Inhaltsanalyse* findet ihre Anwendungsfelder als bewährtes Standardverfahren zur Analyse von Massentexten. Zentrales Merkmal ist die *datenreduzierende* Erzeugung einer quantitativen Deskription der zu analysierenden Texte - üblicherweise auf der Grundlage eines Kategorisierungsvorgangs, der auf der Definition von Kategorien als Wortmengen basiert; gemäß einer klassischen Definition zielt dieses Verfahren auf die Erschließung des "manifesten Inhalts" ab. Anders die *Qualitative Analyse*, die - ausgehend von mannigfaltigen Kritiken betreffend den Objektivitätsanspuch sowie die Reichweite des quantitativen, wortzählend-vermessenden Vorgehens - einen umgekehrten, *induktiv-theoriebildenden* Weg einschlägt, indem die subjektiv-hermeneutischen Verstehensprozesse nicht ausgeblendet, sondern gezielt ins Zentrum des inhaltsanalytischen Vorgehens gestellt werden. An Stelle der vollständig algorithmischen Explikation tritt die *Standardisierung* von *intellektuellen* Erschließungsverfahren, durch die die intersubjektive Reproduzierbarkeit der Ergebnisse möglichst gut gewährleistet werden soll. Die Materialmenge in typischen Anwendungsgebieten ist überschaubar und bleibt als solche stets voll zugänglich; Kategorisierung geht hier *nicht* mit Datenreduktion einher.
Entsprechend der gemeinhin vertretenen Lehrmeinung ist ausschließlich die Quantitative Inhaltsanalyse durch Computer automatisch, d.h. betreffend den Kategorisierungsvorgang algorithmisch bewerkstelligbar; im Rahmen von qualitativen Analysen beschränkt sich die Rolle des Computers auf die eines Werkzeugs zur *Unterstützung* des explorativen Spiels mit den Daten, wobei das Ziel in der Förderung des intellektuellen, hermeneutischen Verstehensprozesses besteht.
Im Rahmen der weiteren Ausführungen wird somit auf die *Kategorien-* oder auch *Themenanalyse* als Standardverfahren sowohl der Qualitativen als auch Quantitativen Inhaltsanalyse fokussiert.

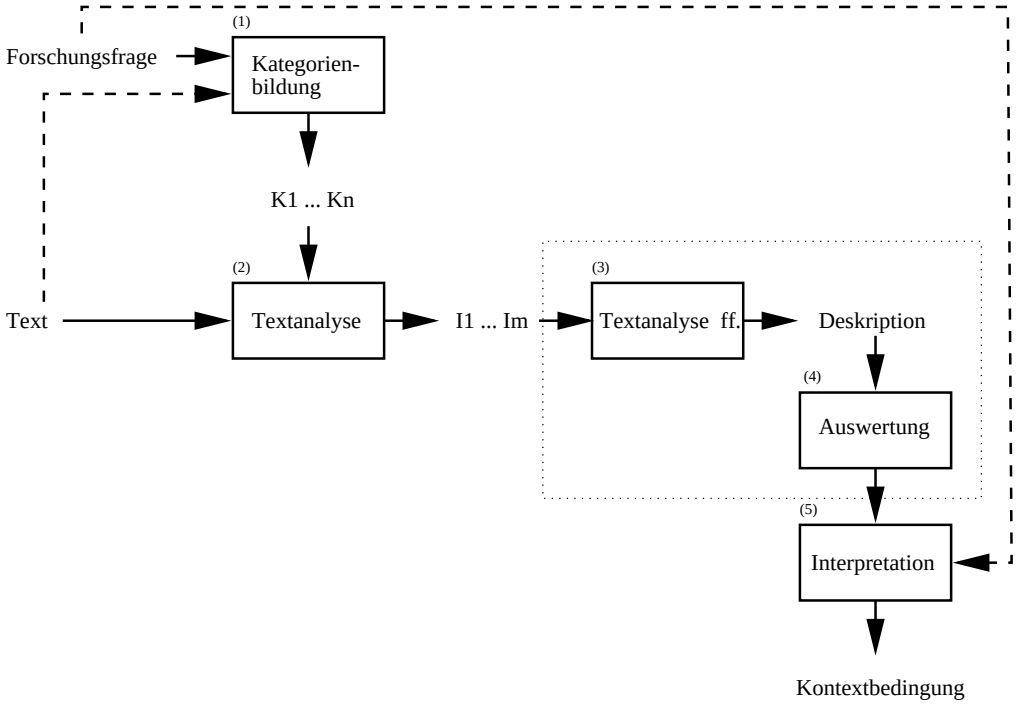

Abbildung 1: Die Kategorienanalyse als Standardverfahren der Inhaltsanalyse

In Abbildung 1 ist das Schema einer solchen Analyse skizziert. Im Speziellen unterscheidet sich die Quantitative Analyse von der Qualitativen Analyse durch die datenreduzierend-quantifizierende Ausrichtung der Schritte (3) und (4).

2 Qualitative und Quantitative Elemente der Analyse stehen in einem Verhältnis der *Komplementarität*

Die folgenden Beobachtungen bilden die Basis für eine differenziertere Bewertung der beschriebenen Dichotomisierung "Qualitative vs. Quantitative Inhaltsanalyse":

1. Das Postulat einer "objektiven" quantitativen, wörterbuchbasierten Inhaltsanalyse, die auf die Erschließung "manifester Inhalte" abzielt, erweist sich als problematisch. Bereits die Formulierung der Kategorisierungskriterien in Form von Wortlisten unterliegt unausweichlich *subjektiven* Elementen, nämlich denen des inhaltsanalysierenden Forschers, dem damit die Fähigkeit zugesprochen würde, "durchschnittliche Bedeutungen" zu definieren.

2. Die Operation des wörterbuchbasierten Kategorisierens ist das elementare Beispiel einer *qualitativen* Operation; "quantitativ" an der Quantitativen Analyse ist die sich üblicherweise anschließende datenreduzierende Deskriptions- und Auswertungsphase, in der auf ein mathematisch-statistischen Instrumentarium zurückgegriffen wird.

3. Mittlerweile gemeinhin akzeptiert ist die Feststellung, dass qualitative und quantitative Elemente in der Inhaltsanalyse nicht in einem Verhältnis des wechselseitigen Ausschlusses, sondern in einer Beziehung der *Komplementarität*, d.h. der gegenseitigen *Ergänzung* stehen. Im Speziellen schlägt sich dies in der Existenz unterschiedlicher typischer Anwendungsfelder der Qualitativen bzw. Quantitativen Inhaltsanalyse nieder.

3 Forschungsfrage: Ist der Codierschritt der Computergestützten Inhaltsanalyse auf algorithmischer Basis verfeinerbar?

Somit erweist sich die eingangs beschriebene Dichotomisierung "qualitativ vs. quantitativ" aus einer vorurteilsfreien Perspektive als ungeeignete Grundlage der - üblicherweise und vorschnell negativen - Entscheidung der Frage nach der Möglichkeit des Computereinsatzes für die automatisierte Inhaltserschließung. Wenn der durch Computer bewerkstelligte Codierschritt der klassischen wörterbuchbasierten Analyse eine elementare *qualitative* Operation darstellt, die jedoch im *Grade* der Inhaltserschließungs-Qualität starken Einschränkungen unterliegt, so legt dies vielmehr die Untersuchung der folgenden Frage nahe:

> *Ist es möglich, den Kategorisierungsschritt der klassischen Computergestützten Inhaltsanalyse algorithmisch zu verfeinern, um von den Vorteilen*
>
> - *ideal reliable, da formal explizierte Kategorisierung,*
> - *automatisierte Verarbeitung großer Textmengen*
>
> *bei gleichzeitig verbesserter Qualität der Inhaltserschließung zu profitieren?*

Falls es gelänge, betreffend die spezifische inhaltsanalytische Aufgabenstellung "automatische Kategorisierung", d.h. in Bezug auf die Algorithmisierung von Schritt (2) des kategorienanalytischen Grundschemas in Abbildung 1 erhebliche Fortschritte zu erzielen, so wäre dies als Evidenz dafür zu bewerten, dass der Anspruch einer computergestützt-*qualitativen* Analyse nicht (unter Rekurs auf unhaltbare Grenzen algorithmischer Explizierbarkeit) als uneinlösbar eingestuft werden kann. Dies ist der Gegenstandsbereich der vorliegenden Arbeit.
Vorab ist zu betonen, dass die Rolle des Computers im Rahmen der maschinellen Textinhaltsanalyse dennoch völlig unprätentiös in der eines *Werkzeugs* zum hundertprozentig reliablen Vollzug bestimmter *symbolischer Transformationen* gesehen wird, das den Forscher von *bestimmten* qualitativen Operationen bei gleichzeitig maximaler, da algorithmischer Explikation der gewählten Vorgehensweise entlastet. Die schlußendliche Interpretation der generierten Beschreibungen vor dem Hintergrund der jeweiligen Forschungsfrage verbleibt dem Forscher überlassen. Demnnach soll die interpretative Interaktion des Forschers mit dem Untersuchungsgegenstand, die zu einem *verstehenden* Erschließen der Inhalte unabdingbar ist, keineswegs eliminiert werden. Der Computereinsatz betrifft ausschließlich den Codierschritt, von dessen intellektueller Durchführung der Forscher unter bestimmten Voraussetzungen befreit werden kann.

4 Der Skopus der klassischen, wörterbuchbasierten Inhaltsanalyse erweist sich aus textlinguistischer Sicht als beschränkt

Analysiert man die wörterbuchbasierte Inhaltsanalyse im Hinblick auf die unmittelbar mit dem Modell der Einzelwortanalyse verbundenen analysetechnischen Probleme, so stellt sich die Frage nach der Algorithmisierbarkeit folgender Aufgabenstellungen: Lösung des Flexionsproblems (Lemmatisierung), Auflösung von Wortform-Ambiguitäten und lexikalischen Mehrdeutigkeiten (Homonymie, Polysemie), Behandlung von Komposita sowie Resolution pronominaler Ausdrücke. Mindestens ebenso wichtig ist jedoch die Diskussion aus der Perspektive der inhaltsanalytischen *Zielsetzung*. Welchen Ansprüchen wird eine Einzelwortanalyse überhaupt gerecht? Anders ausgedrückt: Stellen Aufzählungen einzelner Wörter überhaupt ein geeignetes Instrument zur Spezifikation inhaltlich unterschiedlicher Kategorien dar? Wo liegen die Grenzen derartiger Beschreibungen?
Zur Erörterung dieser Frage bietet sich ein Exkurs in die Textlinguistik an. Textkonstituierender Zusammenhang wird auf vielfältige Weise etabliert. Ein spezifische Variante ist die von A. J. Greimas mit dem Begriff *Isotopie* belegte und auf der lexikalischen Ebene verankerte[1]

> *"Wiederkehr von Wörtern desselben Bedeutungs- bzw. Erfahrungsbereichs in einem Text, z.B. Arzt, Fieber, Spritze, Honorar."*

Ausgehend von dieser Definition läßt sich feststellen, dass die wörterbuchbasierte Operationalisierung der Kategorien in der computergestützten Inhaltsanalyse als Spezifikation einer Menge von Isotopieebenen auffaßbar ist. Jede Kategorie entspricht einer Isotopieebene; die Wortlisten sind vollständige Enumerationen der sprachlichen Ausdrücke, denen ein gemeinsames Merkmal - umschreibbar als *Kategorie_i: +* - zugeordnet wird. Folglich werden per Codierung und Häufigkeitsauszählung unterschiedliche Isotopieebenen eines Textes bestimmt: Kategorien, die häufig instanziiert werden, entsprechen Folgen von Bedeutungseinheiten (Wörtern), die zum textuellen Zusammenhang über das den Vertretern dieser Isotopieklasse gemeinsame Bedeutungsmerkmal beitragen. Der computergestützten Inhaltsanalyse im klassischen Sinne liegt also implizit eine Isotopieanalyse entlang extensionaler Isotopieebenen-Definitionen im Wörterbuch zugrunde. Betrachtet man die Isotopieebenen, in Anlehnung an Bußmann (ibd.)

