Burghard Rieger

Theorie der unscharfen Mengen und empirische Textanalyse*

Mathematisch-statistische Untersuchungen im Bereich von Sprache und Texten sind allemal problematisch, besonders aber dann, wenn sie nicht nur Buchstaben, Silben und Wörter zählen, sondern Einsichten zu vermitteln suchen in so undurchsichtige Prozesse, wie es die Bedeutungskonstitution in der natürlichen Sprache immer noch ist. Ganz offensichtlich treten sie damit in eine gewisse Konkurrenz zu formalen, vornehmlich sprachlogisch ausgerichteten Ansätzen der linguistischen Theorienbildung, für die die Mathematisierung nicht weniger charakteristisch ist: den Semantikmodellen der algebraischen Linguistik:

> Da der spezifische Charakter der Statistik in den quantitativen Angaben liegt, ist die *statistische Linguistik* der kontrollierten Gewinnung von Erfahrung und dem Testen von Hypothesen und Theorien zugeordnet, während der Entwurf von Modellen und ihre formale Kritik in den Bereich der *algebraischen Linguistik* gehört.[1]

Akzeptiert man diese Unterscheidung von algebraischer und statistischer Linguistik, die Hellmut Schnelle (1968) als die beiden Teildisziplinen jenes *exakten* Ausschnitts der Linguistik bezeichnete, der durch Theorie und Experiment gebildet wird, dann kommt man nicht umhin, deren bis heute nahezu unvermitteltes Nebeneinander zu konstatieren.

Der Idealvorstellung jedenfalls einer durch wechselseitige Überprüfung und Kontrolle dieser beiden Teildisziplinen sich ständig korrigierenden Theorienbildung in der Linguistik ist man heute kaum näher als 1968. Dennoch – scheint mir – steht mit der Theorie der unscharfen Mengen inzwischen eine Art *Gelenkstück* bereit, das, *numerisch* flexibel und *formal* befriedigend, algebraische Strukturen einerseits mit empirischen Daten andererseits zu verknüpfen gestattet und möglicherweise doch einen Schritt in Richtung auf eine mathematisch-empirische Sprach- bzw. Textwissenschaft zu tun erlaubt.

Ich möchte im folgenden Ansätze zu einer mathematisch-statistischen Methode, der Bedeutungsanalyse natürlich-sprachlicher Texte vorstellen, deren Resultate sich als *unscharfen* Mengen darstellen lassen und zu einem formalen Semantikmodell beitragen sollen, für das das Phänomen der Vagheit und Verschwommenheit natürlich-sprachlicher Bedeutung konstitutiv ist.

Nach einigen kurzen, allgemeinen Bemerkungen zum Problem der Vagheit natürlich-sprachlicher Bedeutung (1), werden die empirische (2.1) und theoretische (2.2) Angemessenheit ihrer Beschreibung sowie die theoretischen (3.1) und empirischen (3.2) Bedingungen ihrer Analyse diskutiert. Vor diesem Hintergrund wird dann versucht, den Gang der Untersuchung und auch deren abstraktere, formale Seite möglichst anschaulich darzustellen (4.1) und anhand von beispielhaften Ergebnissen zu illustrieren (4.2). Abschließende Bemerkungen möchten sodann (5) noch auf einige mögliche Konsequenzen in der Anwendung wie Theorienbildung hinweisen.

*Mit Unterstützung der Deutschen Forschungsgemeinschaft (DFG).

[1] Schnelle, H.: – Methoden mathematischer Linguistik. In: Enzyklopädie der geisteswissenschaftlichen Arbeitsmethoden, 4. Lieferung: Methoden der Sprachwissenschaft, München/Wien (Oldenbourg) 1968, 135-160; 137 (Hervorh. v. Verf.).

1 Die Vorstellung von der Vagheit, Verschwommenheit und Unschärfe natürlich-sprachlicher Bedeutung ist eine für die Sprachwissenschaft wie Sprachphilosophie schon ehrwürdige Einsicht. Sie hat lange alle Formalisierungsversuche in der Semantik als fruchtlos erscheinen lassen und ihre Dominanz erwies sich auch dann noch, als Bereiche der Syntax schon erfolgreich formalisiert wurden, – erfolgreich freilich nur um den Preis der Ausklammerung des semantischen Aspekts.

Inzwischen ist die Semantik ins Zentrum nicht nur der linguistischen Forschung gerückt, und es fehlt nicht an formalen Ansätzen auch auf diesem Gebiet. Dabei lassen sich – grob gesprochen – zwei Richtungen unterscheiden: einmal eine zunehmend sprachlogisch, Theorie-orientierte linguistische Semantik, die zu formalen Modellbildungen kommt, in denen sie expliziert, wie der logische oder *ideale* Sprecher beim Konstituieren von Bedeutungen verfahren würde oder doch verfahren sollte, und zum anderen eine empirisch-praktisch ausgerichtete Bedeutungsanalyse, die eher Methoden-orientiert ist und herauszufinden sucht, wie wirkliche Sprecher in konkreten Situationen tatsächlich verfahren, wenn sie Bedeutung konstituieren.

Linguistische Theoretiker scheinen sich dabei mit den Logikern zumindest darin einig zu sein, daß (deklarativen) natürlich-sprachlichen Aussagen einer der beiden Wahrheitswerte *wahr* oder *falsch* bzw. in dreiwertig-logischen Systemen auch noch der Wert *unbestimmt* zukommen könne. Logiker führen die Wahrheitsbedingungen normalerweise in Ausdrücken der klassischen Mengentheorie ein. Danach ist die Aussage "Ein Rotkehlchen ist ein Vogel" ($\forall$ *r* Vogel(r)) *wahr* genau dann, wenn alle mit *Rotkehlchen* zutreffend bezeichneten Tiere Element der Klasse derjenigen Tiere sind, die mit *Vogel* bezeichnet wird. Wird nun in gleicher Weise die Zuweisung von Wahrheitswerten auf natürlichsprachliche Aussagen übertragen, muß vorausgesetzt werden, daß die verwendeten Wörter und Begriffe, wie logische Symbole, auf Mengen referieren, über deren zugehörige Elemente und ihre Äquivalenz Eindeutigkeit herrscht, die mithin Klassen bilden, deren Grenzen scharf sind.

