Rezension zu

Willée, Gerd/Schröder, Bernhard/Schmitz, Hans-Christian (eds.) (2002):
Computerlinguistik. Was geht, was kommt? Sankt Augustin.
(= *Sprachwissenschaft, Computerlinguistik und Neue Medien* 4)

Kai-Uwe Carstensen (Zürich)

1 Prolog

Die *Einführung in die Linguistische Datenverarbeitung* von Wilfried Lenders war eines der ersten Bücher (vielleicht sogar das erste?), das ich mir vor geraumer Zeit - zu Beginn meines Studiums der Linguistischen Datenverarbeitung in Trier - auslieh. Möglicherweise wäre die vorliegende Festschrift für Wilfried Lenders das nächste gewesen: Welcher Student interessiert sich nicht für die Perspektiven seiner Disziplin (bzw. für seine Berufsperspektiven)? Und, um den Kreis zu schließen, welcher Wissenschaftler ist nicht an den Einschätzungen seiner Kollegen über die Zukunft seines Fachs (in diesem Fall der Computerlinguistik, CL) interesssiert? Schließlich will man keinen Trend verpassen...

Hierin liegt der Reiz des Buches: Nicht weniger als 53 (!) "Personen [...], die die Gegenwart und Zukunft dieser Disziplin prägen [wurden von den Herausgebern] um kritische Resümees und Bestandsaufnahmen, Perspektiven und Visionen gebeten" (Vorwort, S. 10), die jeweils auf durchschnittlich 5 Seiten dargelegt werden. Das Resultat ist im positiven Sinn divers und implizit kontrovers. Nicht auszudenken, hätten sich *mainstream opinions* wie "Statistik ist notwendig, Korpora sind wichtig, praktische Überlegungen haben Priorität, Ontologien werden benötigt, das Semantic Web ist eine große Herausforderung" 45 mal wiederholt, mit vielleicht 8 differenzierenden Gegenstimmen.

So aber geben die Beiträge Auskunft zu Stand oder (möglicher) Entwicklung der Computerlinguistik in den folgenden Bereichen (gemäß der Klassifikation der Herausgeber, S. 11): Computeretymologie, Sprachdokumentation, Sprachtypologie; Computerlexikographie, Ontologien, Korpuslinguistik; Computerphilologie; Dialogsysteme, natürlichsprachliche Agenten, Verstehenstheorie; Forschungsevaluation, Geschichte der Computerlinguistik; Kognitive Sprachverarbeitung; Lehre; Maschinelle Übersetzung; Sprachschnittstellen; Sprachsignalverarbeitung, akustische Kommunikation; Sprachtechnologie: Robustheit, Netzbarkeit, Markt, gesellschaftliche Implikationen; statistische vs. regelbasierte Ansätze; Texttechnologie; Verarbeitung chinesischer Sprache. Ein Inhaltsverzeichnis sowie ein zusätzlicher Artikel von Yorick Wilks finden sich im Übrigen über www.ikp.uni-bonn.de/cl-wgwk/. Die Autoren setzen das Motto des Buches durchaus unterschiedlich um: Bátori etwa liefert eine theoretische Abhandlung, Calzolaris Text entstammt einem "expression of interest", Endres-Niggemeyer/Ziegert und Gippert stellen ein System vor, Büchel wie auch andere geht auf methodische Feinheiten ein, Haenelt präsentiert die Ergebnisse einer Umfrage, und Hahn versucht, die Entwickung der CL in Deutschland anhand von Publikationsdaten in Fachzeitschriften (1990-1999) national und international nachzuzeichnen.

ISSN 1615-3014

Es liegt auf der Hand, dass das Buch nicht auf die übliche Weise exhaustiv besprochen werden kann. Dementsprechend werde ich beispielsweise auf Beiträge, in denen die prominenten Themen der CL behandelt werden, nicht näher eingehen. Hierzu gehören solche aus dem Bereich der Maschinellen Übersetzung (s. z.B. Boitet, Hutchins, Seewald-Heeg, Zimmermann), der Speech-Technologie (s. Furui, Hess, Pitt) oder der Texttechnologie (s. z.B. Lobin, Rösner), sowie solche aus den Ressourcen-orientierten Bereichen (z.B. Calzolari, Endres-Niggemeyer/Ziegert, Heid). Statt dessen werde ich versuchen, einige meiner Ansicht nach interessante Aspekte anhand weniger Meta-Themen einzuordnen.