> *"In der Anzahl der Isotopieebenen spiegelt sich die thematische Komplexität eines Textes."*

als (elementare, lexikalisch verankerte) Text-*Themen*, so ergibt sich als weiteres Programm die textlinguistische Analyse einer solchermaßen beschränkten Themenerschließung.
Ausgehend von folgender Abgrenzung des Begriffs des *Textthemas*

Definition 1 (Thema) *Ein Thema ist ein textuell etablierter (Teil-)Kontext, der im Text mehr als einmal instanziiert, d.h. wiederaufgegriffen wird.*

[1] Zit. Bußmann: *Lexikon der Sprachwissenschaft*, Verlag Kröner, 1990, S. 357

läßt sich nun zeigen, dass Isotopie (und Koreferenz) lediglich *atomare* Kohäsionsvarianten darstellen. Das Modell einer verfeinerten Themenanalyse ergibt sich unter Einbeziehung von Situationen als komplexere Kontexteinheiten, die teils in der syntaktischen Struktur widergespiegelt werden, in komplex gelagerten Fällen jedoch erst auf der Basis von tiefensemantischen Repräsentationen und Inferenzen erkennbar sind.
Die Limitation der wortlistenbasierten Klassifikation spiegelt sich in der Entwicklung verfeinerter Beschreibungskonzepte für die Kategoriendefinition wider. Jedoch erweisen sich derartige Erweiterungen unter textlinguistischen Gesichtspunkten letztlich als heuristische Ersatzverfahren mit i.a. ungenügender Reichweite: Alleine auf der Basis wortorientierter Kookkurrenztests lassen sich etwa themenrelevante Partizipationsrelationen nicht zuverlässig erkennen, wie sich am Beispiel der Operationalisierung einer Kategorie *"Schaffens-Akte"* auf der Grundlage eines typischen Inventars zusätzlicher Operatoren illustrieren läßt:

```
<ARCHITEKT> = Behrens ODER Utzon ODER Asplund ODER Gropius ...
<SCHAFFEN>  = entwickeln ODER erarbeiten ODER
              erschliessen ODER verwirklichen ...

001         <ARCHITEKT> VOR <SCHAFFEN>
```

deckt eben nur bestimmte Standardrealisierungen ab:[2]

(1a) Behrens entwickelte eine architektonische Konzeption.
(1b) Behrens erschloß als Formgestalter ein neues Spezialgebiet.

Wird die Kernaussage anders ausgedrückt, so greift auch das verfeinerte Kriterium nicht:

(1c) Als Formgestalter erschloß Behrens ein neues Spezialgebiet.

Teils werden auch Auswahleinheiten, die irrelevant sind, fälschlicherweise kategorisiert:

(1d) Behrens entwickelte sich zu einem Designer von Weltrang.

In diesen problematischen Fällen setzt eine korrekte Klassifikationsentscheidung zumindest die Berücksichtigung des syntaktischen Kontexts voraus.
Natürlich kann die obige Kategoriedefinition auf der Basis des erweiterten Operatoreninventars weiter verfeinert werden. Jedoch wird mit solchen Reparaturversuchen letztendlich nur das Symptom kuriert: Eine adäquate Theorie, mit Hilfe derer ein Bezug zwischen sprachlichen Ausdrücken und kommunizierten Inhalten hergestellt wird, liegt einer derartigen Vorgehensweise nicht zugrunde. Erfahrungen aus konkreten Anwendungen belegen, dass die Beschreibungen ab einem gewissen Punkt derart komplex werden, dass weitere Verfeinerungen nicht mehr praktikabel erscheinen. Die praktisch unbegrenzte Vielfalt sprachlicher Ausdrücke, mit der eine bestimmte inhaltliche Einheit kommuniziert werden kann, sprengt den Rahmen der Möglichkeiten einer

[2]Im Rahmen des Beispiels wird vereinfachend unterstellt, dass das Flexionsproblem gelöst ist.

einzelwortorientierten Analyse, da die Themenanalyse durch das erweiterte Operatoreninventar nicht auf *linguistische fundierte* Weise generalisiert wird. Somit gilt

These 1 *Die Beschränkungen der klassischen maschinellen/Quantitativen sozialwissenschaftlichen Inhaltsanalyse sind primär auf die unzureichende textlinguistische Fundierung des einzelwortorientierten Codierschritts zurückzuführen, durch den die "Themen" des zu analysierenden Texts nur auf einer elementaren lexikalischen Ebene erschlossen werden. Das Ziel einer thematischen Analyse wird somit nur in sehr eingeschränktem Umfang erreicht. Die in der Literatur vorgeschlagenen "Verfeinerungen" der maschinellen Inhaltsanalyse sind linguistisch unfundiert; die Grenzen der klassischen Verfahren werden nicht in entscheidendem Ausmaß transzendiert.*

5 Die Computergestützte Inhaltsanalyse kann auf der Grundlage gegenwärtiger Technik algorithmisch verfeinert werden

Zunächst stehen zumindest für einen Teil der zuvor identifizierten modellbezogenen Probleme der einzelwortorientierten Inhaltsanalyse bereits ausgereifte computerlinguistische Lösungen zur Verfügung. So kann das Flexionsproblem mittlerweile als gelöst angesehen werden; auch für die Wortarten- sowie lexikalischen Disambiguierung stehen Algorithmen mit ausgezeichneten bzw. guten Erfolgsquoten zur Verfügung. Um der Unzulänglichkeit der wörterbuchbasierten Inhaltsanalyse abzuhelfen, ist jedoch das Problem des restringierten Inhaltsbezug des auf Isotopie basierenden Kategorisierungskriteriums zu adressieren. Inwieweit ist es möglich, diese Beschränkung auf der Basis tiefergehender Analysealgorithmen zu transzendieren? Das Verfahren der Wahl hat dabei den Rahmenbedingungen der computergestützten Inhaltsanalyse zu genügen, die sich unter dem zentralen Kriterium der *Massendatentauglichkeit* subsumieren lassen - Verfügbarkeit einfließenden Wissens, robuste Verarbeitung, Genre- und Domänenunabhängigkeit, adäquate Analysegeschwindigkeit - sowie nicht zuletzt natürlich eine hinreichende Güte der Ergebnisse.
Bereits die syntaktische Analyse, die ja in komplex gelagerten Fällen bezüglich der Themen-Erschließung noch immer zu kurz greift, erweist sich aus algorithmischer Sicht als problematisch, da das zur strukturellen Disambiguierung notwendige Wissen in typischen Anwendungsszenarien für die Mehrzahl der zu analysierenden Sätze nicht zur Verfügung stehen dürfte. So ergeben sich aus den Interpretationsvarianten der Präpositionalphrasen (jeweils adverbial oder attributiv) im Standardbeispiel

(2) Peter beobachtet den Mann im Park mit dem Teleskop.

insgesamt fünf syntaktische Lesarten;[3] aufgrund der Multiplizität der Mehrdeutigkeiten sind für komplexe Satzgefüge durchaus mehrere tausend Lesarten möglich. Somit ist bereits die Syntaxanalyse ein Problem, für das allenfalls *partielle* algorithmische Lösungen existieren, deren Ergebnis in der Regel aus einer *Menge syntaktischer Fragmente* bestehen wird. Anhand der Diskussion wird ferner deutlich, dass sich spezifische Anforderungen robuster Verarbeitung gegenüber Analysemoduln (Komponenten) ergeben, die auf der Verfügbarkeit syntaktischer Beschreibungen

[3]Die Lesart, die der Kombination ("*adverbial*","*attributiv zu Mann*") entspricht, ist aus strukturellen Gründen ausgeschlossen: Die entsprechende syntaktische Struktur ist - in diesem Fall unzulässigerweise - *nichtprojektiv*.

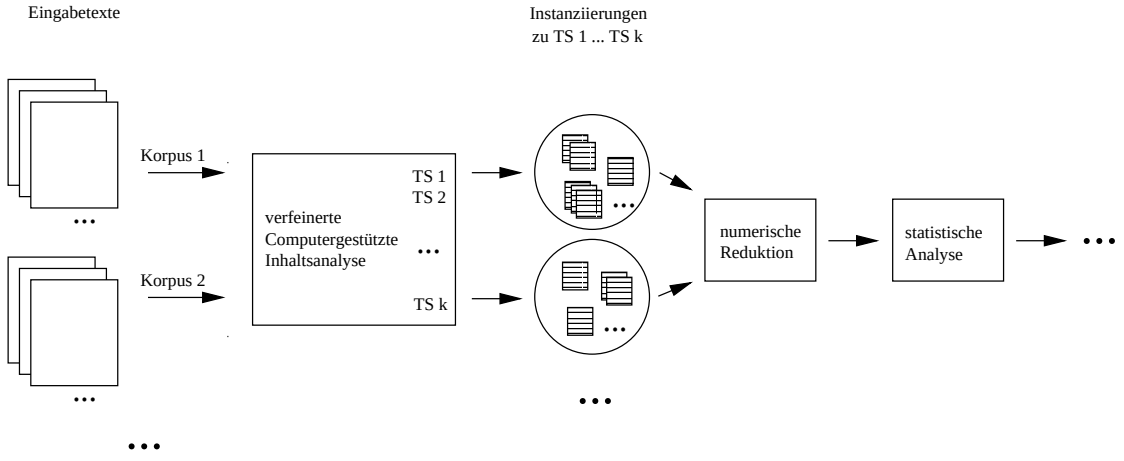

Abbildung 2: Verfeinerte Computergestützte Inhaltsanalyse mit thematischen Schablonen

aufbauen - ein Punkt, der sich im Rahmen der Algorithmisierung des Pronomenproblems als bedeutsam erweisen wird.
Wenn eine syntaktische Analyse einerseits nur einen Ausschnitt textueller Themen erschließen kann, und wenn andererseits bereits für die Syntaxanalyse derzeit nur partielle algorithmische Lösungen existieren, ist dann das Ziel einer algorithmischen Inhaltserschließung, die eine weiterreichende Erschließung des Textinhalts losgelöst von den spezifischen Ebenen der linguistisch-strukturellen Konstitution leistet, überhaupt realistisch? Eine nähere Analyse neuerer Entwicklungen in der Computerlinguistik - genauer: auf den Spezialgebieten des *Information Extraction* bzw. *Message Understanding* - zeigt auf, dass es tatsächlich auf dem gegenwärtigen Stand der Technik möglich ist, komplexe inhaltliche Beschreibungen (im informatischen Sinne: *Frames*) aus Texten automatisch, d.h. vollständig algorithmisch zu extrahieren. Die o.g. Rahmenbedingungen der Computergestützten Inhaltsanalyse in Bezug auf die Massendatentauglichkeit - Robustheit und Analysegeschwindigkeit - werden dabei eingelöst. Die Idee besteht nun darin, die *Frames* als komplexe thematische Schablonen aufzufassen, die als *inhaltsorientierte*, d.h. *von der konkreten linguistischen Realisierungsform an der Oberfläche abstrahierende* Beschreibungsgrundlage verfeinerter Kategorienschemata dienen. In Abbildung 2 ist das Prinzip einer solchermaßen bewerkstelligbaren algorithmischen Verfeinerung des Codierschritts (ST = *Scenario Template*, d.h. thematische Schablone) im Rahmen der klassischen Computergestützten Inhaltsanalyse skizziert. Gezählt werden demnach nicht mehr Wörter, sondern Instanzen der *inhaltsorientierten* thematischen Schablonen.
Zentrales Merkmal der "*Information Extraction*"-Systeme, die für die Teilnahme an der *Message Understanding Conference* MUC-6 (1996) entwickelt wurden, ist die aufwandsarme Portierbarkeit zu einer breiten Klasse inhaltlicher Domänen; dies wurde durch eine spezifische Rahmenbedingung von MUC-6 forciert, dergemäß den Teilnehmern nur der vergleichsweise kurze Zeitraum von einen Monat für die Konfiguration der Systeme auf die evaluationsrelevante Anwendungsdomäne zur Verfügung stand. Durch diese Anforderung wurde die softwaretechnische Trennung von domänenunabhängigen (insbesondere linguistischem) und domänenspezifischen Analysekomponenten gefördert - eine Bedingung, deren Essenzialität gerade mit Blick auf die zuvor identifizierten Probleme linguistisch unfundierter Verfeinerungen der wörterbuchbasierten Inhaltsanalyse auf der Hand liegt. In einer *adaäqat* verfeinerten Computergestützten Inhalts-