Da über die Angemessenheit bzw. Unangemessenheit einer solchen Voraussetzung nicht mehr formal innerhalb des logischen Systems entschieden werden kann, das diese Voraussetzung schon macht, bleiben Fragen nach seiner Adäquatheit, wenn auf dieser Ebene überhaupt gestellt, unbeantwortbar. Sie lassen sich nur *empirisch* beantworten und dies nur *relativ zum Anwendungsbereich.*

2.1 Ich möchte deswegen hier die Ergebnisse einer empirischen Untersuchung von Eleanor Rosch Heider mitteilen, die im Rahmen der experimentell arbeitenden Psychologie angestellt und von Lakoff (1973) referiert wurde.[2] Sie liefert eine Reihe von Befunden, die als empirisch belegte Indizien der Unangemessenheit gelten können, die Konstitution natürlich-sprachlicher Begriffsklassen adäquat als Klassen im mengentheoretisch-logischen Sinne zu rekonstruieren. Durch Tests sollte heraus gefunden werden, ob Sprachteilhaber die Zugehörigkeit bestimmter ihnen geläufigen Wörter und deren Bedeutungen zu bestimmten ihnen geläufigen Begriffen als Ja/Nein-entscheidbar oder eher als graduell empfinden.

[2]Lakoff, G.: – Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. In: Journal of Philosophical Logic 2 (1973), 458-508; 458 ff.

Gebeten, eine Anzahl vorgegebener Tiernamen (wie *Huhn, Kuh, Gans, Adler, Rotkehlchen,* etc.) danach zu ordnen, inwieweit die mit ihnen benannten Tiere dem Begriff bzw. der *Idealvorstellung* von *Vogel* entsprächen, lieferten die Probanden weder eine einfache Zweiteilung in Vogel/Nicht-Vogel, noch eine völlig regellose Zuordnung. Dagegen entstand eine durch hohe intersubjektive Übereinstimmung ausgezeichnete Rangordnung

Rotkehlchen
Adler
Huhn, Ente, Gans
Pelikan, Pinguin
Fledermaus

wonach *Rotkehlchen* als typischster Vogel, *Adler* als ein Raubvogel schon weniger, *Huhn, Ente* und *Gans* als noch weniger typisch erschienen, während *Fledermaus* kaum noch und *Kuh* und *Pferd* überhaupt nicht mehr als zugehörig empfunden wurden.

Im allgemeinen lassen sich solche Rangfolgen durchaus auch über klassische Mengenkonzepte darstellen und in diesem Sinne zweiwertig-logisch rekonstruieren. So könnte man etwa eine Menge komponentieller Deskriptoren aufstellen (für *Vogelhaftigkeit* z. B.: *legt Eier, kann fliegen, ist klein, hat zwei Beine, zwitschert*, etc.) um den Rang, den ein Tiername einnehmen soll, abhängig zu machen von der Menge der nun wieder Ja/Nein-entscheidbaren Deskriptoren. Danach würde *Rotkehlchen* deswegen der typischste Vogel genannt werden können, weil für ihn die Zahl der positiv entschiedenen Deskriptoren am höchsten liegt. – Diese Tatsache der Rekonstruierbarkeit kann aber nicht als Einwand gelten, der das Ergebnis des Heiderschen Experiments berührt. Denn der Test besagt ja nicht, daß eine resultierende graduelle Gewichtung etwa *unmotiviert* sei (ganz im Gegenteil), sondern er besagt nur daß die Probanden quasi integral über virtuelle Komponenten gradieren, *ohne* sich dieser Komponenten bewußt sein zu müssen. Daß sie, falls nötig, ihre Rangfolgen nachträglich motivieren, explizieren und auch verteidigen könnten, zeigt dagegen, daß sie in der Regel zu einer logischen Rekonstruktion ihrer zunächst *unbewußten* Graduierung darüber hinaus *auch* fähig sind.

Dies wird bestätigt von *immediate-response*-Tests, in denen vordergründig gerade eine alternative Entscheidung gefordert wurde, in denen der eigentliche Testparameter aber die Reaktionszeit war, die der Proband brauchte, um auf Sätze der Form „(Individuenvariable) x ist ein (Begriffsklasse) A“ mit *wahr* oder *falsch* zu antworten. Es zeigte sich dabei, daß die Antwort-Zeiten deutlich kürzer ausfielen bei Sätzen mit sehr typischen Individuen einer Begriffsklasse (etwa „Ein *Rotkehlchen* ist ein Vogel“) gegenüber deutlich längeren Antwortzeiten bei weniger typischen in Sätzen wie („Eine *Fledermaus* ist ein Vogel“).

Mir scheint, diese – und ähnliche hier nicht referierte – Befunde sind überzeugende empirische Belege zumindest dafür, daß begriffliche Zugehörigkeit in natürlicher Sprache von Sprachteilhabern eher als *gradueller Übergang* denn als *abrupter Sprung* konstituiert wird. Dies aber spricht eher *gegen* als *für* die Versuche, im Bereich der natürlichsprachlichen Semantik mithilfe zwei- oder auch dreiwertig-logischer, scharfer Systeme zu *formalisierten* Modellen zu kommen, die auch *empirisch* adäquat sind. Denn die Aufstellung empirischer Theorien und Modelle ist mehr als ein bloßes Anwendungsfeld von

Mathematik und formaler Logik. Hier muß sich im Gegenteil – wie Dieter Wunderlich (1974) schreibt –:

> die Art des zu entwickelnden formalen Systems [...] an den Bedürfnissen der betreffenden Wissenschaft orientieren, anstatt daß sich diese Wissenschaft nur an den vorhandenen Logiksystemen orientiert.[3]

2.2 Die von Lotfi A. Zadeh (1965) vorgelegte und seither in theoretischer wie praktischer Hinsicht ausgebaute und erweiterte Theorie der *unscharfen Mengen* (Fuzzy Sets Theory) ist eine solche an den Bedürfnissen orientierte Entwicklung, deren Fruchtbarkeit sich durch eine zunehmend größere Zahl von Publikationen ausweist.[4] Neben den theoretischen Arbeiten, die das Konzept der *Unschärfe* auf die verschiedensten mathematischen Strukturen, Systeme, Topologien etc. ausdehnen, machen inzwischen jene Publikationen schon den größeren Teil aus, die im Bereich der empirisch arbeitenden Disziplinen im besonderen mit solchen Phänomenen befaßt sind, die bei teilweise hoher Komplexität empirisch nur unvollkommen und vage zugänglich sind. Das reicht von der automatischen Zeichenerkennung über System- und Automatentheorie in Verfahrenstechnik und Unternehmensforschung bis hin zum gerade in letzter Zeit wieder aktuellen Forschungsschwerpunkt der *künstlichen Intelligenz*, in dem diese Ansätze mit logischen, psychologischen und linguistischen Überlegungen zusammentreffen.[5]