2 Subdisziplinen und disziplinäre Nischen

Nicht alles, was als CL-Forschungsgebiet gedacht oder geeignet ist, hat auch kurz- oder längerfristig angesetzte praktische Relevanz im Sinne vermarktungsfähiger Software. Nichtsdestotrotz handelt es sich dabei in der Regel um philologisch interessante oder sogar kulturell wichtige Gebiete. Hierzu gehören die *computational etymology* (Bátori) sowie der Aufbau von literarischen (Boggs) und linguistischen (Gippert) historisch-dokumentarischen Korpora und Wissensbasen und deren Nutzung (Gippert, Klein). Gibbon und Müller weisen darauf hin, dass der CL angesichts der Existenz vieler *endangered languages* auch aktuell eine wichtige Rolle in der Sprachdokumentation zukommt. Dokumentation erfordert Beschreibungsmittel und -kriterien, was eine Annäherung an die Typologie-Forschung sinnvoll erscheinen lässt. Tatsächlich schlägt Schulze vor, dass Typologen und CLer in einen dialektischen Diskurs miteinander treten, mit dem Ziel, voneinander zu lernen. Denkt man hier weiter, so würde eine umfassende Sprachdokumentation mit einem im Hinblick auf Diachronie und Synchronie ausgewogenen Theoriegerüst die Möglichkeit eröffnen, *Sprachwandel* zu formalisieren und zu komputationalisieren.

Während Harbusch eine CL-Nische mit eher großer praktischer Relevanz vorstellt (*Unterstützte Kommunikation*), macht Schwanke den Vorschlag, einen Studiengang "Computergestützte Sprachdidaktik" an den Universitäten einzuführen. Obwohl ich den Sinn und die Perspektiven dieser Subdisziplin anhand ihrer Charakterisierung nicht letztgültig einschätzen kann, macht mich schon der Titel skeptisch, sensibilisiert durch einen Satz von Schmitz (der für die Bezeichnung "Computerlinguistik" gedacht ist): "Dass ein Werkzeug überhaupt einer wissenschaftlichen Disziplin seinen Namen geben konnte, rührt von einer eigentümlichen Faszination her, die Computer im vergangenen Jahrhundert ausübten" (S. 251). Auch der abschließende Satz "Absolventinnen dieses Studiengangs finden Arbeitsplätze an Hochschulen und in der Sprachindustrie (z.B. Verlage, Software-Entwicklung, Sprachlehrinstitute)" (S. 262) ist eher ein mir aus eigener Erfahrung bekannter Studierendenköder als eine realistische Berufsperspektive. Spontaner Alternativvorschlag: Hauptfach Mediendidaktik mit zweitem Hauptfach CL oder Hauptfach CL mit Schwerpunkt Lehre bzw. Nebenfach Mediendidaktik. Ansonsten leiden (wieder) die Studierenden. Universitäre Entwicklungen sind zwar sowohl im Hinblick auf Überlebenskampf als auch im Hinblick auf Nischenbildung evolutionär. Dies ist aber eher eine Zustandsbeschreibung als ein Prinzip, das gepflegt werden sollte.