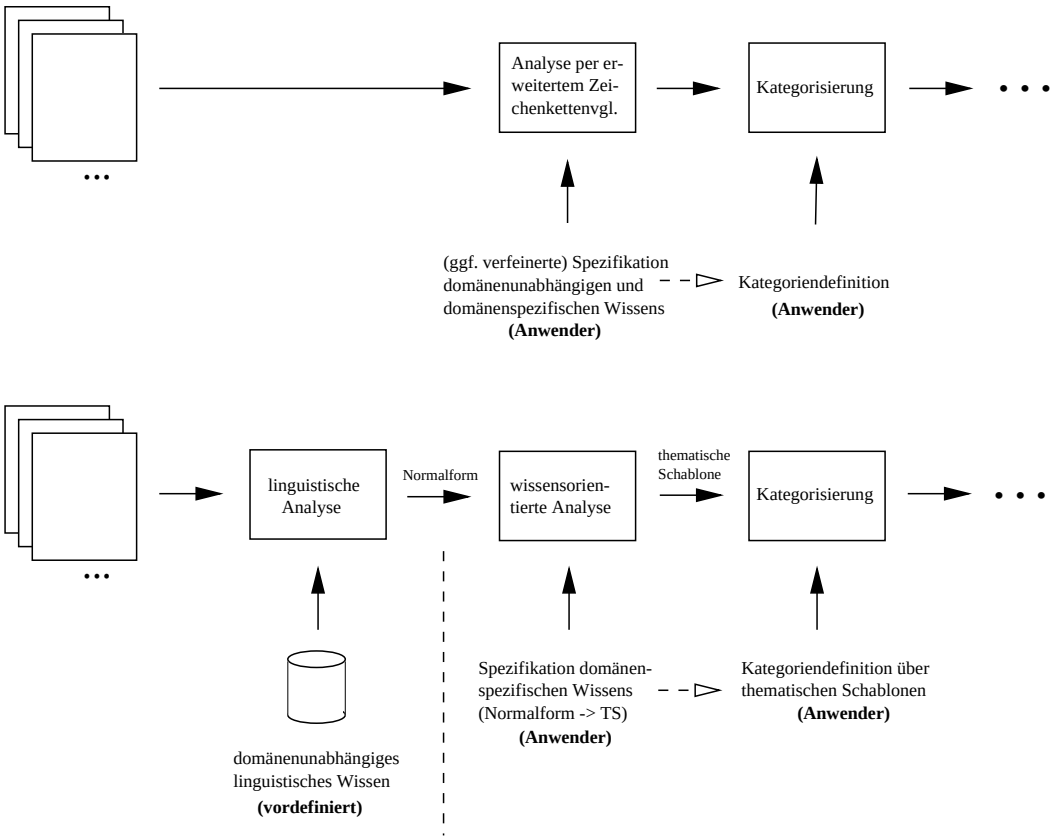

Abbildung 3: herkömmliche versus verfeinerte, portable Computergestützte Inhaltsanalyse

analyse, die auf den MUC-Vorarbeiten aufbaut, können somit per Herausfaktorisierung der linguistischen Analysekomponente *inhaltliche* Kategorienspezifikation und - wiederverwertbare - Spezifikation syntaktischer Ausdrucksvarianten an der sprachlichen Oberfläche künftig entkoppelt werden (vgl. die Gegenüberstellung in Abbildung 3).
Auch die im Rahmen von MUC-6 in Bezug auf die *Qualität* der Inhaltserschließung ermittelten Gütewerte (i.d.R. aufgegliedert nach *Precision* und *Recall*)[4] lassen anwendungsbezogene Experimente mit "*Information Extraction*"-Technik in der Computergestützten Inhaltsanalyse der Sozialwissenschaften Erfolg versprechend erscheinen. So gelingt die Entscheidung, *ob* ein bestimmtes Dokument in Bezug auf eine vorgegebene Inhaltsschablone relevent ist, mit einer Genauigkeit und einer Abdeckung von deutlich über 80 %; die Ermittlung aktueller Schablonen-*Instanzen* gelingt immerhin noch mit *Precision*-Werten oberhalb von 60 % und *Recall*-Werten oberhalb von 40 %. Eine genauere Untersuchung dieser Erschließungstechnik thematisch-inhaltlicher Information vor dem Hintergrund spezifischer, sozialwissenschaftlich einschlägiger Anwendungsszenarien wäre nun der nächste Schritt.
Die Ergebnisse der Betrachtungen lassen auf die Gültigkeit der folgenden These schließen:

[4]Die allgemeinen Definitionen sind *Precision* = Anzahl *korrekter* gefundener geteilt durch Gesamtanzahl gefundener Einzelergebnisse; *Recall* = Anzahl *korrekter* gefundener geteilt durch Gesamtanzahl gesuchter (korrekter) Einzelergebnisse. I.d.R. besteht zwischen diesen beiden Größen eine Tradeoff-Beziehung.

These 2 *Bereits auf der Basis des gegenwärtigen Stands der Forschung sind die wesentlichen Voraussetzungen für eine algorithmische, insbesondere auf linguistischem Wissen basierende Verfeinerung des Codierschritts als qualitative Operation der maschinellen Inhaltsanalyse erfüllt. Weder theoretische noch technische Gründe sprechen für eine Begrenzung maschineller Themenanalysen auf die einzelwortorientierte, lexikalische Ebene. Die Grenzen der klassischen Verfahren sind somit in entscheidendem Ausmaß transzendierbar.*

6 Pronomeninterpretation ist ein zentrales und inhärent schwieriges Teilproblem der Computergestützten Inhaltsanalyse

Es soll nun anhand eines in der einschlägigen Literatur immer wieder als prototypische Instanz algorithmischer Nichtexplizierbarkeit zitierten Kernproblems verfeinerter computergestützter Kategorisierung, der Interpretation anaphorischer (insbesonderer pronominaler) Ausdrücke, der konkrete Nachweis der Möglichkeit einer algorithmischen Expkation erbracht werden, die den zuvor identifizierten Anforderungen betreffend Massendatentauglichkeit, Robustheit, domänenübergreifende Gültigkeit und Analysegüte entspricht. Die zentrale Relevanz dieses Problems im Rahmen von wörterbuchbasierter Inhaltsanalyse sowie jeglichen Verfeinerungen wird mit Blick auf folgendes Beispiel ersichtlich:

(3) Peter Behrens ist berühmt. Er baute expressionistische Gebäude.

Um etwa auf die Anwendbarkeit der oben betrachteten Kategoriendefinition für Schaffensakte entscheiden zu können, ist zunächst der Bezug des Pronomens "*Er*" zu bestimmen, der hier durch das Antezedens "*Peter Behrens*" determiniert ist. Für die Kategorisierungsentscheidung sollte die konkrete sprachliche Realisierung der spezifizierten inhaltlichen Entität (pronominal vs. nichtpronominal) keine Rolle spielen. Aus textlinguistischer Sicht geht es um die Inhaltserschließung auf der *referentiellen* Ebene, der neben der oben identifizierten lexikalisch-semantischen Isotopie zweiten elementaren Spielart textueller Kohäsion und damit Themenentwicklung.
Folgendes Paar von Beispielen illustriert, dass das Problem der referentiellen Interpretation nicht in jedem Falle unter alleinigem Rekurs auf linguistisches Wissen lösbar ist und somit beliebig schwierig sein kann:

(4a) Peter Behrens besuchte Walter Gropius in Weimar.
Er informierte ihn über eine Künstlergruppe in Berlin.

versus

(4b) ...
Er informierte ihn über eine Künstlergruppe am Bauhaus.

In letzterer Variante könnte *Weltwissen* über die Beziehungen zwischen den mit den Ausdrücken "*Walter Gropius*", "*Weimar*" und "*Bauhaus*" bezeichneten inhaltlichen Entitäten - Gropius

lehrte am Bauhaus zu Weimar - eine im Vergleich zum ersteren Text umgekehrte Zuordnung der Pronomen nahelegen.
Wie bereits bei der Wahl der bisherigen Beispiele zugrunde gelegt, wird der Skopus der nachfolgenden Algorithmisierungsanstrengungen eingeschränkt auf (1) die Behandlung von als Nominalphrasen realisierten Ausdrücken, (2) die Erschließung von Beziehungen der *Ko*-Referenz, (3) die Interpretation *nicht-elliptischer* Ausdrücke. Durch diese drei Annahmen wird einerseits auf die typische Probleminstanz der computergestützten Inhaltsanalyse - Pronomen - fokussiert und andererseits die Anforderungen in der intellektuellen Erstellung der Referenzdaten (als Basis einer formalen Evaluation) in einem Rahmen gehalten werden, der eine - zur Gewährleistung der Aussagekraft der Ergebnisse unerlässliche - hinreichend hohe Intercoderreliabilität sicherstellt. Nicht zuletzt stehen die genannten Vorgaben im Einklang mit der Definition der entsprechenden *MUC-CO-Task*, die somit als Ausgangspunkt dienen kann.
Unter der vorläufigen Annahme, dass die Teilmenge referenzierender sprachlicher Ausdrücke (sog. *Vorkommen*) eines referentiell zu interpretierenden Textes bereits bestimmt und ferner bekannt ist, *welche* dieser Vorkommen *Anaphern* sind (d.h. auf eine textuell bereits eingeführte Entität verweisen), stellt sich das Problem der Interpretation anaphorischer Ausdrücke aus abstrakter Perspektive als *Entscheidungsproblem* dar:

> *Gegeben ist eine Menge V von Vorkommen und eine Teilmenge $A \subseteq V$ anaphorischer Vorkommen. Für alle Anaphern $a \in A$ ist unter Rekurs auf verfügbare Wissensquellen ein koreferenzierendes Antezedens $v_a \in V$ auszuwählen.*

7 Für die Algorithmisierung der Pronomeninterpretation erweist sich ein *Mehrstrategie*-Ansatz als adäquat

Somit stellt sich nun die Frage, auf welche Weise die ausstehenden Antezedensentscheidungen algorithmisch getroffen werden können. Carbonell und Brown schagen ein Analyseparadigma vor, derzufolge das Problem der computergestützten Koreferenz-Resolution durch eine Kombination elementarer Interpretationsstrategien anzugehen sei, die sich in die beiden Klassen *Restriktionen* als (quasi-)stringente Strategien und *Präferenzen* mit heuristischem Stellenwert unterteilen lassen. Bekannte Beispiele stringenter Strategien sind die *morphologische Kongruenzbedingung*, dergemäß in folgendem Beleg

> *(5) Behrens sagte seiner Kollegin, dass er sie schätze.*

die Antezedenten der drei Pronomen "*seiner*", "*er*" und "*sie*" alleine auf der Grundlage der distinktiven Merkmale grammatischer Genus und Numerus bestimmt sind, sowie *syntaktisch-konfigurationale Bedingungen*, durch die der Bezug der Pronomen in den folgenden Belegen