Entscheidend für die breite Aufnahme der neuen Theorie ist dabei wohl vor allem, daß sich mit ihr die Möglichkeit ergibt, sehr komplexe Gegenstandsbereiche auch formal in befriedigender Weise anzugehen, *ohne* dabei diese Komplexität entweder *reduzieren* oder aber gar *hinwegpräzisieren* zu müssen. Denn – so Zadeh (1972) –:

> in general, complexity and precision bear an inverse relation to one another in the sense that, as the complexity of a problem increases, the possibility of analysing it in precise terms diminishes.[6]

Gerade in Bereichen, für die menschliches Handeln und Verhalten konstitutiv ist, bedeutet aber Präzisierung nicht immer schon den entscheidenden Schritt zur wissenschaftlichen Klärung eines Phänomens, sondern oft nur dessen Verschwinden. Dies gilt sicherlich in besonderem Maße für die Analyse und Beschreibung *natürlich-sprachlicher*, im Unterschied zu *standard-sprachlicher Bedeutung*.

Der Grundgedanke der Theorie der unscharfen Mengen, die die traditionelle Mengentheorie umfaßt, ist denkbar plausibel und einfach.

Im Unterschied zur klassischen oder scharfen Mengentheorie, in der ein Individuum *alternativ* im Hinblick auf eine Menge entweder Element ist oder nicht, kann man in der neuen Theorie die Zugehörigkeit eines Individuums zu einer deswegen *unscharf* genannten Menge *graduell* angeben. Diese Zugehörigkeit wird dabei durch eine reelle, positive Zahl

[3] Wunderlich, D.: – Grundlagen der Linguistik, Reinbek (Rowohlt) 1974; 153.

[4] Zadeh, L. A.: – Fuzzy Sets. In: Information and Control 8 (1965), 338-353.

[5] Zimmermann, H. J.: – A Bibliography of the Theory and Application of Fuzzy Sets, Technical Report 75/16, Inst. für Wirtschaftswissenschaften der RWTH Aachen.

Gaines, B. R./Kohout, L.: – The Fuzzy Decade: A Bibliography of Fuzzy Systems and Related Topics. Erscheint in: International Journal of Man-Machine Studies, voraussichtlich Ende 1976.

[6] Zadeh, L. A.: – Fuzzy Languages and their Relation to Human Intelligence. In: Proceedings of the Intern. Conference on Man and Computer, Basel (Karger) 1972, 130-165; 131.

des Intervalls zwischen 0 und 1 ausgedrückt, wobei der Wert exakt 0.0 der klassischen Nicht-Zugehörigkeit, der Wert exakt 1.0 der klassischen Zugehörigkeit eines Elements zu einer Menge entspricht. Das soll im folgenden anhand eines einfachen Begriffs wie *Mittelklassewagen* verdeutlicht werden. Dieser Begriff, aus der Automobilwerbung jedem von uns geläufig, ist *unscharf* in bezug auf die Menge derjenigen Fahrzeugtypen, die gemeint sind, wenn von *Mittelklassewagen* die Rede ist.

Einem Fiat 500 beispielsweise würde deshalb in dieser Menge der *Mittelklassewagen*, ein äußerst geringer Zugehörigkeitswert zukommen, weil er als ausgesprochener Kleinwagen gilt; wenn man seine „Mittelklassehaftigkeit“ anhand des Kaufpreises bewerten sollte, würde man ihm, sagen wir, den Wert 0 = 0.0 geben müssen. Ein VW-Golf, beinahe schon ein Mittelklassewagen, zumindest dem Preis nach, hätte einen deutlich höheren Wert, sagen wir = 0.5, ein Opel-Rekord, als typischer Mittelklassewagen, den Wert = 1.0, ein Mercedes, beinahe schon Luxuswagen, einen wieder deutlich niedrigeren Wert, beispielsweise = 0.3, während ein Rolls Royce, als reine Luxuslimousine, einen Zugehörigkeitswert von exakt = 0.0 zur unscharfen Menge M der *Mittelklassewagen* aufwiese. Trägt man nun zur Veranschaulichung (Abb. 1) den Individuenbereich X der Einfachheit halber als kontinuierliche Preisskala von links nach rechts auf der Abszisse – und die einzelnen den verschiedenen Fahrzeugtypen (subjektiv) zugeschriebenen Zugehörigkeitswerte $\mu_M(x)$ auf der Ordinate auf, so ergibt sich die unscharfe Menge M als Kurve, die als Darstellung der referentiellen Bedeutung von *Mittelklassewagen* über X gelten kann.

DM	500	1100	1700	3200	6300
X	Fiat	VW-Golf	Opel-Record	Mercedes	Rolls-Royce
	(f)	(g)	(o)	(m)	(r)
$\mu_M(X)$	0.0	0.5	1.0	0.3	0.0

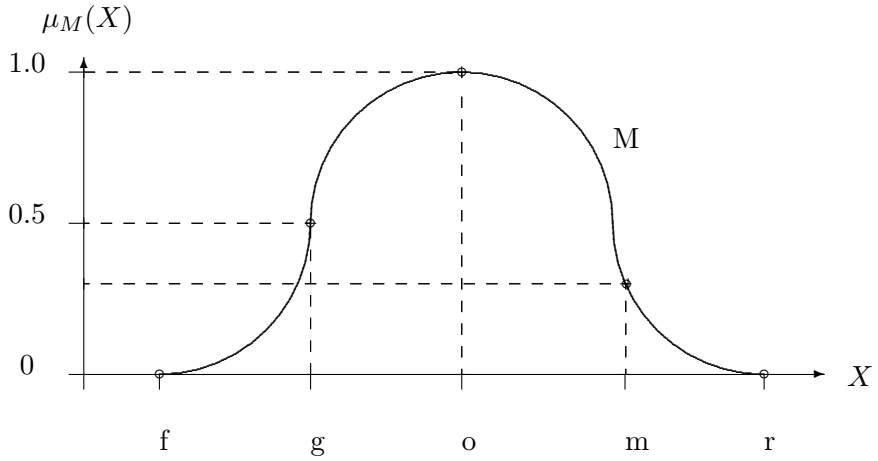

Abb. 1

Allgemein: eine unscharfe Menge A in X wird charakterisiert durch die Zugehörigkeitsfunktion

