Tagungen

First International Quantitative Linguistics Conference (QUALICO 91)

In der Woche vom 23. bis 27. September 1991 fand an der Universität Trier die erste internationale *Quantitative Linguistics Conference* (QUALlCO 91) statt. Veranstalter dieser von der DFG finanziell sowie von zahlreichen nationalen und internationalen wissenschaftlichen Gesellschaften und Vereinigungen organisatorisch unterstützten Konferenz waren die GLDV, die im Rahmen der QUALlCO 91 auch ihre Jahrestagung abhielt (s. deren in diesem LDV-Forum, S.2-37, abgedruckte Beiträge), und die Herausgeber der Buchreihe *Quantitative Linguistics* (QL). Auf Initiative des derzeitigen Vorsitzenden der GLDV wurden beide Veranstaltungen als *joint conference* vom Fach Linguistische Datenverarbeitung/Computerlinguistik der Universität Trier ausgerichtet.

An der Konferenz nahmen über 100 Wissenschaftler aus 16 Ländern Europas, Asiens und Amerikas teil, wobei erfreulicherweise auch die osteuropäische quantitative Linguistik - wenngleich, gemessen an Umfang und Bedeutung allein der in den Staaten der ehemaligen Sowjetunion betriebenen quantitativ-linguistischen Forschung, stark unterrepräsentiert - mehrfach vertreten war. In acht Sektionen ("Dialectometry" , "Models and Explanation", "Phonemics and Phonetics", "Process Dynamics and Semiotics", "Quantification and Measurement", "Reports, Projects and Results", "Statistical Studies", "Textual Structures and Processing") wurde auf dieser Konferenz das gesamte Spektrum quantitativ-linguistischer Forschungsrichtungen und -ansätze präsentiert und diskutiert, wobei insbesondere die eingeladenen Vorträge die derzeitigen Schwerpunkte der weltweiten Forschungsaktivitäten hervortreten ließen:

- Prof. Dr. Gabriel Altmann (Universität Bochum), seit mehr als zwei Jahrzehnten die zentrale Figur (nicht nur) der deutschen quantitativen Linguistik, der in seinem Vortrag "Science and Linguistics" eine wissenschaftstheoretische Standortbestimmung der quantitativen Linguistik und eine Analyse ihrer Relation zu anderen sprachwissenschaftlichen Disziplinen unternahm;
- Dr. Kenneth W. Church (AT&T Bell Labs, Murray Hill, USA), der Möglichkeiten und Fruchtbarkeit der Anwendung quantitativer Methoden in der Lexikographie anhand sehr großer Corpora demonstrierte: "Using Statistics in Lexicographic Analysis";
- Prof. Dr. Hans Goebl (Universität Salzburg, Österreich), mit einem Vortrag zu Möglichkeiten und Methoden der rechnerunterstützten Dialektometrie: "Computational Dialectometry";
- Prof. Dr. John S. Nicolis (University of Patras, Griechenland), zur Modellierung dynamischer Zeichenprozesse und ihrer chaostheoretischen Erklärung: "Chaotic Dynamics of the Linguistic Processes: At the Syntactical, Semantic and Pragmatic Levels";
- Prof. Dr. Mildred G. Shaw und Brian R. Gaines (University of Calgary, Alberta, Kanada), die in ihrem Vortrag" A Methodology for Analyzing Terminological and Conceptual Differences in Language Use across Communities", die Resultate ihrer langjährigen Arbeiten zur Entwicklung einer kognitionspsychologisch fundierten Methodologie der quantitativ-empirischen, semantischen Analyse sprachlicher Vermittlung und deren Implementierung vorstellten.