> *(6a) Behrens bemerkt, dass der Friseur sich rasiert.*
> *(6b) Behrens bemerkt, dass der Friseur ihn rasiert.*

aufgrund der jeweiligen Ausdrucksform (reflexiv vs. nichtreflexiv) auf ein (in noch festzulegendem Sinne) lokales bzw. nichtlokales Antezedens eingegrenzt werden kann. Während den

Restriktionen somit in Algorithmen zur Koreferenz-Resolution die Rolle eines A-Priori-Filters zur Elimination definitiv inkorrekter Antezedenskandidaten zukommt, bilden die Präferenzstrategien die Basis zur Auswahl unter den verbleibenden Kandidaten: Fokussierungstheoretisch gerechtfertigte und praktisch nutzbringende Heuristiken sind etwa die Regel der *Präferenz des syntaktischen Subjekts*, die in Beleg

(7) Gropius betrat den Raum. Behrens wartete bereits auf ihn.

zur (richtigen) Auswahl "*Gropius*" (anstelle des ebenfalls möglichen Antezedens "*Raum*") führt, sowie die Regel der *syntaktischen Rollenträgheit*, die in Fällen wie etwa

(8) Behrens besuchte Gropius in Dessau. Der Architekt bat ihn um Unterstützung.

zur (korrekten) Entscheidung für das (als transitives Objekt) syntaktisch rollengleiche Antezedens "*Gropius*" führt. Dass es sich hierbei nicht um stringente Kriterien, sondern um Heuristiken handelt, wird z.B. daran ersichtlich, dass die beiden genannten Präferenzstrategien in einigen Fällen zu gegensätzlichen Vorhersagen führen.
Ausgehend von der zuvor motivierten Unterscheidung zwischen stringenten und nichtstringenten Kriterien besteht die Idee nunmehr darin, das Entscheidungsproblem "Anaphern-Interpretation" unter Anwendung eines *dreiphasigen Rahmenalgorithmus* zu lösen, der in die Schritte Restriktionsanwendung, Präferenzbewertung und Antezedensauswahl unterteilt ist. Um zu einer operationalen Lösung zu gelangen, die den Rahmenbedingungen einer robusten, massendatentauglichen Inhaltsanalyse genügt, ist nun eine Teilmenge von Interpretationsstrategien zu identifizieren, die tatsächlich *algorithmisierbar* ist. In Bezug auf die unterschiedlichen in der Literatur vorgeschlagenen Restriktionen und Präferenzkriterien sind somit folgende Fragestellungen zu untersuchen: (1) Auf welcher Art von *Wissen* baut die jeweilige Strategie auf? (2) Ist die Strategie demnach operational umsetzbar? Welche ggf. *zusätzlichen Anforderungen bzw. Einschränkungen* bestehen? (3) Wie hoch ist die erwartete *Relevanz* der Strategie für das Problem der Anapherninterpretation auf anwendungstypischem, unrestringiertem Text?
Die Ergebnisse dieser Untersuchung sind in Abbildung 4 zusammengefaßt. Der obere Teil der Tabelle beschreibt die *Restriktionen*: morphologische Kongruenz, syntaktische Konfiguration, semantischer Skopus, globaler (intentionaler) Fokus, lokaler (attentionaler) Fokus, kataphorische Bezugnahmen, selektionale Bedingungen, Prä-Post-Bedingungen; im unteren Teil werden die *Präferenzkriterien* betrachtet: globaler (intentionaler) Fokus, lokaler (attentionaler) Fokus, Distanz, Topikalisierung, Subjektpräferenz, syntaktische Rollenhierarchie, syntaktischer/semantischer Parallelismus, kataphorische Bezugnahmen. Im linken Teil der Tabelle sind die einfließenden *Ressourcen* (Wissensquellen) aufgeschlüsselt: OBERF (Zeichenketten an der sprachlichen Oberfläche sowie deren Anordnung), MOLEX (morphologische Analyse und Lexikon, umfassend u.a. Stammformen, Wortarten, ggf. Genus), SELEX (lexikalisch-semantische Information), SYNPAR (Analysemodul zur Bestimmung der syntaktischen Oberflächenstruktur und die zugehörige Grammatik), SEMWI und PRAWI (semantische bzw. pragmatisches Kontextwissen sowie entsprechende Inferenzmechanismen). Ressourcen, die im Allgemeinen einfließen, sind mit "+" gekennzeichnet; solche, die nur in wenigen Fällen einfließen bzw. auf deren Verfügbarkeit ohne entscheidende Einbußen verzichtet werden kann, tragen die Markierung "•".[5]. Im rech-

[5]Um die Darstellung überschaubar zu halten, werden offensichtliche Subsumptionsverhältnisse zwischen Ressourcen in der Tabelle nicht expliziert. Z.B. basiert die Syntaxanalyse (SYNPA) in aller Regel auf einer morpho-

STRATEGIE	RESSOURCEN						BEWERTUNG	
	OBERF	MOLEX	SELEX	SYNPA	SEMWI	PRAWI	RELEV	ALGOR
MORKONGRU		+		•	•		⊕ S	⊕/⊙ S
SYNKONFIG				+			⊕	⊙
SEMSKOPUS				+	+	+	⊖	⊖
INT-FOKUS				+	+	+	⊙ (?) T	⊖
LOK-FOKUS				+			⊖	⊙
DEFKATAPH	+				•		⊙ (?)	⊕/⊙
SEMROLLEN			+	+	•	•	⊙ (?)	⊙/⊖
PRAE-POST					+	+	⊖	⊖
INT-FOKUS				+	+	+	⊙ (?) T	⊖
LOK-FOKUS				+			⊕	⊙
DISTANZ	+						⊕	⊕
TOPIKALIS				+			⊙	⊙
SUBJEKT				+			⊕	⊙
SYNROLLE				+			⊙ (?)	⊙
SYNPARALL				+			⊕	⊙
SEMPARALL			+	+	•		⊕	⊙
KATAMALUS	+						⊕	⊕

Abbildung 4: Ressourcen, Algorithmisierbarkeit und Relevanz einzelner Strategien

ten Teil der Tabelle erfolgt eine graduelle Einstufung von *Algorithmisierbarkeit* (ALGOR) bzw. *Relevanz* (RELEV): "⊕" = "*problemlos*" bzw. "*hoch*", "⊙" = "*unter bestimmten Voraussetzungen*" bzw. "*mittel*", "⊖" = "*auf dem gegenwärtigen Stand der Technik problematisch*" bzw. "*gering*". Eine Einstufung ist mit "(?)" gekennzeichnet, falls die Angabe auf einer vorläufigen Schätzung basiert, die (ggf.) per Evaluation zu verifizieren ist; "S" bzw. "T" steht für Sprach- bzw. Anapherntyp-Abhängigkeit. Mehrfachangaben beziehen sich auf Fälle unterschiedlicher Schwierigkeit, die i.d.R. mit fakultativem Ressourcenbedarf einhergehen.

8 Die Algorithmisierbarkeit der syntaktisch-konfigurationalen Restriktionen unterliegt zusätzlichen Bedingungen

Am Beispiel der syntaktisch-konfigurationalen Bedingungen soll nun expliziert werden, wie die Ergebnisse in der zuvor angegebenen Tabelle zustande kommen. Die Algorithmisierung dieser Restriktionsklasse kann auf der Grundlage der *Bindungstheorie* angegangen werden, einem theoretischen Modell, das als Modul der Chomsky'schen GB-Theorie formuliert wurde. Durch die drei *Bindungsprinzipien A, B und C* für unterschiedliche Typen sprachlicher Ausdrücke werden folgende (hier nur verkürzt wiedergegebenen) Bedingungen postuliert:

Definition 2 (Bindungsprinzipien A, B und C)

(A) Ein Reflexiv- oder Reziprok-Pronomen ist lokal gebunden.

(B) Ein nichtreflexives Pronomen ist nicht lokal gebunden.

logischen Voranalyse (MOLEX).

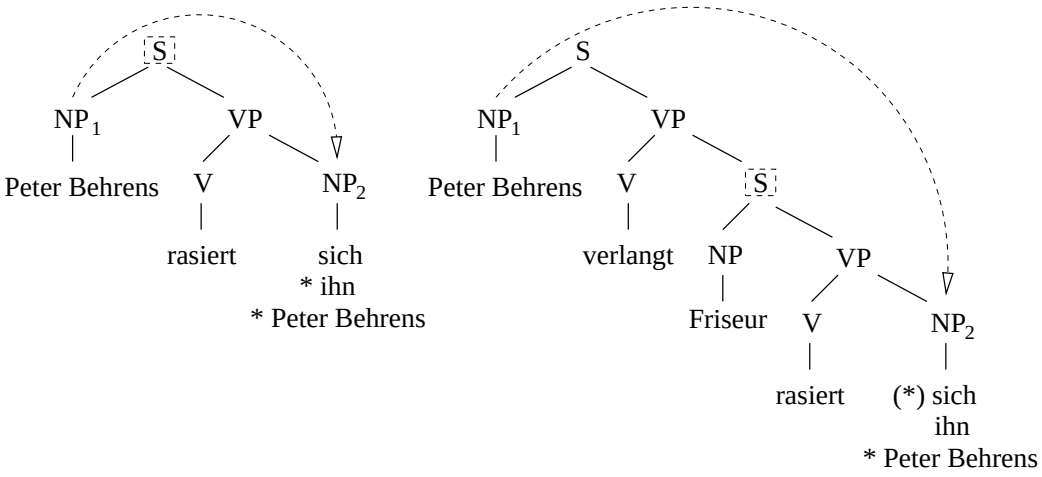

Abbildung 5: Lokale vs. nichtlokale Bindung in den Belegen *(9a)* bzw. *(9b)*

(C) Eine nichtpronominale Anapher ist weder lokal noch nichtlokal gebunden.

Was unter "*lokal*" bzw. "*nichtlokal*" sowie unter Bindung zu verstehen ist, wird unter Rekurs auf Konstrukte der GB-Theorie formal modelliert. Von zentraler Bedeutung ist hierbei, dass die genannten Begriffe unter Rekurs auf Beschreibungen der *syntaktischen Oberflächenstruktur* formalisiert werden.
In Abbildung 5 ist die oberflächenstrukturelle Beschreibung der Belege

(9a) Peter Behrens rasiert sich / ihn / Peter Behrens.
(9b) Peter Behrens verlangt, dass der Friseur sich / ihn / Peter Behrens rasiert.

skizziert. Augenscheinlich ist im Beispiel des Objektsatzes (Beleg (9b)) die Wahl des Hauptsatzvorkommens von "*Peter Behrens*" als Antezedens des anaphorischen Ausdrucks ausschließlich im Falle des nichtreflexiven Pronomens "*ihn*" grammatisch akzeptabel. In der Bindungstheorie wird dieser Sachverhalt per Postulierung eines Lokalitätsbegriffs modelliert, der (u.a.) an den durch finite Verben induzierten S-Knoten (als sog. *Bindende Kategorie*) festmacht. Demnach kommt für das Reflexivum als (unmittelbares) Antezedens ausschließlich ein Vorkommen innerhalb des Komplementsatzes in Frage (Bindungsprinzip A, "*Friseur*" als lokaler Binder). Bindungsprinzip C schließt die Wahl des ausdrucksidentischen Hauptsatzvorkommens für die nichtpronominale Anapher "*Peter Behrens*" aus, da eine nichtlokale Bindungsbeziehung entstehen würde. Die formale Definition der Bindungsrelation rekurriert auf die GB-theoretisch zentrale, ebenfalls an der oberflächenstrukturellen Beschreibung festmachende *K-Herrschafts-Relation*. In dem betrachteten Syntaxbaum etwa besteht eine K-Herrschafts-Beziehung zwischen den Knoten NP1 und NP2 (hervorgehoben durch Pfeil).
Die Algorithmisierung der syntaktisch-konfigurationalen Restriktionen basiert somit auf der Verfügbarkeit *vollständiger* und *unambiger* syntaktischer Beschreibungen; die Strategie SYN-KONFIG ist deshalb im Ressourcen-Teil von Abbildung 4 als abhängig von der Wissensquelle SYNPAR (syntaktisches Parsing) gekennzeichnet. Da algorithmisches Parsing, wie zuvor illustriert, i.d.R. nur *partielle* syntaktische Beschreibungen liefert, ergibt sich somit in Bezug auf