$$\mu_A : X \to [0, 1]$$

die jedem x aus X einen (und nur einen) Zugehörigkeitswert $\mu_A(x)$ zuordnet, der den Grad angibt, mit dem das Individuum x als Element der unscharfen Menge A zu gelten hat. Die Menge A besteht also aus der Menge der geordneten Paare

$$A := \{(x, \mu_A(x))\}$$

Für beliebige unscharfe Mengen A, B und C lassen sich nun wie für die klassischen, scharfen Mengen *Gleichheit* und *Enthaltensein* sowie die Verknüpfungen *Durchschnitt, Vereinigung* und *Komplementbildung* definieren, deren detaillierte Einführung sich hier erübrigt.[7] Diese Definitionen sind mit denen für klassische Mengen konsistent und entsprechen den logischen Operatoren Konjunktion, Adjunktion und Negation. Soweit zur Grundlage der Theorie der unscharfen Mengen.

Wie das Beispiel verdeutlicht, wird die oben angesprochene Forderung der Adäquatheit der Beschreibung natürlich-sprachlicher Bedeutung durch unscharfe Mengen nur zum Teil, nämlich bloß *formal* erfüllt. Diese formale Darstellung hängt aber entscheidend ab

1. von dem jeweils zugrundegelegten Individuenbereich (hier: Autotypenpreise), in dem die unscharfe Menge (Mittelklassewagen) definiert wird und

2. von dem Verfahren, aufgrund dessen jedem Individuum (z.B. Golf = 0.5) ein Zugehörigkeitswert in bezug auf diese Menge zugeschrieben wird.

Beides, die Bestimmung des Individuenbereichs wie das Verfahren zur Ermittlung von Zugehörigkeitswerten, betrifft aber schon Fragen einer *Analyse* natürlich-sprachlicher Bedeutung, deren Bedingungen uns im folgenden beschäftigen sollen.

3.1 Man könnte versucht sein, in den Verfahren der experimentellen Psychologie eine auch für die empirische Linguistik anwendbare Methodik zu sehen, die natürlich-sprachliche Bedeutungen über Probandenbefragungen empirisch zu ermitteln erlaubt. Dies um so eher, als das in anderem Zusammenhang angeführte Heidersche Experiment ja zu Resultaten geführt hatte, die bei geringfügig abgeänderter Versuchsanordnung leicht als *unscharfe Menge* sich würde darstellen lassen. So könnte man etwa die Menge der vorgegebenen Tiernamen als Individuenbereich deuten, über dem sich die *Bedeutung* des Begriffs *Vogel* als unscharfe Menge definieren ließe. Dazu hätten die Probanden, anstatt eine Rangordnung unter den Tiernamen zu bilden, jedem dieser Tiere nur einen Zahlenwert zwischen 0 und 1 zuzuordnen, der als Zugehörigkeitsgrad dann angäbe, in welchem Maße ihrer Meinung nach die Idealvorstellung *Vogel* sich mit jedem der Tiernamen verbindet, d. h. diese zum Begriff *Vogel* beitragen.

Was spricht gegen solche Befragungsverfahren im Rahmen der Analyse und Beschreibung natürlich-sprachlicher Bedeutungen?

[7] Vgl. Zadeh (1965), 340-342.

Sieht man einmal von den in jeder Befragungsmethodik gelegenen, eher technischen Problemen ab, dann bleiben im wesentlichen zwei Einwände, die diese Verfahren als ungeeignet erscheinen lassen:

1. alle Probanden befragenden Bedeutungsanalysen – von denen das Osgoodsche *Semantische Differential* die vielleicht bekannteste ist – geben in der einen oder anderen Form einen Individuenbereich vor, der als quasi komponentielle Zerlegung in der Regel schon den Analyserahmen absteckt, innerhalb dessen dann nurmehr die Gewichtungen der einzelnen Komponenten oder polaren Begriffspaare ermittelt werden;

2. alle Probanden befragenden Verfahren zur Bedeutungsanalyse sind – kommunikationstheoretisch gesprochen – sterile Laborversuche. Damit ist gemeint, daß die Fähigkeit, die einem Probanden während einer Befragung abverlangt wird, gerade nicht jene ist, die das Experiment zu testen vorgibt: nicht auf sein Vermögen wird rekurriert, natürlich-sprachliche Bedeutungen in ihrem tatsächlichen kommunikativen Zusammenhang zu verstehen, sondern auf seine Fähigkeit, dieses Verstehen in künstlichen Zusammenhängen, die vom Probanden selbst erst hergestellt werden müssen, zu simulieren.

3.2 Eine empirisch fundierte und formal befriedigende Analyse natürlich-sprachlicher Bedeutung muß diesen Einwänden Rechnung tragen. Bedeutung wird daher nicht außerhalb des pragmatischen Zusammenhangs analysiert werden können, in dem sie sich konstituiert, d.h. in der *Kommunikation*, und Bedeutung wird anhand von Gegebenheiten analysiert werden müssen, die diesen Zusammenhang repräsentieren, d. h. anhand von *Texten*.

Als *Text* gilt dabei jede Folge von sprachlichen Zeichen, die im Zusammenhang (d. h. Kontext) einer konkreten Situation von tatsächlichen Sprechern/Hörern zum Zweck der Kommunikation geäußert/erkannt werden.

Unter *Kommunikation* wird der Prozeß zunehmender Einschränkung von Wahlmöglichkeit verstanden, den die daran Beteiligten über Zeichen und Zeichenfolgen (d. h. Texte) wechselseitig initiieren und nachvollziehen.

Setzt man in diesem handlungsorientierten Sinne einmal voraus, dass das Resultat einer solchen über *Texte* initiierten Kommunikation *primär* nicht das *richtige* oder *falsche Verstehen* ist, sondern eben jener *Abbau von Unsicherheit*, den die miteinander Kommunizierenden je nach pragmatischen Erfordernissen in *größerem* oder *geringerem* Grade anstreben und/oder erzielen, dann läßt sich Bedeutung nicht länger mehr als eine *statische* Qualität beschreiben, die Zeichen und Zeichenfolgen auf geheimnisvolle Weise zukommt. Vielmehr muß die Analyse und Beschreibung der Bedeutung eines Zeichens, Wortes, etc. als eine Art *Momentaufnahme* verstanden werden, die den im Prinzip andauernden *dynamischen* Prozeß der Bedeutungskonstitution quasi unter dem Blickwinkel dieses betreffenden Zeichens, Wortes, etc. aber einer Pragmatik abbildet. Unter *Pragmatik* wird dabei im folgenden ein sowohl Ort und Zeit wie Gegenstand und Beteiligte umfassender situativer Kommunikationsrahmen verstanden.