Die 31 Vorträge der acht QUALICO-Sektionen können in folgender Weise thematisch gruppiert und charakterisiert werden:

i. allgemeine Fragen quantitativ-linguistischer Methodologie: R. Grotjahn diskutierte exemplarisch die wesentlichen methodologischen Probleme der Modellierung der Verteilung sprachlicher Einheiten am Beispiel der Wortlänge. Der Vortrag von J. Krilik behandelte Methoden der Skalierung und Klassifikation von Texten.

ii. quantitative Beiträge 'unter Gesetzesniveau' (zu Metrisierung, Klassifikation, Tests, Methoden der Datenerhebung und -repräsentation, deskriptiv-statistischen Befunden etc.) zu einzelnen sprachwissenschaftlichen Teilbereichen und zu 'verwandten' Bereichen: Dialektometrie: multi dimensionale Skalierung als dialektometrische Methode (S. Embleton); Morphosyntax: Metrisierung dekodierungsrelevanter Aspekte von Kongruenz, Rektion etc. zu sprachtheoretischen wie sprachtypologischen Zwecken (P. Schmidt); Verbvalenz/Satzbaupläne: sprachstatistische

Resultate eines polnischen Lexikonprojekts (M. Swidzinski); Verblexikon/-semantik: Merkmalsanalyse und quantitative Klassifikation englischer und deutscher Verben auf der Basis von Merkmalen mehrerer Sprachebenen (G. Sil'nickij); Analyse von Bedeutungswörterbüchern: Charakterisierung von Güte und Repräsentativität auf der Basis 'externer' (Seitenmengen pro Initial, Lemmatamenge pro Initial, etc.) quantitativer Indikatoren (S. Schierholz/ E. Windisch); Sprachgütebeurteilung/ Audiometrie: quantitative Parameter der phonetischen/phonologischen Ausbalanciertheit von Testmaterialien (W. Sendlmeier); konnektionistische Modelle der Sprachproduktion: Modellierung von Selbstkorrekturen (" Reparaturen") von Sprechern in Äußerungen (U. Schade); Mensch-Maschine-Kommunikation: empirische Untersuchungen zum menschlichen Sprachverhalten in Mensch-Maschine-Dialogen (C. Womser-Hacker); quantitative lexikalische Kollokationsanalyse als Hilfsmittel historisch-soziologischer Forschung (M. Olsen); Klassifikation literarischer Prosatexte mit clusteranalytischen Methoden (N. Bolz); Entzifferung: quantitativer Annäherungsversuch an den berühmt-berüchtigten Diskos von Phaistos (D. Rumpel).

iii. quantitative Studien 'auf explanatorischem Niveau', d.h. solche Beiträge, die bereits bekannte probabilistische Sprachgesetze involvieren oder gesetzesartige probabilistische Hypothesen formulieren: Im Beitrag von K. Ejiri und A. Smith wurde ein 'genetisch' auf ZIPFS Gesetz (Frequenzrang-Frequenz-Verteilung von Einheiten) zurückgehendes Maß zur Charakterisierung des inhaltlichen Reichtums von Texten formuliert und validiert. A. Fenk und G. Fenk-Oczlon präsentierten eine sprachtypologische Studie zum MENZERATHschen Gesetz ("Je komplexer eine Einheit, desto kürzer ihre Konstituenten.") als Explanans der Interrelation von Kernsatzlänge, Wortlänge und Silbenkomplexität und einen kognitivistischen Motivationsversuch des MENzERATHschen Gesetzes als Explanandum aus informationstheoretischen Ökonomieprinzipien. L. Hrebiicek skizzierte seinen Ansatz zur Beschreibung von Texten als Strukturen von Koreferenzketten ("Aggregationen") als Textkonstituenten und seine darauf basierenden Untersuchungen (Aggregationen und MENZERATHsches Gesetz, systemtheoretische Betrachtungen zur Textstruktur und -dynamik). A. Polikarpov stellte ein quantitatives Modell des Bedeutungswandels von Wörtern als 'Alterungsprozeß' vor, bei dem deren Polysemiepotential (Fähigkeit zur Annahme neuer Bedeutungen) mit steigender Polysemie abnimmt.

iv. statistische Modelle (hier: MARKOV-Modelle im weiteren Sinne) in der automatischen Sprachverarbeitung: Algorithmen zur Wortartenklassifikation (E. Dermatas/ G. Kokkinakis, R. Kneser/ H. Ney), Effizienzsteigerung von Spracherkennungssystemen durch textabhängige Dynamisierung der Information zu Übergangswahrscheinlichkeiten von Einheiten (U. Essen/ H. Ney), *Hidden Markov Models* der Phonem-Graphem-Zuordnung (P. Rentzepopoulos/ A. Tsopanoglou/ G. Kokkinakis), quantitative Parameter zur Bewertung probabilistischer Sprachmodelle (M. Refice / M. Savino). v. systemtheoretisch (inspiriert)e Ansätze:

a) Der von B. Rieger seit Anfang der 80er Jahre entwickelte prozedural-rekonstruktive Ansatz, der Sprecher-Hörer-Prozesse der Bedeutungskonstitution und des Verstehens in selbstorganisierenden Systemen modelliert und deren kognitive Leistungen als strukturbildende und -verändernde Resultate dynamischer Informations- und Wissensverarbeitung erklärt, war repräsentiert durch Riegers programmatische Darlegung des Forschungsansatzes einer *dynamischen Semiotik* und durch die Präsentation eines kategorientheoretischen Modells der Entstehung und Entwicklung lexikalischer Bedeutung und seiner Implementierung (B. Rieger/C. Thiopoulos).

b) Die Mitte der 80er Jahre von R. Köhler und G. Altmann begründete *synergetische Linguistik*, die einen quantitativ-linguistischen Forschungsansatz zur Modellierung natürlicher Sprachen als selbstorganisierende dynamische Systeme bietet mit der Aussicht, zu einer erklärenden Theorie ihrer Strukturen und Entwicklungen zu gelangen, war vertreten durch Köhlers programmatischen Abriß des synergetischen Ansatzes in der Linguistik, durch eine Variante eines synergetischen Modells der Lexik von R. Hammerl und durch Sambor/Hammerl (s.u).

vi. Überblicksberichte (über größere Projekte oder nationale quantitativ-linguistische Forschung): F. Dupuis, D. Gosselin und B. Habert zu einem Korpus des Mittelfranzösichen, F. Qian zum durch typologische Eigenarten des Chinesischen motivierten Modell der *C[hinese] P[hrase] S[tructure]* G[rammar], J. Reitsma zum Projekt eines Korpus des Friesischen, J. Sambor und R. Hammerl zu Arbeiten zum Polnischen im Rahmen des Bochumer Projekts "Sprachliche Synergetik", P. Saukkonen zur quantitativen Linguistik in Finnland.

Die Organisation und Durchführung einer internationalen und interdisziplinären Konferenz zur quantitativen Linguistik, welche 1990-91 die GLDV und die Herausgeber der QL-Reihe dankenswerterweise übernommen hatten, stellte zweifellos seit längerer Zeit ein Desiderat dar. Denn über eine Reihe von Jahren ließ sich in verschiedenen theoretischen wie angewandten sprach- und kognitionswissenschaftlichen Disziplinen eine zunehmende Tendenz zur Entwicklung von im weitesten Sinne *quantitativ* zu nennenden Ansätzen beobachten. Hierfür scheinen sehr unterschiedliche Motivationen (Effizienzgesichtspunkte, prinzi-

pielle qualitative Schwellen bei der Modellierung von Sprachverhalten, psychologischer oder sogar biologischer Realismus, Einsicht in die essentielle- makro- wie mikroskopische - Plastizität, Variabilität, Unschärfe, Adaptivität und Dynamik natürlicher Sprachen) bestimmend zu sein, womit diese Neuansätze erkennbar teils komplementäre, teils alternative Entwicklungen gegenüber konventionellen Positionen traditioneller Orientierung in den verschiedenen Disziplinen vertreten und verfolgen. Speziell scheint das in verschiedenen konkreten Manifestationen auftretende und sich verbreitende systemtheoretische Paradigma zum ersten Mal die Möglichkeit der Konstruktion einer Sprachtheorie zu bieten, die zugleich erklärende Kraft beanspruchen und diesen Anspruch gemäß den Standards der exakten Wissenschaften auch einzulösen vermag, wobei sie das bestenfalls partielle Erklärungsmuster des CHOMSKYschen Innatismus komplementieren und einbetten könnte.