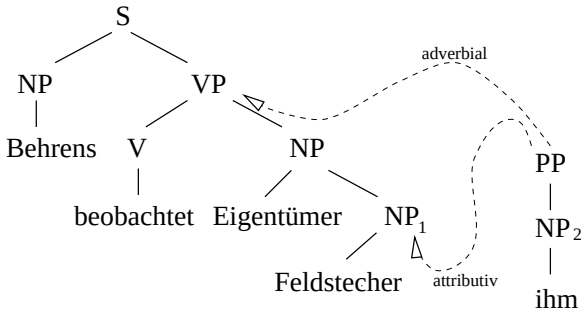

Abbildung 6: Anknüpfungs-Ambiguität in Beleg *(10)*

SYNKONFIG nur eine *eingeschränkte Algorithmisierbarkeit*, zu deren Gewährleistung zusätzliche Maßnahmen zu ergreifen sind. Da andererseits bekannt ist, dass SYNKONFIG einen potenziell hohen Beitrag für die algorithmische Anapherninterpretation leistet, lohnt sich eine nähere Untersuchung der Möglichkeiten einer *robusten*, auch in typischen Anwendungsszenarien operationalen Algorithmisierung der syntaktisch-konfigurationalen Bedingungen.
Nicht zuletzt sind die Bedingungen der Bindungstheorie insofern von hoher Relevanz, als sie einer Theorie *mit universalgrammatischem Anspruch* entspringen, wodurch die *sprachübergreifende Tauglichkeit* als Interpretationsstrategie für das Problem der Anapherninterpretation impliziert wird. Da v.a. auf *linguistischem*, nicht kontextdeterminiertem Wissen aufgebaut wird, kann die solchermaßen generierte referentielle Evidenz als invariant gegenüber hermeneutischen Bedeutungsvariationen angesehen werden. Somit erweist sich die Strategie SYNKONFIG als idealer Kandidat für die Inkorporation in ein Interpretationsverfahren mit *domänenübergreifendem* Skopus.

9 Betreffend die bindungstheoretischen Restriktionen ist die Anforderung der robusten Algorithmisierung erfüllbar

Im typischen Anwendungsszenario einer uneingeschränkt operationalen, massendatentauglichen Koreferenz-Resolution kann allenfalls von der Verfügbarkeit *fragmentarischer* syntaktischer Beschreibungen ausgegangen werden. In Abbildung 6 sind die oberflächenstrukturellen Gegebenheiten im Falle einer aus Sicht des Syntaxanalysealgorithmus ambigen Präpositionalphrase skizziert. Anhand des Belegs

(10) Behrens beobachtet den Eigentümer des Feldstechers mit ihm.

lässt sich veranschaulichen, dass die algorithmische Verifikation der bindungstheoretischen Bedingungen durch die syntaktische Fragmentierung tangiert wird. Die referentielle Identifikation des in der PP enthaltenen Pronomens "*ihm*" mit dem Ausdruck "*Feldstecher*" - formal: die Koindexierung der entsprechenden NP-Knoten des Syntaxbaums - ist im Falle der (hier offenkundig intendierten) adverbialen Interpretation statthaft, bei einer attributiven Interpretation hingegen nicht, da eine unzulässige Konfiguration entstünde, die gegen eine weitere bindungstheoretische

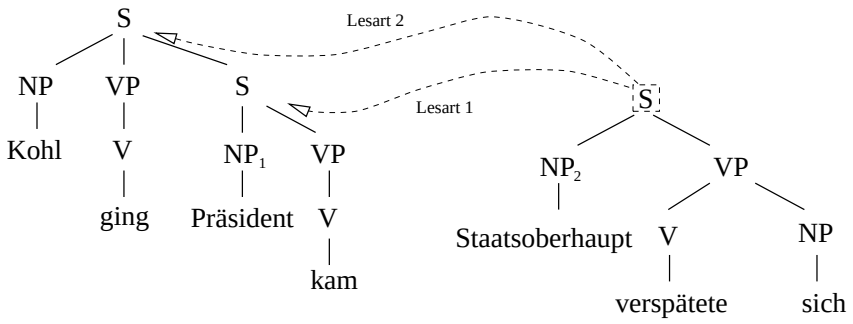

Abbildung 7: syntaktische Fragmente für Beleg *(11)*

Bedingung, den sog. i-über-i-Filter, verstieße.
Eine nähere theoretische Analyse zeigt auf, dass das beschriebene Problem im Allgemeinen nicht alleine auf fehlendes Weltwissen im Parsingvorgang zurückgeführt werden kann, sondern grundsätzlicherer Natur ist. Tatsächlich erweist sich das stillschweigend zugrunde gelegte *sequenzielle* Prozessmodell, derzufolge die referentielle Interpretation auf den Ergebnissen der syntaktischen Analyse *aufbaut*, als inadäquat: In der Chomsky'schen GB-Theorie wird vielmehr eine *Interaktion* der bindungstheoretischen Prinzipien mit den übrigen Modulen der Theorie angenommen, die durch ein (quasi-)*paralleles* Verarbeitungsmodell einzulösen wäre. In einer solchen theoretisch adäquaten Verschränkung der Interpretationsprozesse ist es insbesondere möglich, dass auch umgekehrt die referentielle Interpretation einen Beitrag zur strukturellen Disambiguierung des Oberflächenstrukturbaums leistet.
Dass ein solches (anspruchsvolles) Programm einer im engeren Sinne Menzels[6] robusten Interpretation qua Interaktion von syntaktischer und referentieller Analyse algorithmisch formalisiert werden kann, wurde in Kap. 10 der vorliegenden Dissertation nachgewiesen. Folglich gilt:

These 3 *Von zentraler Bedeutung für die algorithmische Explikation qualitativer Prozesse ist die Entwicklung robuster Analyseverfahren. Der Zielsetzung der robusten Verarbeitung wird durch eine möglichst weitreichende Verschränkung unterschiedlicher Teilprozesse der Analyse Rechnung getragen. Betreffend das Problem der Koreferenzinterpretation ist dieser Anspruch über ein Prozeßmodell einlösbar, in dem die Teilprozesse der syntaktischen und referentiellen Analyse interagieren.*

Auch unter Beibehaltung des programmiertechnisch einfacher umsetzbaren sequenziellen Verarbeitungsmodells ist jedoch im (Regel-)Falle fragmentarischer syntaktischer Beschreibungen oftmals mehr möglich als ein heuristisches Akzeptieren eines vorgeschlagenen, in einem anderen syntaktischen Fragment befindlichen Antezedens. Wenn beispielsweise die Syntaxanalyse für folgenden Beleg

(11) Kohl ging, bevor der Präsident kam, weil das Staatsoberhaupt sich verspätete.

auf Anknüpfungsambiguität des Kausalsatzes befindet, so kann für das Reflexivpronomen "*sich*" dennoch definitiv entschieden werden, dass der Ausdruck "*das Staatsoberhaupt*" das bindungs-

[6]Vgl. Wolfgang Menzel: *Robust Processing of Natural Language*, Universität Hamburg, 1995

[F1] $\surd$ $\{ \ldots F_i = [\ldots bk(\gamma)(\ldots \gamma_{typ\,B} \ldots) \ldots] , \ldots , F_j = [\ldots bk(\alpha)(\ldots \alpha_{typ\,B} \ldots) \ldots] \ldots \}$

[F2] $*$ $\{ \ldots F_i = [\ldots nvk(\gamma)(\ldots \gamma_{typ\,A/B/C} \ldots) \ldots] , \ldots , F_j = [\ldots bk(\alpha)(\ldots \alpha_{typ\,A} \ldots) \ldots] \ldots \}$

[E1a] $\surd$ $\{ \ldots F_d = [\ldots \gamma_{typ\,A/B/C} \ldots] , \ldots , F_e = [\ldots bk(\alpha)(\ldots \alpha_{typ\,B} \ldots) \ldots] \ldots \}$

[E1b] $\surd$ $\{ \ldots F_d = [\ldots \alpha_{typ\,B/C} \ldots] , \ldots , F_e = [\ldots bk(\gamma)(\ldots \gamma_{typ\,B} \ldots) \ldots] \ldots \}$

[E2] $*$ $\{ \ldots F_d = [\ldots \gamma_{typ\,A/B/C} \ldots] , \ldots , F_e = [\ldots bk(\alpha)(\ldots \alpha_{typ\,A} \ldots) \ldots] \ldots \}$

[E3a] $*$ $\{ \ldots F_d = [\ldots \gamma_{typ\,A/B/C} \ldots] , \ldots , F_e = [\ldots \alpha_{typ\,C} \ldots] \ldots \}$, falls gilt:
γ k-beherrscht α unabhängig von der Entscheidung der strukturellen Ambiguität

[E3b] $*$ $\{ \ldots F_d = [\ldots \alpha_{typ\,A/B/C} \ldots] , \ldots , F_e = [\ldots \gamma_{typ\,C} \ldots] \ldots \}$, falls gilt:
α k-beherrscht γ unabhängig von der Entscheidung der strukturellen Ambiguität

[E4] $*$ $\{ \ldots F_d = [\ldots \alpha_{typ\,A} \ldots] , \ldots , F_e = [\ldots nvk(\gamma)(\ldots \gamma_{typ\,A/B/C} \ldots) \ldots]$

Abbildung 8: Regeln für die Bindungsprinzip-Verifikation auf fragmentarischer Syntax

theoretisch einzig zulässige (unmittelbare) Antezedens ist (vgl. Abbildung 7). Diese Beobachtung fußt u.a. auf dem Sachverhalt, dass der den Kausalsatz beschreibende Syntaxbaum lokal abgeschlossen in dem Sinne ist, dass dessen oberster (S-)Knoten bindende Kategorie des Reflexivums ist und somit eine lokale Bindungsdomäne konstituiert, innerhalb derer das Pronomen qua Bindungsprinzip A ein Antezedens haben muss; aus ebendiesem Grund scheiden die weiteren Kandidaten "*Kohl*" und "*Präsident*" sowie ggf. zusätzliche intersententielle Kandidaten aus. Dieses Derivat von Bindungsprinzip A für fragmentarische Syntax lässt sich durch folgende Regel operational explizieren:[7]

$$* \quad \{ \ldots F_d = [\ldots \gamma_{typ\,A/B/C} \ldots] , \ldots , F_e = [\ldots bk(\alpha)(\ldots \alpha_{typ\,A} \ldots) \ldots] \ldots \}$$

Analog lassen sich Regeln für weitere Bedingungen der Bindungstheorie angeben (vgl. Abb. 8).