Diesen kommunikationstheoretisch beschreibbaren Zusammenhang, der einer prag-

matischen Fundierung der Semantik entspricht,[8] kann eine mathematisch-statistische Analyse natürlich-sprachlicher Bedeutung durchaus berücksichtigen. Wie an anderer Stelle näher ausgeführt,[9] läßt sich dabei Gegenstand und Ziel einer Untersuchung mit der weitgehend operationalisierten Begriffsbildung von Stichprobe und Grundgesamtheit in Verbindung bringen, wie dies innerhalb statistischer Methodik geschieht.

Jede textstatistische Untersuchung kann zwar davon ausgehen, daß sich in Texten (im Unterschied etwa zu bloßen Wörter- bzw. Zeichenansammlungen) *Ordnungsrelationen* und *regelhafte Beziehungen* zwischen den verwendeten Zeichen, Wörtern, Lexemen, etc. aufdecken, beschreiben und messen lassen.[10] Damit aber die so ermittelten Beziehungen nicht ihrerseits eine bloße Ansammlung numerischer Fakten und uninterpretierter Daten bleiben, muß ein Untersuchungsgegenstand zusätzlichen Forderungen genügen, wenn seine quantitativ-statistische Analyse sinnvoll sein soll: er muß sich im Sinne statistischer Methodik als *zufällige Stichprobe* aus einer *Grundgesamtheit* deuten lassen, über die Aussagen gemacht werden sollen.

Zufällig heißt eine Stichprobe dann, wenn die Operation der Auswahl eines Untersuchungsgegenstandes (etwa einer Textmenge) bei – im Prinzip beliebig häufiger – Wiederholung auf die Grundgesamtheit hin konvergiert.

Die *Grundgesamtheit*, die im textwissenschaftlichen Bereich im allgemeinen fiktiv sein wird, läßt sich nur vom Untersuchungsziel her bestimmen. Denn aus der Sicht der Statistik ist ein *Untersuchungsziel* identisch mit dem Vorhaben, intersubjektiv nachprüfbare Aussagen über eine *Grundgesamtheit* zu machen aufgrund der Analyse von daraus entnommenen zufälligen *Stichproben*, die den *Untersuchungsgegenstand* bilden.

Ist – wie im vorliegenden Fall – das Untersuchungsziel eine Analyse natürlich-sprachlicher Bedeutung im Zusammenhang der sie fundierenden Pragmatik, dann können nur solche natürlich-sprachlichen Texte den Untersuchungsgegenstand bilden, die von bestimmten tatsächlichen Sprechern/Verfassern in einer konkreten, dabei *gleichartigen* Kommunikationssituation geäußert worden sind. Nur eine solche Textmenge, die wir ein *pragmatisch-homogenes* Textkorpus nennen wollen, kann als zufällige Stichprobe aller derjenigen Äußerungen gelten, die in einer bestimmten Pragmatik tatsächlich gemacht wurden oder hätten gemacht werden können und so eine *fiktive* Grundgesamtheit bilden.

Bei der statistischen Analyse kann nun im wesentlichen von einem Verfahren Gebrauch gemacht werden, das erlaubt, etwa vorhandene Regularitäten der Abhängigkeit zwischen Wörtern/Lexemen festzustellen und deren unterschiedliche Intensitäten (von wechselseitiger Abstoßung über Beziehungslosigkeit bis zur wechselseitigen Anziehung) *graduell* in numerischen Werten des Intervalls von - 1 bis + 1 zu präzisieren. Dies leistet

[8]Schneider, H. J.: – Pragmatik als Basis von Semantik und Syntax, Frankfurt/Main (Suhrkamp) 1975,112-128, bes. 116, 121 ff.

[9]Rieger, B.: – Warum mengenorientierte Textwissenschaft? Zur Begründung der Statistik als Methode. In: LiLi 8 (1972) 11-28.

[10]Harris, Z. S.: – Mathematical Structures of Language (Interscience Tracts in Pure and Applied Mathematics), New York/London (J. Wiley) 1968.

Salton, G.: – Automatic Text Analysis. In: Science 168 (1970) 335-343.

ders.: – On the Role of Words and Phrases in the Automatic Content Analysis of Texts, Vortrag auf der 2. Intern. Conference on Computers and the Humanities (ICCH/2) Los Angeles 1975, erscheint voraussichtlich 1977.

der Korrelationskoeffizient. Er mißt die Beziehung eines jeden Wortes zu jedem anderen verwendeten Wort, und zwar aufgrund des Gebrauchs den die Sprecher/Verfasser von ihnen in den analysierten Texten machen.

Es wird gezeigt werden, daß diese Korrelationswerte eines Wortes zu sämtlichen anderen Worten des Vokabulars als Grundlage dienen können für eine empirisch adäquate Zumessung von Zugehörigkeitswerten, die dann die Verwendungsstruktur eines Wortes, d. h. seine Bedeutung als unscharfe Menge über dem Individuenbereich des verwendeten Vokabulars abbilden und damit Kommunikations-abhängig für eine bestimmte Pragmatik definieren und beschreiben läßt.

4.1 Es soll im folgenden versucht werden, den Gang der Analyse möglichst anschaulich darzustellen. Aus diesem Grunde wähle ich den einfachen Fall eines Textkorpus T, das aus einer Anzahl Texten t bestehe und die oben aufgeführten Bedingungen erfülle. Die Gesamtzahl der darin verwendeten Wörter (*token*) möge aber nur aus drei Worttypen (*types*) i, j und k bestehen, die das Vokabular V ausmachen.

Es braucht nicht betont zu werden, daß diese Vereinfachung nicht die Komplexität des zu analysierenden Phänomenbereichs reduziert, sondern nur der größeren Anschaulichkeit dient. Obwohl wir in natürlichsprachlichen Texten mit sehr großen Vokabularen zu tun haben, bleibt die Mathematik die gleiche, wenn wir uns zunächst mit einem Vokabular von nur drei Worttypen beschäftigen.