QUALICO 91 war gekennzeichnet durch ihre Interdisziplinarität und die repräsentative Vielfalt der vertretenen Aspekte der quantitativ-linguistischen Forschung - darunter verschiedene vielversprechende quantitative Ansätze paradigma*tischen*, forschungsleitenden Zuschnitts-, durch fruchtbaren und regen wissenschaftlichen Austausch, eine perfekte Organisation und positive Konferenzatmosphäre (wobei hier der attraktive Rahmen der Konferenz in ihrem reichhaltigen *Social* Program mit Empfang durch die Stadt Trier, Stadtbesichtigungen, Weinprobe, Dampferfahrt nach Saarburg und abendlichem Diner auf der Burg ausdrücklich zu erwähnen und einzuschließen ist). Der beschriebenen, sich verstärkenden quantitativ-linguistischen Tendenz ist mit QUALICO ein internationales und interdisziplinäres wissenschaftliches Forum entstanden, von dem zu hoffen ist, daß es sich institutionalisieren wird und Folgekonferenzen dieser Art erleichtert.

PETER SCHMIDT, *Universität Konstanz*

Tagung:

Information und Klassifikation

Die 16. Jahrestagung der Gesellschaft für Klassifikation fand vom 1.-3. April 1992 an der Universität Dortmund statt. Unter dem Motto 'Information und Klassifikation. Konzepte, Methoden und Anwendungen' fanden eine Reihe von Disziplinen zueinander, die ansonsten kaum miteinander in Berührung kommen: Mathematik, Statistik, Informatik, Medizin, Biologie, Bibliothekswissenschaft, Informationswissenschaft, Psychologie, Sprachwissenschaft, Computerlinguistik, Wirtschaftswissenschaft, Sozialwissenschaft, Musikwissenschaft, Archäologie. Der organisatorische Rahmen umfaßte Plenar- und Übersichtsvorträge, parallele Sektionsvorträge, Workshops, Tutorials und Softwaredemonstrationen-ein kompaktes Programm für nur drei Tage. Die Verfasser sehen sich deshalb außerstande, darüber in aller Ausführlichkeit zu berichten. Der Tagungsbericht gibt deshalb nur die subjektive Auswahl der Verfasser wieder. Wir haben uns auf Beiträge aus den Bereichen Informatik, Informationswissenschaft, Musikwissenschaft, Sprachwissenschaft und Computerlinguistik konzentriert.

Der insgesamt doch recht heterogene Charakter der Tagung spiegelte sich u.a. in der Zusammensetzung der einzelnen Sektionen wieder. So vereinte z.B. das Tutorium "Grundlagen und Nutzungsmöglichkeiten von Computergrammatiken und semantischen Repräsentationsformalismen" einen Bericht über die in den letzten Jahrzehnten gesammelten Erfahrungen in der thesaurusbasierten Auswertung von medizinischen Befunden (W. Giere), einen Überblick über neuere Grammatikformalismen (S. Naumann) und die Vorstellung eines sprachverarbeitenden Systems (SMART) zur semantischen Analyse medizinischer Nominalgruppen und -komposita (J. Ingenerf).

Im Mittelpunkt der beiden Workshops zur medizinischen Linguistik standen Beiträge, in denen über Erfahrungen im Einsatz von medizinlinguistischen Systemen (wie z.B. MUMPS und SNOMED) berichtet wurde. Über zwei in diesem Bereich angesiedelte computerlinguistische Projekte, deren Ziel die praktische Evaluierung semantischer und texttheoretischer Ansätze bildet, berichteten H. Kranzdorf "Automatische Generierung von Sprachtherapieberichten" und M. Schulz "Ein natürlichsprachliches Interface für eine komplexrelationale Datenbank".

Jürgen Kristophson versuchte in 'Ein neuer Beitrag zur Sprach bund diskussion' die Definition von 'Sprachbund: auf eine quantitative Grundlage zu stellen. Über die Textfrequenz bestimmter Merkmale (z.B. postdeterminierend, prädeterminierend) lassen sich Indizien gewinnen, die auf einen möglichen Sprachbund hinweisen.

Stefan J. Schierholz 'Zur Klassifikation von Substantiven nach ihren Determinatoren' lieferte eine Beschreibung der Kookkurrent von Substantiven und Artikeln im Text. Diese Beschreibung steht in Verbindung mit der Entwicklung eines maschinellen Grammatik-Checkers. Als Ergänzung zu einer Klassifikation auf struktureller Grundlage schlug Schierholz eine Angabe von Vorkommenswahrscheinlichkeiten vor.