10 Es kann nun ein robuster Algorithmus zur Koreferenz-Interpretation angegeben werden

Somit lassen sich die bindungstheoretischen Restriktionen auf *fragmentarischen* syntaktischen Beschreibungen, wie sie durch verfügbare robuste Parser generiert werden, algorithmisieren. Eine Reihe weiterer Restriktionen (MORKONGRU, DEFKATAPH) und Präferenzheuristiken (DISTANZ, TOPIKALIS, SUBJEKT, SYNROLLE, KATAMALUS, SYNPARALL, SEMPARALL) lassen sich mit vergleichsweise geringerem Aufwand robust operationalisieren. Der auf diese Weise entstehende dreiphasige Algorithmus (vgl. Anhang A) genügt den Anforderungen

[7] Folgende notationelle Konventionen liegt der Beschreibung der Regelmuster zugrunde: α = zu resolvierende Anapher, γ = Antezedenskandidat; Syntaktische Konstituenten werden durch runde Klammern begrenzt; eckige Klammern heben die Fragmentgrenzen hervor; $bk(X)$ kennzeichnet die Bindende Kategorie eines Knotens X; $nvk(X)$ kennzeichnet den nächstverzweigenden Knoten eines Knotens X im Sinne der Definition des K-Herrschafts-Begriffs; der Index von $X_{typ\,Y}$ spezifiziert die bindungstheoretische Klasse des Vorkommens-Stifters X als Y mit $Y \in \{A, B, C\}$, z.B. $\gamma_{typ\,B}$ ist ein nichtreflexives Pronomen; $\surd/*$ zeigt an, ob das jeweilige Regelmuster die Koindexierung von γ und α zuläßt oder ausschließt.

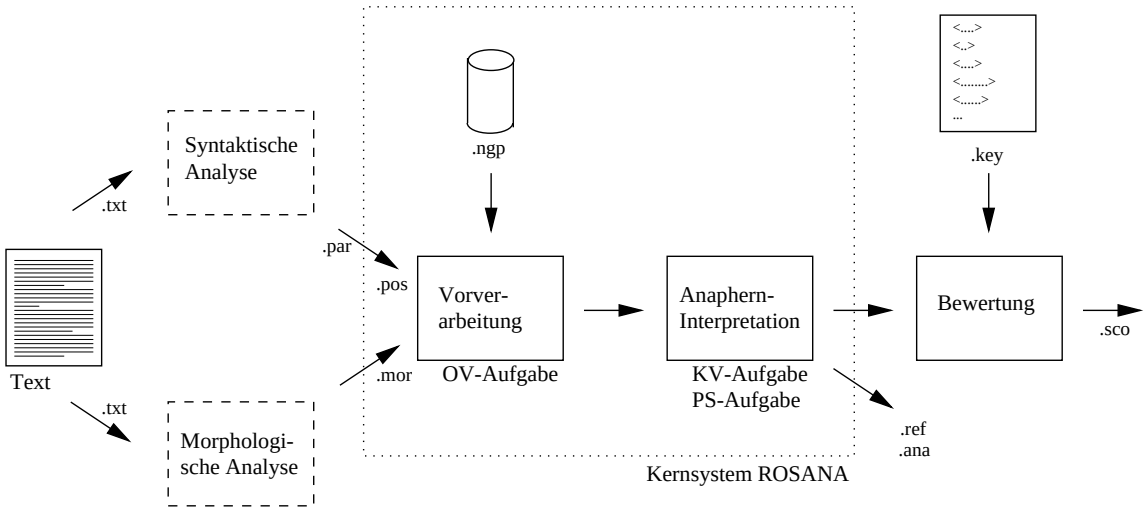

Abbildung 9: ROSANA-System, Verarbeitungsschema von Implementierung und Evaluation

der robusten, massendatentauglichen Verarbeitung, die im Rahmen eines Einsatzes in der Computergestützten Inhaltsanalyse der Sozialwissenschaften bestehen.

11 Design, Implementierung und Evaluation des ROSANA-Systems: Das qualitative Problem "Koreferenzanalyse" ist algorithmisch explizierbar

Dass der auf der beschriebenen, in abgeschwächtem Sinne robusten Operationalisierung der syntaktisch-konfigurationalen Restriktionen basierende Mehrstrategie-Algorithmus in praktischen Anwendungen sehr gute Ergebnisse liefert, wurde per Implementierung und formaler Evaluation des *Anaphernresolutions-Systems ROSANA* auf zwei Korpora, bestehend aus Nachrichtenagenturmeldungen bzw. Operntexten (Inhaltsbeschreibungen zu Mozart-Opern), nachgewiesen. Das Akronym ROSANA steht für

Ro*buste* **s***yntaxbasierte Interpretation* **an***aphorischer* **A***usdrücke*

In Abbildung 9 ist das Verarbeitungsschema des Interpretations- und Evaluationsvorgangs skizziert: Inputs sind morphologische und syntaktische Voranalysen, die extern unter Rückgriff auf kommerziell verfügbare massendatentaugliche Softwarekomponenten (*ENGTWOL* bzw. *Dependency Parser for English*) generiert wurden. Das ROSANA-Kernsystem, das im Rahmen der vorliegenden Dissertation entworfen und implementiert wurde, besteht aus den Komponenten Vorverarbeitung (u.a. umfassend die Ermittlung referenzierender sprachlicher Ausdrücke (OV-Aufgabe)) und eigentliche Anapherninterpretation, in deren Rahmen die Äquivalenzklassen der Koreferenzrelation berechnet (KV-Aufgabe) bzw. nichtpronominale Substitute für Pronominalanaphern ermittelt werden (PS-Aufgabe). Die Evaluation der Interpretationsleistung von RO-

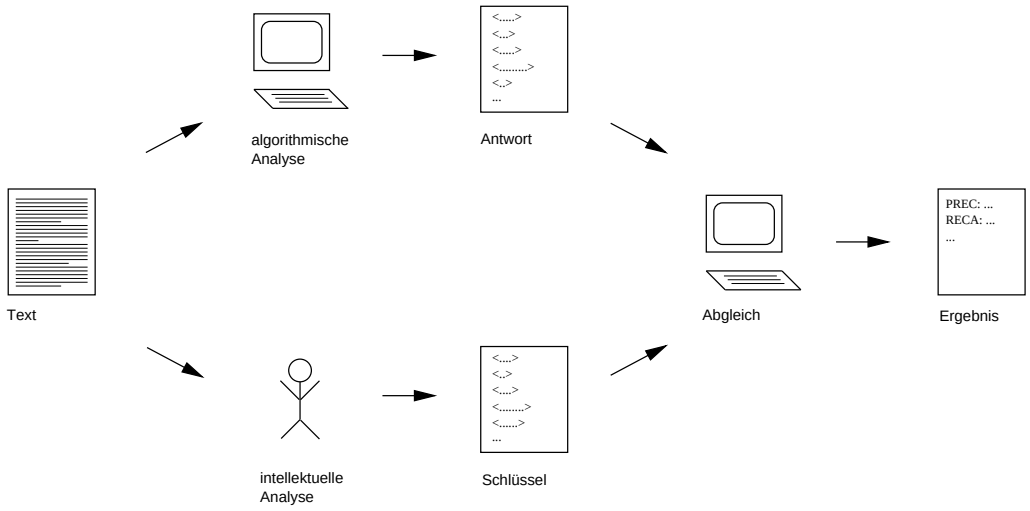

Abbildung 10: Szenario der Evaluation von Systemen der algorithmischen Inhaltserschließung

SANA geschieht per automatischem Abgleich mit intellektuell erzeugten Referenzdaten, sog. Ergebnis-*Schlüssel* (vgl. Abbildung 10).
Um eine geeignete Grundlage zur formalen Evaluation zu schaffen, war es u.a. notwendig, die im Rahmen der *Message Understanding Conferences* entwickelten äquivalenzklassenbezogenen Evaluationsmaße (KV-Aufgabe) zu verfeinern und um die (schwierigere) Aufgabenstellung der Pronominalsubstitut-Bestimmung (PS-Aufgabe) zu ergänzen, die die typischen Anforderungen im Rahmen einer verfeinerten Computergestützten Inhaltsanalyse reflektiert. Ferner wurden geeignete Referenzdaten für die beiden Korpora, umfassend 24712 bzw. 2522 Wörter, intellektuell erzeugt und ein Software-Modul zur Bewertung implementiert.
Anhang B zeigt das Ergebnis der ROSANA-Interpretation einer typischen Nachrichtenagentur-Meldung; in Anhang C sind die Ergebnisse der Evaluation auf dem Korpus von Nachrichtenagenturmeldungen wiedergegeben. In der Bestimmung der *Vorkommen*, d.h. der im Rahmen der referentiellen Interpretation relevanten sprachlichen Ausdrücke (OV-Aufgabe) erzielt ROSANA (*Precision*,*Recall*)-Wertekombinationen von (0.94,0.96) (Nachrichtenagenturmeldungen) bzw. (0.95,0.98) (Operntexte). Für die Interpretationsleistung in der äquivalenzklassenbezogenen KV-Disziplin (Ermittlung der Koreferenzrelation) wurden Ergebnisse von (0.81,0.68) (Nachrichtenagenturmeldungen) bzw. (0.88,0.81) (Operntexte) ermittelt. Für die aus der Anwendungssicht zentralen, da häufigsten und schwieriger zu resolvierenden Personal- und Possessivpronomen in dritter Person bestimmte ROSANA ein korrektes Antezedens in 71 bzw. 76 (Operntexte: 79 bzw. 77) Prozent aller Fälle. Legt man die verschärfte Zielsetzung der Ermittlung nichtpronominaler Antezedenten zugrunde (PS-Disziplin), so erzielt ROSANA Precision-Werte von 68 bzw. 66 (Operntexte: 70 bzw. 76) Prozent. Dass letzteres Problem schwieriger ist, lässt sich u.a. fokustheoretisch begründen.
Die sehr guten Ergebnisse von ROSANA auf den pronomenreiche(re)n Operntexten belegen die domänen- und textgenre-übergreifende Gültigkeit der verwendeten Interpretationsstrategien, die auf der Grundlage eines Trainingskorpus von Pressemeldungen optimiert wurden. Somit ergibt sich ein wichtiges Indiz für die generelle Verwendbarkeit von ROSANA. Nicht Genre oder inhaltliche Domäne, sondern vielmehr oberflächlich-formale Charakteristika der zu verarbeitenden Dokumente wie etwa die relative Häufigkeit von Pronomen erweisen sich als die bestimmenden

Faktoren der Interpretationsqualität.
Die Qualität der Ergebnisse reicht nahe an die im Rahmen der MUCs erzielte Interannotator-Übereinstimmung in der Schlüsselerstellung (81 Prozent) und damit an die Grenze des prinzipiell Messbaren heran. Als von zentraler Bedeutung für die Aussagekraft der Resultate ist zu bewerten, dass die Evaluationskorpora keiner vorherigen Korrektur etwaiger orthographischer und syntaktischer Fehler unterzogen wurden. ROSANA stellt somit den Stand der Forschung auf dem Gebiet der robusten algorithmischen Koreferenz-Interpretation dar, und es gilt ferner:

These 4 *Für den Teilprozeß "Koreferenzanalyse" des qualitativen Prozesses der referentiellen Interpretation ist eine algorithmische Explikation auf der Basis des gegenwärtigen Stands der Forschung möglich.*

12 Konklusion

Die Ergebnisse der Untersuchungen belegen, dass der Kategorisierungsschritt der Computergestützten Inhaltsanalyse auf der Basis verfügbarer Technik algorithmisch verfeinert werden kann. Einmal mehr zeigt sich, dass das etwa seitens der philosophischen Hermeneutik i.d.R. vertretene Postulat der generellen Unmöglichkeit qualitativer Analysen durch Computer als widerlegt anzusehen ist. Für einen wichtigen Teilbereich des Problems der Interpretation indexikalischer Ausdrücke, das von sozialwissenschaftlichen und philosophischen Kritikern der Künstlichen Intelligenz immer wieder als Beispiel prinzipieller algorithmischer Nichtexplizierbarkeit herangezogen wird, wurde der Nachweis der algorithmischen Lösbarkeit erbracht. Die Ergebnisse der Dissertation legen folglich eine differenzierte Neubewertung der Möglichkeiten einer computergestützten qualitativen Analyse nahe:

These 5 *Algorithmen stellen die geeignete Basis zur formalen Explikation qualitativer Prozesse der Textinhaltserschließung dar. Aus den Postulaten der hermeneutischen Philosophie ist keine generelle Inadäquatheit computergestützter qualitativer Analysen ableitbar, sondern lediglich die Aussage, dass die algorithmische Explikation qualitativer Prozesse im Einzelfall schwierig sein kann. Die Frage nach der Adäquatheit von Computeranalysen kann somit nicht a priori negativ beschieden werden; sie ist vielmehr vor dem Hintergrund konkreter inhaltsanalytischer Probleme sowie mit Blick auf den State-of-the-Art maschineller Textanalyseverfahren zu beantworten.*

Erneut ist hierbei zu betonen: Die Rolle des Computers in der qualitativen Analyse wird als die eines *Werkzeugs* gesehen, das bestimmte, vom Menschen als inhaltserschließend verstandenen Symboltransformationen auf einer rein *syntaktischen* Ebene vollzieht; die wesentlichen Aufgaben des *sinnzuweisenden Verstehens* der computergenerierten Ergebnisse vor dem Hintergrund der jeweiligen Forschungsfrage verbleibt im Verantwortungsbereich des Forschers.
Die obigen Ausführungen verdeutlichen die *interdisziplinäre Verortung* des Programms einer inhaltsorientierten, die Einzelwortebene transzendierende Verfeinerung der Computergestützten Inhaltsanalyse: Unter Rekurs auf Elemente der *(Text-)Linguistik* erschließen sich die Defizite des wörterbuchbasierten Analysemodells; mit Hilfe moderner *computerlinguistischer* Analysestrategien sind diese Defizite transzendierbar; *kognitionstheoretische* Forschungsergebnisse bilden die

Grundlage adäquater, im Speziellen domänenübergreifend gültiger referentieller Interpretationsstrategien; Techniken der *Informatik* ermöglichen die Realisierung konkreter Software-Systeme wie im Rahmen der vorliegenden Dissertation ROSANA; den globalen Bezugsrahmen jedoch bilden die spezifischen Bedingungen der Inhaltsanalyse als zentrale Methode der *Sozialwissenschaften*, weshalb sich die bekannten Anforderungen in der Verarbeitung von Massentexten - Robustheit und domänenübergreifende Anwendbarkeit - in allen übrigen Ebenen der Untersuchung niederschlagen. So wurde u.A. die Möglichkeit einer Verfeinerung des Kategorisierungsschritts erarbeitet, die auf thematischen Schablonen basiert, für deren Erschließung im Kontext der *Message Understanding Conferences* massendatentaugliche Computerprogramme realisiert wurden, die - qua Separierung *linguistischen* Wissens - schnell auf neue Anwendungsdomänen portierbar sind. In Bezug auf das Problem der Koreferenz-Interpretation schlug sich die Robustheitsbedingung in der Entwicklung eines Algorithmus nieder, in der syntaktische und referentielle Analyseebenen auf theoretisch adäquate Weise interagieren; der Rekurs auf das universalgrammatische Modul der Bindungstheorie als zentrale Interpretationsstrategie sorgt auch hier für die domänenenübergreifende Anwendbarkeit des Algorithmus.
Betreffend ROSANA ergeben sich folgende Perspektiven für eine weitere Verfeinerung:

1. Eine vergleichende Gegenüberstellung der Interpretationsleistung auf den beiden Evaluationskorpora (Nachrichtenagenturmeldungen, Operntexte) legt die *textgenre-spezifische Konfiguration der Präferenzheuristiken* in Entsprechung zu bestimmten formalen Charakteristika der Kohäsionsstruktur nahe; dies könnte auf der Grundlage einer Computeranalyse intellektuell annotierter Schlüsseltexte erfolgen.

2. *Verfeinerte Behandlung nichtpronominaler Anaphern* (Brückenanaphern) unter Rekurs auf statistische Information (als heuristisches Substitut für Weltwissen), die durch Analyse annotierter Korpora gewonnen wird.

Zumindest für Pronomen bewegen sich die Ergebnisse des Kernsystems ROSANA bereits jetzt *nahe am erzielbaren Optimum*, da weit mehr als die Hälfte aller Fehlentscheidungen unmittelbar durch Fehler in der morphologischen oder syntaktischen Voranalyse bedingt sind.
Im Rahmen des Forschungsprogramms einer verfeinerten Computergestützten Inhaltsanalyse in den Sozialwissenschaften besteht indes der nächste und zentrale Schritt in der *Untersuchung von Möglichkeiten, Reichweite und Grenzen der neuen Techniken in konkreten inhaltsanalytischen Anwendungsszenarien.* Ausschließlich auf der Basis konkreter Experimente läßt sich entscheiden, wieviel etwa durch die erweiterte Themenerschließung à la MUC oder durch die Einbeziehung einer Komponente zur Pronomeninterpretation gewonnen werden kann. Letztlich spielt neben den rein technischen Gesichtspunkten auch die Akzeptanz der verfeinerten Verfahren durch den Benutzer eine entscheidende Rolle.

A ROSANA-Algorithmus zur Koreferenz-Interpretation

1. Für jede anaphorische Okkurrenz Y, **ermittle die Menge der möglichen Antezedenten** X:
   (a) Anapherntypabhängige Anwendung von Kongruenzrestriktionen:
      - (MORKONGRU) Falls Y pronominal ist, verifiziere die morphologische Kongruenz mit X.
      - (ZEICHKETT) Falls Y nichtpronominal, d.h. Name oder definite NP ist, verifiziere Namens- bzw. sonstige Übereinstimmung der lexikalischen Stammform mit X.

   (b) (SYNKONFIG) Falls der Antezedenskandidat X intrasententiell ist, verifiziere, dass die Bindungsbedingungen von Y *und* X erfüllt sind: für die vorgeschlagene Koindexierung
      i. verifiziere, dass das Bindungsprinzip von Y *konstruktiv erfüllt* ist,
      ii. verifiziere, dass das Bindungsprinzip von X *nicht verletzt* ist,
      iii. verifiziere, dass keine i-über-i-Konfiguration vorliegt.

   (c) (SEMROLLEN) Verifiziere die selektionale Verträglichkeit des Kandidaten X mit der Rolle, die das anaphorische Vorkommen Y ausfüllt.

   (d) (DEFKATAPH) Falls Y ein Typ-B-Pronomen ist, der Antezedenskandidat X intrasententiell ist und X bezüglich der Oberflächenanordnung auf Y *folgt* (d.h. es liegt *kataphorischer* Wiederaufgriff vor), dann verifiziere, dass X *definit* ist.

2. **Bewertung und Sortierung nach Plausibilität**:
   (a) Für jedes verbleibende Paar (Y_i, X_j) von Anapher und Antezedenskandidat: ermittle - in Abhängigkeit des Typs von Y_i - einen numerischen Plausibilitätsgrad $v(Y_i, X_j)$, der X_j relativ zu Y_i bewertet, unter Rückgriff auf die Faktoren DISTANZ, TOPIKALIS, SUBJEKT, SYNROLLE, KATAMALUS, SYNPARALL und SEMPARALL.

   (b) *(lokales Sortieren)* Für jede Anapher Y: sortiere ihre individuellen Antezedenskandidaten X_j gemäß absteigender Plausibilität $v(Y, X_j)$.

   (c) *(globales Sortieren)* Sortiere die Anaphern Y gemäß absteigender Plausibilität ihrer individuell besten Antezedenskandidaten.

3. **Antezedensauswahl**: betrachte die Anaphern Y in der in Schritt 2c ermittelten Reihenfolge. Schlage die Antezedenskandidaten $X(Y)$ in der in Schritt 2b ermittelten Reihenfolge vor. Wähle $X(Y)$, falls keine Interdependenz besteht, d.h. genau dann wenn
   (a) (MORKONGRU) die morphosyntaktischen Eigenschaften von Y und $X(Y)$ noch kompatibel sind,

   (b) (SYNKONFIG) für alle Okkurrenzen Z_1 und Z_2, die dem Diskursreferenten zu $X(Y)$ bzw. Y im *aktuellen* Lauf der Analyse zugeordnet wurden, gilt: die Koindexierung von Z_1 und Z_2, die transitiv aus der Wahl von $X(Y)$ als Antezedens für Y resultiert, *verletzt nicht* die Bindungsprinzipien und fällt nicht unter den i-über-i-Filter (Der Spezialfall $Z_1 = X(Y) \wedge Z_2 = Y$ braucht nicht erneut betrachtet zu werden.),

   (c) Y und $X(Y)$ noch unterschiedlichen Diskursreferenten zugeordnet sind (Vermeidung zyklischer Wiederaufgriffsrelationen).

Basisalgorithmus

1. Für jede anaphorische Okkurrenz Y, ermittle die Menge der möglichen Antezedenten X:
   ...
   (b) (F-SYNKONFIG) Falls der Antezedenskandidat X intrasententiell ist, verifiziere, dass die Bindungsbedingungen von Y *und* X erfüllt sind.
   - Falls X und Y in demselben oberflächenstrukturellen Fragment vorkommen, so verfahre gemäß Schritt 1b des Basisalgorithmus
   - **Falls X und Y in unterschiedlichen Fragmenten vorkommen**, so
     i. versuche, die *Bindungsprinzipien* von X und Y durch Anwendung eines der Regelschemata definitiv zu verifizieren:
        – ist eine Einbettungsbeziehung zwischen den beiden Fragmenten bekannt, so versuche Regelschemata [E1a], [E1b], [E2], [E3a], [E3b] und [E4];
        – ansonsten versuche Regeln [F1] und [F2].

        Falls keines der Regelschemata anwendbar ist, so *nimm heuristisch an*, dass die Bindungsprinzipien von X und Y *erfüllt* sind;
     ii. verifiziere ferner, dass keine (definitive) *i-über-i-Konfiguration* vorliegt:
        – im Falle einer Einbettungsbeziehung über die Regelschemata [IEa], [IEb];
        – ansonsten über Regel [IF].

   ...
2. Bewertung und Sortierung nach Plausibilität:
   ...
   (a') (Plausibilitätsbewertung ff.) Für jedes Paar (Y_i, X_j) von Anapher und Antezedenskandidat, für das in Schritt 1(b)i die Erfüllung der Bindungsprinzipien *heuristisch angenommen* wurde: reduziere die Plausibilitätsbewertung $v(Y_i, X_j)$.
   ...
3. Antezedensauswahl: betrachte Anaphern Y und jeweilige Antezedenskandidaten $X(Y)$ in derselben Reihenfolge wie in Schritt 3 des Basisalgorithmus. Wähle $X(Y)$, falls keine Interdependenz besteht, d.h. genau dann wenn
   ...
   (b) (F-SYNKONFIG) für alle Okkurrenzen Z_1 und Z_2, die dem Diskursreferenten zu $X(Y)$ bzw. Y im *aktuellen* Lauf der Analyse zugeordnet wurden, gilt: die Koindexierung von Z_1 und Z_2, die transitiv aus der Wahl von $X(Y)$ als Antezedens für Y resultiert, *verletzt nicht* die Bindungsprinzipien und fällt nicht unter den i-über-i-Filter -
   - falls Z_1 und Z_2 in demselben oberflächenstrukturellen Fragment vorkommen, so verfahre gemäß der näheren Spezifikation von Schritt 3b des Basisalgorithmus (nichtkonstruktive Verifikation der Bindungsprinzipien qua Ausblendung von BP A);
   - **falls X und Y in unterschiedlichen Fragmenten vorkommen**, so verfahre wie in den Schritten 1(b)i und 1(b)ii, wobei jedoch die Regelschemata [F2], [E2] und [E4], die BP A konstruktiv verifizieren, nicht angewendet werden.