Der Korrelationskoeffizient α mißt nun die Beziehung eines jeden Wortes i, j und k zu jedem der anderen Worte des Vokabulars, und zwar aufgrund ihrer Verwendung in den Texten des Korpus. Das ergibt für jedes Wort drei Meßwerte, für die Korrelationen von i beispielsweise die Werte ii, ij und ik. Diese Meßwerte werden nun als Koordinaten interpretiert, die für jedes Wort i, j, k einen Punkt α_i, α_j und α_k in einem Raum definieren, der durch die drei den Wörtertypen entsprechenden Achsen i, j, k aufgespannt wird (Abb. 2).

Die Lage eines Punktes α_i in diesem Raum wird demnach bestimmt durch das Tripel der Korrelationswerte, d. h. durch die *Verwendungsregularitäten* des Wortes i zu allen anderen Wörtern in den Texten des Korpus. α_i heißt daher *Korpuspunkt von* i, im α- oder *Korpusraum*. Zwei α-Punkte in diesem Raum werden folglich dann enger benachbart sein, wenn ihre jeweiligen Verwendungsregularitäten nicht sehr unterschiedlich sind. Als Maß dieser Unterschiedlichkeit der Verwendungsregularitäten kann die Entfernung zwischen zwei α-Punkten im Korpusraum (gepunktete Linien in Abb. 2) gelten, die als Distanz- oder δ-Werte gemessen werden können.

Diese δ-Werte stellen nun eine neue Charakteristik dar. Sie läßt sich auf zweierlei Weise interpretieren:

1. man faßt die (gepunkteten) δ-Abstände eines (α-Punktes zu sämtlichen *anderen* als neue Koordinaten auf: dann definieren diese Koordinaten wiederum einen Punkt in einem neuen Raum, der δ- oder *Bedeutungsraum* heiße. Die Lage dieses Punktes wird dabei bestimmt von *allen* Unterschiedlichkeiten (δ- oder Distanzwerte) *aller* Verwendungsregularitäten (α- oder Korrelationswerte) eines Wortes in den analysierten Texten.

2. man faßt die (gepunkteten) Abstände *eines* α-Punktes zu sämtlichen *anderen* als Zugehörigkeitsgrade auf: dann definieren (nach geeigneter Umformung der δ-Werte

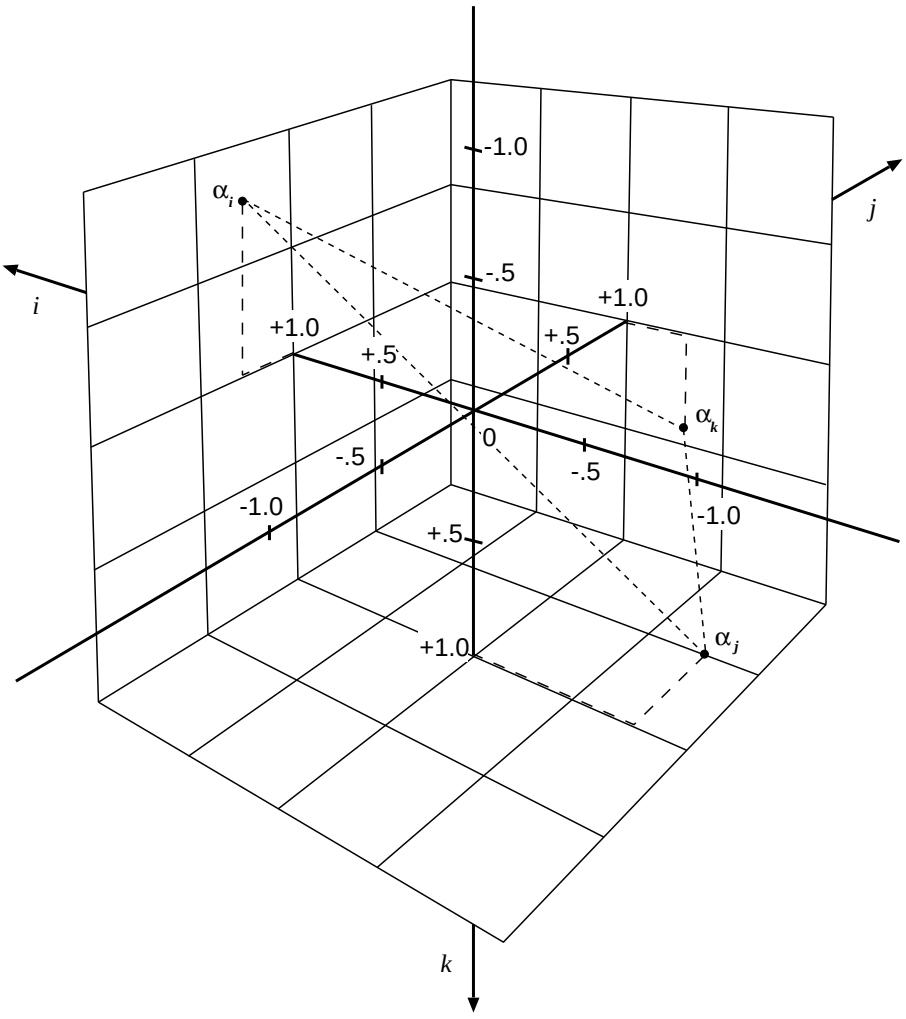

Abbildung 2

in μ- Werte) diese Zugehörigkeitswerte eine *unscharfe Menge*, die die Bedeutung eines Wortes über dem Vokabular als dem Individuenbereich darstellt.

Beide Auffassungen der Distanzwerte, sowohl als Koordinaten eines *Punktes* im Bedeutungsraum als auch als Zugehörigkeitswerte einer *unscharfen Menge* im Vokabular sind gleichwertig: sie bilden Bedeutung eines Wortes ab als Funktion *aller* seiner Unterschiedlichkeiten in allen seinen Verwendungsregularitäten, wie sie sich in einem homogenen Textkorpus als einer zufälligen Stichprobe aus einer bestimmten Pragmatik konstituieren.

Die Darstellung von Bedeutungen als unscharfen Mengen erlaubt darüber hinaus aber eine sehr wesentliche Erweiterung unseres zunächst nur analytischen Modells. Durch die Übertragung der in der Theorie der unscharfen Mengen gegebenen Verknüpfungsoperationen lassen sich nämlich neue Bedeutungen dadurch generieren, daß man die empirisch ermittelten Bedeutungen nun aufgrund der formal definierten Operationen (Komplement, Durchschnitt, Vereinigung) *negiert*, oder miteinander durch *Konjunktion* oder *Adjunktion* verknüpft. Diese Bedeutungen lassen sich dann wiederum als unscharfe Mengen über dem Vokabular bzw. als neue Punkte im Bedeutungsraum darstellen. Erst diese Möglichkeit der Verknüpfung schon definierter Bedeutungen zu *neuen* Bedeutungen vermag das in der natürlich-sprachlichen Bedeutungskonstitution wirksame Moment der *Kreativität* abzubilden.