Heinz J. Weber gab in 'Generierung themenbasierter Links in einem Hypertext-System für Pressenachrichten' einen Überblick über das experimentelle System t-X-t. Auf der Grundlage einer durch Textparsing erstellten Topik-Hierarchie für jeden Nachrichtentext werden inhaltlich ähnliche Topiks ermittelt und miteinander vernetzt.

Eine in Teilen vergleichbare Zielsetzung stellte der Übersichtsvortrag von Gerard Salton 'Automatic Text Linking and Text Grouping Methods' vor. Teiltexte (z.B. Kapitel, Paragraph, Fußnote) aus einer großen Datenbasis (25.000 Artikel einer Enzyklopädie) sollen aufgrund von statistisch ermittelten inhaltlichen Ähnlichkeiten in Affinitätsklassen zusammengefaßt und durch Links miteinander verbunden werden. In einer Retrieval- Umgebung bieten diese Verbindungen Zugänge zu spezifischen Informationen.

In der Sektion *Informationssysteme* 5 stellten D. Fensel und J. Klein drei Algorithmen *Relax, H*-Relax und *I-Relax zur* Generierung allgemeiner Regeln aus einer Menge positiver und negativer Beispiele vor. Ausgehend von den positiven Beispielen wird eine maximal-allgemeine Beschreibung der Zielklasse erzeugt. Im Gegensatz zu anderen Verfahren wie AQ, CABRO oder ID3 verwendet Relax Generalisierung als Suchstrategie. Darüber hinaus wurde noch aufgezeigt, wie Verfahren des maschinellen Lernens mit statistischen Verfahren kombiniert werden können.

In der Sektion *Informationssysteme* 6 - *Wissen-sakquisition* berichtete A. Ultsch über die Integration von selbst-organisierenden neuronalen Netzen (insb. KOHONEN-feature-map) und regelbasierten Expertensystemen. Der Einsatz neuronaler Netze beschränkt sich bei dem vorgestellten Ansatz auf die Extraktion gewisser Regularitäten in den Beispieldaten. Diese Regularitäten werden dann von einem Regelextraktor in PROLOG-Regeln überführt, die dann zusammen mit dem Netzwerk in das Expertensystem als Wissenbasen integriert werden. J. Schrepp beschrieb ein Verfahren, das aufgrund einer Menge natürlichsprachlicher Texte eine kontextfreie Syntax für diese Textmenge erzeugt.

In der Sektion *Neural Networks for EDA and Classification* beschrieb A. Ultsch die Anwendungsmöglichkeiten selbst-organisierender neuronaler Kohonen-Netzwerke im Bereich der explorativen Datenanalyse. Eine fundamentale Eigenschaft der KOHONEN-Netze ist die strukturerhaltende Abbildung des n-dimensionalen Datenraumes auf den 2-dimensionalen Raum der Verarbeitungselemente (units) des Netzes. Um die extrahierte Struktur sichtbar zu machen, wurde die U-Matrix-Methode vorgestellt, die es erlaubt, die Eingabedaten zu klassifizieren.

In der leider nur sehr spärlich besetzten Abteilung *Musikwissenschaft* diskutierte M. G. Boroda die Zweckmäßigkeit verschiedener Gleichheitskriterien für F-Motive. F-Motive sind elementare musikalische Einheiten, die im Gegensatz zum herkömmlichen intuitiven Motivbegriff eindeutig definiert sind. Das ZIPFsche Gesetz liefert ein Indiz dafür, daß Motive nur dann als gleich angesehen werden können, wenn sie bis auf Transposition übereinstimmen. Aus anderen Gleichheitsrelationen, wie z.B. Gleichheit bei rhythmischer Identität resultiert eine Häufigkeitsverteilung, die nicht mit dem ZIPFschen Gesetz vereinbar ist.

Ulrich Franzke demonstrierte recht eindrucksvoll, wie aus einfachen Melodien verschiedene Merkmale abstrahiert werden können, die es erlauben, neue Melodien maschinell zu synthetisieren, ohne daß ihre Künstlichkeit wahrnehmbar ist. Dies ist ein wichtiger Beitrag zur formalen Beschreibung von Eigenschaften, die Melodien als spezielle Klasse von Tonfolgen auszeichnen.

E. LEOPOLD, S. NAUMANN, J. SCHREPP, H.-J. WEBER, *Universität Trier*