   (Der Spezialfall $Z_1 = X(Y) \wedge Z_2 = Y$ braucht nicht erneut betrachtet zu werden.)
   ...

Der robuste ROSANA-Algorithmus

B ROSANA-Interpretation einer Pressemeldung

A Punch is Sometimes as Good as a Speech in Taiwanese Democracy.
Taiwanese legislators are finding that good oratory is not the only
skill needed to survive in Taiwan's blooming democracy: a powerful
right-hook also helps. Opposition lawmaker Huang Chao-hui was back
at work Wednesday after being felled by a punch in the latest of the
legislative brawls that have marked the island's transition from virtual
dictatorship to democracy since 1987. Huang, of the Democratic Progressive
Party, had sought hospital treatment for suspected concussion Tuesday after
Lin Ming-yi of the ruling Nationalists punched him during a debate. Lin
said he could not stomach Huang's taunts that Nationalist lawmakers have
recently started attending more legislative sessions only to try to ensure
victory in elections later this year. Lin later apologized for his violent
outburst, and Huang was released from a hospital with bruises on his head.
On Monday, more than 10 lawmakers traded punches during a brawl started when
opposition New Party legislator Ju Gau-jeng, nicknamed ``Rambo'' by Taiwanese
newspapers, jumped onto the legislative speaker's desk. Tensions are running
high in the legislature because lawmakers are debating a bill to govern a
presidential election next March. The vote is seen as a milestone in Taiwan's
march to democracy because it will mark the first time that the island's
president is elected by universal suffrage.

```
A.R.-ENTSCHEIDUNGEN
++    2 25894: democracy  --->    1 25872: democracy
++    3 25918: punch      --->    1 25862: punch
++    3 25926: that       --->    3 25925: brawl
++    3 25936: democracy  --->    2 25894: democracy
++    4 25966: him        --->    4 25942: huang
++    5 25975: he         --->    5 25973: lin
++    5 25979: huang      --->    4 25942: huang
++    6 26005: lin        --->    5 25973: lin
++    6 26009: his        --->    6 26005: lin
++    6 26014: huang      --->    5 25979: huang
+-    6 26023: his        --->    6 26009: his
++    9 26095: taiwan     --->    2 25892: taiwan
+-    9 26096: march      --->    8 26084: march
++    9 26098: democracy  --->    3 25936: democracy
++    9 26100: it         --->    9 26088: vote
++    9 26108: island     --->    3 25930: island

NICHTPRONOMINALE SUBSTITUTION
++    3 25926: that       ==    3 25925: brawl
++    4 25966: him        ==    4 25942: huang
++    5 25975: he         ==    5 25973: lin
++    6 26009: his        ==    6 26005: lin
+-    6 26023: his        ==    6 26005: lin
++    9 26100: it         ==    9 26088: vote
```

C Ergebnisse ROSANA-Evaluation, Korpus "*Presse-Meldungen*"

```
EVALUATIONSERGEBNIS OKKURRENZEN:
 - NUR ANA: 243
 - NUR KEY: 150
 - ANA UND KEY: 3831
=> PRECISION: 0.9404
=> RECALL: 0.9623

-----------------------------------------------

 A.R.-ERGEBNIS-PARTITIONIERUNG
 - SCHNITTE: 256
 - MOEGLICH: 1334
=> PRECISION: 0.8081

 KEY-PARTITIONIERUNG
 - SCHNITTE: 496
 - MOEGLICH: 1572
=> RECALL: 0.6845

 ANKNUEPFUNGS-ENTSCHEIDUNGEN

             | PRECIS |  ++  |  +-  |  +?  |  +_  |  +*  |  ?+  |  ?_  |
 ------+-----+--------+------+------+------+------+------+------+------+
 PRON | PE-3 | 0.7143 |  145 |   48 |   10 |    1 |    0 |   18 |    0 |
      | PE12 | 0.9474 |   18 |    1 |    0 |    7 |    6 |    0 |    0 |
      | PO-3 | 0.7634 |  100 |   28 |    3 |    0 |    0 |    0 |    0 |
      | PO12 | 1.0000 |    3 |    0 |    0 |    1 |    1 |    0 |    0 |
      | REFL | 1.0000 |    3 |    0 |    0 |    1 |    0 |    0 |    0 |
      | RELA | 0.7789 |   74 |   18 |    3 |    6 |    0 |    7 |    4 |
      +------+--------+------+------+------+------+------+------+------+
             | 0.7555 |  343 |   95 |   16 |   16 |    7 |   25 |    4 |
 ------+-----+--------+------+------+------+------+------+------+------+
 NOMN | VNOM | 0.7014 |  357 |  136 |   16 | 1973 |   43 |   31 |  133 |
      | NAME | 0.9390 |  308 |   15 |    5 |  368 |    5 |    5 |   28 |
      +------+--------+------+------+------+------+------+------+------+
             | 0.7945 |  665 |  151 |   21 | 2341 |   48 |   36 |  161 |
 ------+-----+--------+------+------+------+------+------+------+------+

 DURCHSCHNITT JE ANAPHER
=> PRECISION: 0.7808

-----------------------------------------------

 NICHTPRONOMINALE SUBSTITUTION

      | PRECIS | RECALL |  ++  |  +-  |  +?  |  +_  |  +*  |  ?+  |  ?_  |
 -----+--------+--------+------+------+------+------+------+------+------+
 PE-3 | 0.6766 | 0.6667 |  136 |   54 |   11 |    3 |    0 |   18 |    0 |
 PE12 | 0.9091 | 0.3846 |   10 |    1 |    0 |   15 |    6 |    0 |    0 |
 PO-3 | 0.6641 | 0.6641 |   87 |   39 |    5 |    0 |    0 |    0 |    0 |
 PO12 | 1.0000 | 0.5000 |    2 |    0 |    0 |    2 |    1 |    0 |    0 |
 REFL | 1.0000 | 0.7500 |    3 |    0 |    0 |    1 |    0 |    0 |    0 |
 RELA | 0.7667 | 0.6832 |   69 |   18 |    3 |   11 |    0 |    7 |    4 |
 -----+--------+--------+------+------+------+------+------+------+------+

 DURCHSCHNITT JE PRONOMEN
=> PRECISION: 0.7009
=> RECALL: 0.6532
```

D Definitionen einiger zentraler Begriffe

D.1 Textlinguistik

Definition (Kohäsion) *Ein Text heißt kohäsiv, wenn er einen durch lexikalisch-semantische oder formalgrammatische Mittel hergestellten Zusammenhang aufweist.*

Definition (Kohärenz) *Ein Text heißt kohärent, wenn er relativ zu einem bestimmten Verwendungszusammenhang pragmatisch sinnvoll erscheint.*

Definition (Thema) *Ein Thema ist ein textuell etablierter (Teil-)Kontext, der im Text mehr als einmal instanziiert, d.h. wiederaufgegriffen wird.*

D.2 Referentialität / Koreferenz

Definition (Anapher, Antezedens, anaphorischer Wiederaufgriff) *Eine Anapher ist ein sprachlicher Ausdruck (evtl. eine Ellipse), der als inhaltliche Entität zu interpretieren ist, die bereits durch (mindestens) ein weiteres sprachliches Element, das Antezedens, entweder direkt oder indirekt in den Diskurs eingeführt wurde.*
Auf der Ebene des sprachlichen Ausdrucks - zwischen Anapher und Antezedens - besteht eine Kohäsionsrelation, die als anaphorischer Wiederaufgriff bezeichnet wird.
Referenz-Identität bzw. Koreferenz stellt nur einen Spezialfall anaphorischen Wiederaufgriffs dar. Die Relation kann auch indirekt sein (konzeptueller Aufgriff des Sinns, Summation, Teil-Ganzes usw.) und sich ggf. erst unter Rekurs auf Weltwissen erschließen.

Definition (Vorkommen (Okkurrenz)) *Ein Vorkommen (auch Okkurrenz genannt) ist ein sprachlicher Ausdruck, der "spezifiziert", d.h. der ein Spezifikat in der projizierten Welt symbolisiert.*

Definition (Diskursreferent) *Der Begriff Diskursreferent ist ein Synonym für das Spezifikat, das ein sprachlicher Ausdruck eines bestimmten Textes (Diskurses) in der projizierten Welt spezifiziert (symbolisiert).*

D.3 Syntaktisch-konfigurationale Bedingungen

Definition (Relation K-Herrschaft) *Knoten N_1 k-beherrscht N_2 genau dann, wenn folgende Bedingungen erfüllt sind:*

1. *Der nächste verzweigende Knoten, der N_1 dominiert, dominiert auch N_2.*
2. *Weder wird N_2 von N_1 dominiert, noch wird N_1 von N_2 dominiert, noch ist $N_1 = N_2$.*

Definition (Bindungsrelation) *Knoten N_1 bindet Knoten N_2 genau dann, wenn N_1 und N_2 koindexiert sind und N_2 von N_1 k-beherrscht wird.*

Definition (Bindungsprinzipien A, B und C)

(A) Ein Reflexiv- oder Reziprok-Pronomen ist in seiner Bindenden Kategorie gebunden.

(B) Ein nichtreflexives Pronomen ist in seiner Bindenden Kategorie frei (d.h. nicht gebunden).

(C) Eine nichtpronominale Anapher[8] *ist in jeder Domäne frei (d.h. nicht gebunden).*

Definition (Bindende Kategorie) *Knoten N_1 ist Bindende Kategorie von Knoten N_2 genau dann, wenn N_1 der nächste Knoten ist, der N_2 dominiert und folgende Bedingungen erfüllt:*

1. *N_1 dominiert ein SUBJEKT*[9] *N_3, das N_2 k-beherrscht.*
2. *Falls N_2 ein Reflexiv- oder Reziprok-Pronomen ist, fällt eine Koindexierung von N_2 und N_3 nicht unter den i-über-i-Filter.*

Definition (i-über-i-Filter) *Oberflächenstrukturelle Koindexierungs-Konfigurationen von der Form $[\alpha \ldots [\beta \ldots]_i]_i$ sind ausgeschlossen.*

[8] In der Theorie findet sich hier der Begriff des *R-Ausdrucks*, d.h. des "selbsttätig referenzierenden" sprachlichen Ausdrucks. In der Literatur finden sich unterschiedliche Definitionen; im Rahmen dieser Arbeit wird die Sichtweise Haegemans zugrundegelegt, die hierunter neben *Namen* u.a. auch sonstige *nichtpronominale NP* subsumiert (vgl. Haegeman: *Introduction to Government and Binding Theory*, Oxford, 1991).

[9] Für die Entwicklung eines prinzipiellen Verständnisses, das den Zielsetzungen der vorliegenden Arbeit gerecht wird, ist es hinreichend, diesen Begriff des SUBJEKT (*"big subject"*) vereinfachend mit dem zuvor erläuterten Begriff des verallgemeinerten Subjekts gleichzusetzen. Im Rahmen einer präziseren Darstellung der Bindungstheorie wäre jedoch feiner zu unterscheiden zwischen finiten und infiniten Konstruktionen, für die sich empirische Unterschiede nachweisen lassen, die durch eine entsprechende Verfeinerung des SUBJEKT-Begriffs theoretisch erfaßt werden können: in finiten Konstruktionen fällt SUBJEKT mit einer künstlichen Kategorie AGR zusammen, die vom entsprechenden S-Knoten direkt dominiert wird und wie das gewöhnliche Subjekt den Inhalt von S k-beherrscht; in allen anderen Fällen koinzidiert es mit dem herkömmlichen Subjekt (vgl. von Stechow und Sternefeld: *Bausteine syntaktischen Wissens*, 1988).