4.2 Soweit der vielleicht etwas verwirrende Gang von der Ermittlung der grundlegenden *Verwendungsregularitäten* (α -Werte) über die Messung ihrer *Unterschiedlichkeiten* (δ-Werte) bis hin zur Abbildung von so definierten Bedeutungen (und deren Verknüpfungen) in einem *Bedeutungsraum*. Die im folgenden vorgelegten Resultate einer empirischen Bedeutungsanalyse mögen dies illustrieren.

Dabei handelt es sich um ein literarhistorisches Textkorpus gleicher Pragmatik, nämlich um $T = 600$ Gedichttexte deutscher Studenten des frühen 19. Jhs., das mir aus früheren Untersuchungen[11] schon in maschinenlesbarer Form vorlag. In diesem Korpus wurde ein Vokabular von $V = 315$ Worttypen (types) analysiert, welche insgesamt (token) 21 000 mal vorkommen.

Da die anschauliche Darstellung eines 315-dimensionalen Bedeutungsraumes auf echte Schwierigkeiten stößt und auch die Abbildung der Bedeutung eines Wortes, Lexems, etc. als unscharfer Menge über dem Vokabular V nur ein n-Tupel von $n = 315$ δ-Werten zeigt (Abb. 3), bin ich auf eine andere Möglichkeit der Darstellung verfallen.

[11] Rieger, B.: – Literarische Massenphänomene und mengenorientierte Textanalyse. Zu Gegenstand und Methode der Trivialliteraturforschung. In: Das Triviale in Musik, Literatur und bildender Kunst, hg. von H. de la Motte-Haber, Frankfurt/M. (Klostermann) 1972, 42-62.

δ-Werte n-Tupel, $n = 315$, von
FRÜHLING

1.921	2.558	2.050	2.351	2.139	2.142	2.172	2.593	2.184	1.984
2.302	2.142	2.063	2.173	1.995	1.872	2.252	0.000	2.302	2.001
1.756	1.828	2.239	1.962	2.069	2.013	2.490	1.896	2.097	2.579
2.263	1.425	1.845	2.316	2.160	2.058	1.823	1.821	2.238	1.820
1.959	2.163	2.038	2.412	2.088	1.753	2.436	2.202	2.038	1.849
2.122	1.908	2.179	2.546	2.253	2.010	1.952	1.946	2.202	2.551
1.636	1.978	2.105	2.040	2.450	2.873	1.871	2.115	2.399	2.011
2.111	2.315	2.073	2.453	2.056	2.707	2.093	1.704	1.993	$^{+}0.000$
2.066	2.048	1.970	2.171	2.250	1.978	1.993	2.389	2.182	2.040
2.193	2.206	2.354	2.858	2.157	1.881	2.486	1.888	2.233	1.875
2.110	2.276	2.423	2.328	1.949	1.977	2.255	2.035	2.652	2.060
2.329	2.248	2.191	1.830	2.117	2.016	2.033	2.002	2.225	2.469
2.083	1.929	1.912	2.032	2.051	2.083	2.014	2.309	2.026	2.704
1.931	2.151	2.313	2.244	2.300	2.152	2.121	2.177	1.770	2.309
2.229	1.935	2.191	2.211	2.135	1.997	1.829	2.086	1.963	2.013
2.006	2.134	2.121	2.470	1.973	1.551	1.861	2.364	2.018	2.038
2.072	1.923	1.928	2.039	2.146	1.967	1.929	2.163	1.960	1.952
1.684	2.356	2.220	2.330	2.102	0.000	2.308	2.172	2.185	1.911
2.177	2.378	1.830	2.278	2.280	1.908	2.215	2.098	2.174	1.729
2.289	2.215	2.119	1.976	1.898	1.861	1.993	2.114	1.926	1.868
2.129	1.845	2.152	2.012	2.017	2.230	1.993	2.278	2.208	1.990
1.973	2.004	1.787	1.907	2.126	2.033	2.229	2.171	2.015	2.561
2.461	2.013	1.982	2.153	2.007	2.319	2.077	1.930	2.057	2.281
2.087	1.832	2.602	2.505	2.072	1.993	2.004	2.039	2.016	2.171
2.004	2.153	2.005	2.187	2.141	2.055	1.969	2.265	2.667	2.246
1.983	2.084	1.803	2.043	1.991	1.978	2.761	2.405	2.061	2.187
2.154	2.142	2.148	2.017	2.176	0.000	2.119	1.328	2.063	2.285
2.415	2.282	2.072	2.566	2.347	1.387	2.039	1.942	1.993	2.028
2.091	1.991	1.890	2.084	2.151	1.992	2.113	2.091	2.114	2.170
1.815	2.189	2.261	1.885	2.062	2.318	2.191	2.318	2.317	1.948
2.081	1.961	2.367	2.013	2.309					

Abb. 3

Um eine Vorstellung zu geben von der Lage eines Punktes im Bedeutungsraum und damit einen Eindruck zu vermitteln von der durch diesen Punkt repräsentierten Bedeutung eines Wortes, Lexems, etc. in den historischen Texten des Korpus, lassen sich diejenigen Punkte feststellen, die dem darzustellenden Bedeutungspunkt im semantischen Raum am nächsten sind. Die Konfiguration einer Bedeutung durch Angabe derjenigen Bedeutungspunkte, die innerhalb eines solchen lexikalischen Systems in unmittelbarer Nachbarschaft liegen, hat ihre sprachwissenschaftliche Entsprechung im paradigmatischen oder semantischen Feld. Dieses läßt sich – wie an anderer Stelle ausgeführt[12] – topologisch als *Umgebung* innerhalb einer Lexikonstruktur explizieren.

Die folgenden Konfigurationen, welche die Lage der Bedeutungspunkte FRÜHLING

[12]Rieger, B.: – Eine tolerante Lexikonstruktur. Zur Abbildung natürlichsprachlicher Bedeutung auf

(Abb. 4), GRAB/GRUFT (Abb. 5), FRÜHLING ∧ GARTEN (Abb. 6) und FRÜHLING ∨ GARTEN (Abb. 7) im Bedeutungsraum erkennen lassen, sind daher als solche Umgebungen bzw. Felder zu verstehen.

FRÜHLING

LENZ	2.698	BLÜTE	2.768	WINTER	2.817
DUFT	3.269	FELD	3.384	NEU	3.424
BERG	3.450	NACHTIGALL	3.479	MAI	3.513
VOGEL	3.540	LERCHE	3.620	ROSE	3.663
PRACHT	3.663	GRAS/HALM	3.713	ZART	3.733
BAUM	3.736	SONNE	3.751	QUELLE	3.780
SILBER	3.793	BACH	3.812	SCHÖN	3.853
WIESE/AUE	3.855	HOLD	3.861	BUSEN	3.879

Abb. 4

GRAB/GRUFT

FRIEDHOF	2.945	TOD	3.696	FINSTER	3.758
KALT	3.997	STUNDE	5.032	SCHWARZ	5.142
FAHL/WELK	5.269	BLEICH	5.756	HOHL	5.922
HEILIG	5.968	GRAU	6.005	SCHEIN	6.038
ABGRUND	6.241	ANGST	6.265	SCHWEBEN	6.819
GELB	6.877	BLASS	6.933	WEISS	6.972
DUMPF	7.166	SCHATTEN	7.290	GLANZ	7.533
JUNG	7.919	HAIN	7.933	LEID	7.934

Abb. 5

FRÜHLING ∧ GARTEN

BLÜTE	2.048	ZWEIG/AST	2.161	FRÜHLING	2.165
WUNDER	2.214	DUFT	2.242	BAUM	2.296
ROSE	2.326	FRÜH	2.434	TRAUM	2.463
LEISE	2.498	SONNE	2.513	HOLD	2.560
STILL	2.628	VOGEL	2.632	LENZ	2.642
SCHÖN	2.683	WOLKE	2.705	BLATT	2.733
STIRN	2.744	BLAU	2.745	WIESE/AUE	2.745
SCHNEE	2.747	FROST	2.765	LUFT	2.783

Abb. 6

unscharfe Mengen in Toleranzräumen. In: LiLi 16 (1974) 31-47.

ders.: – On a Tolerance Topology Model of Natural Language Meaning, Vortrag auf der 2. Intern. Conference on Computers and the Humanities (ICCH/2), Los Angeles 1975.

FRÜHLING ∨ GARTEN

GARTEN	2.980	BLÜTE	3.528	FRHÜH	3.577
LENZ	3.645	PRACHT	3.650	WUNDER	3.664
BAUM	3.674	WINTER	3.689	NACHTIGALL	3.743
BACH	3.745	ZART	3.770	MAI	3.778
FRUCHT	3.785	DUFT	3.790	ABEND	3.813
FELD	3.827	NEU	3.832	LILIE	3.853
SILBER	3.918	ZWEIG/AST	3.926	BERG	3.927
LERCHE	3.965	TRAUM	3.967	GRAS/HALM	3.971

Abb. 7

5 Mit einigen Hinweisen zur theoretischen wie zur eher anwendungsorientierten Seite dieses Ansatzes möchte ich schließen:

1) Was diese Korrelationsanalyse von Wörtern/Lexemen in einem pragmatisch-homogenen Textkorpus für eine semantische Modellbildung interessant macht, ist der Umstand, daß die in der *linearen* (eindimensionalen) Ordnung jedes einzelnen Texts gegebenen Beziehungen (Saussure's *rapports syntagmatiques*) sich nutzen lassen zur Ermittlung *relationaler* (vieldimensionaler) Beziehungsstrukturen, (Saussure's *rapports associatifs*), von denen jeder einzelne Text, qua pragmatischem Kommunikationszusammenhang, quasi immer schon Gebrauch macht, durch solchen Gebrauch aber auch modifiziert. Dieser Zusammenhang wird methodisch über die Menge der im Korpus zusammengefaßten pragmatisch-homogenen Texte in die statistische Analyse einbezogen und als kommunikativer Zusammenhang syntagmatischer und paradigmatischer Relationen faßbar.

2) Die scheinbare Präzision, mit der Bedeutungen als n-Tupel von Zahlenwerten angegeben und als Konfigurationen von Bedeutungspunkten dargestellt werden, darf nicht übersehen machen, daß es sich hierbei um relativ *präzise* Abbildungen sehr *unpräziser* Gegebenheiten handelt (nämlich um unscharfe Mengen im Sinne Zadehs).

3) Die Darstellung von Bedeutungen als unscharfer Mengen bzw. Punkten im Bedeutungsraum ist als eine momentane *synchrone* Zustandsbeschreibung eines Systems zu verstehen, dessen strukturelle Bedeutungszusammenhänge sich *diachron* im Prozeß der Bedeutungskonstitution ständig verändern.

4) Dieses System von Bedeutungspunkten läßt sich – wie an anderer Stelle ausgeführt[13] – als ein formales Modell einer Lexikonstruktur deuten, in dem Sinnrelation wie Synonymität, Hyponymität etc. über die in der Theorie der unscharfen Mengen gegebenen Definitionen der Gleichheit, des Enthaltenseins etc. formal befriedigend und Pragmatikabhängig sich explizieren lassen.

5) Es wäre sicherlich wünschenswert, die semantischen Strukturzusammenhänge (Lexikonstrukturen) für die verschiedensten Pragmatiken zu ermitteln, d. h. differenziert nach Kommunikationssituationen und/oder Kommunikationsgegenständen und/oder Kommunikationspartnern. Dabei könnten sich gerade im Bereich der soziolektisch wie idiolektisch ausgerichteten semantischen Untersuchungen interessante Resultate ergeben, möglicherweise aber auch im Rahmen so spezieller Fragestellungen, wie sie etwa von der

[13]Rieger, B.: – Fuzzy Structural Semantics. On a Generative Model of Vague Natural Language Meaning, Vortrag auf dem 3. European Meeting on Cybernetics and Systems Research (EMCSR 76) in Wien, April 1976, erscheint 1977.

Aphasieforschung aufgeworfen werden.

6) Vielleicht kann die Verbindung von mathematisch-statistischen Verfahren mit der Theorie der unscharfen Mengen dazu beitragen, daß man über *unpräzise* Gegebenheiten *präziser* als bisher sich wird verständigen können, was in der Wissenschaft von den natürlich-sprachlichen Bedeutungen – wie mir scheint – nicht wenig wäre.