3 Methodische Fragestellungen

Ein Thema, das seit dem "statistical turn" verstärkt diskutiert wird, betrifft die jeweilige Rolle statistischer und regelbasierter Verfahren im Methodeninventar der CL. "Dass die statistischen Verfahren zur Zeit so *en vogue* sind und die regelbasierten Verfahren aussehen lassen wie eine alte Dallas-Folge" (Kiss, S. 170), mag Kiss nicht so stehen lassen und stellt aus diesem Grund einen empirischen Vergleich beider Verfahren an. Das Ergebnis: "Von einer Überlegenheit statistischer Verfahren sollte zumindest im Bereich des Tagging eigentlich nicht gesprochen werden" (S. 170). In einem theoretischen Vergleich der Verarbeitung von Musik und Sprache bringt Hausser den Nachteil eines ausschließlichen Einsatzes statistischer Verfahren auf den folgenden Punkt: "the search space of statistical methods will continue to be so large that it constitutes the biggest single obstacle to ever obtaining successsful automatic speech recognition" (S. 127). Eine Lösung dieses Problems (der CL) scheint ihm nur möglich, wenn der Bezug zu konzeptuellem Wissen im Rahmen eines Sprach*verstehens* hergestellt wird (eine Sichtweise, die komplementär zu dem von Hess vorgebrachten Plädoyer für eine *inhaltsgesteuerte Sprachsynthese* (content-to-speech) ist).

"Im Bereich der Semantik warten sicherlich mittelfristig die größten Herausforderungen auf die Computerlinguistik" (Heid, S. 131), eine Auffassung, die von weiteren Autoren geteilt wird (Eberle, Görz, Helbig, Stede, Weber). Hier setzt das Verlangen nach geeigneten Ontologien und nach Standardisierung der Ausdrucksmittel, nach Robustheit der Verarbeitung (Stede) und nach kognitionsorientierten Semantikformalismen (Helbig, Weber) ein. Görz geht noch einen Schritt weiter, indem er eine pragmatische Wende in der CL einfordert.

4 Spannungsfeld Theorie und Praxis/Anwendung

"Wer Sprachprodukte entwickeln möchte, stelle keinen Computerlinguisten ein" (Heyer, S. 150). Diese Konsequenz Heyers aus der von ihm wahrgenommenen Diskrepanz zwischen "dem Anspruch [der CL], einen Beitrag zur Entwicklung von Sprachprodukten zu leisten, und der rauen Marktwirklichkeit" (ibid.), ist für ihn ein sicheres Indiz dafür, dass die CL nicht, wie scheinbar weithin üblich, in der Philologie angesiedelt sein sollte, sondern in der Angewandten Informatik, aus der solche Produkte momentan im Wesentlichen hervorgehen. Zugegeben, Heyer identifiziert punktgenau die Probleme der disziplinären Einordnung der CL und, daraus folgend, die überwiegende Ausbildungsmisere. Nur ist m.E. seine Lösung falsch. CL in der Angewandten Informatik? O tempora, o mores!

Richtig wäre es, einen angemessenen Status in der Universitätslandschaft sowie eine zufriedenstellende Ausbildungssituation für CLer zu fordern (s. Hellwig). Entsprechende AbsolventInnen dürften einem Angewandten Informatiker dann in fast jeder Hinsicht überlegen sein.

Eberle fordert in einer ähnlichen Stoßrichtung die Abkehr von universitätsüblichen "Spielsystemen" (Prototypen). Auch diese Forderung hat ihre Berechtigung, sind diese Systeme doch in der Regel unfertig und unpraktikabel. Allerdings gibt es wohl kaum Fortschritt ohne ein *Ausprobieren*, dem gegenüber *Spielen* angemesseneren Terminus. Wer

ausgereiftere universitäre Systeme verlangt, der möge bitte langfristigere Planungen, Kontinuität und Perspektiven an den Universitäten ermöglichen.

5 Verhältnis zur Kognition/Kognitionswissenschaft

Die Frage, ob die CL eher der *cognitive science* oder dem *language engineering* zuzuordnen ist, wird explizit von Crocker diskutiert. Klenk ist optimistisch, was den Nutzen sprachorientierter Hirnforschung für die CL anbetrifft. Klabunde verweist auf die Vorteile, die die CL aus einer Berücksichtigung psycholinguistischer Ergebnisse ziehen kann. Umgekehrt argumentieren Schade/Eikmeyer, dass auch die Psycholinguistik von der CL profitieren kann: "Die anwendungsorientierten Systeme können als Ausgangsmodelle in der kognitionswissenschaftlichen Arbeit dienen" (Schade/Eikmeyer, S. 248). Das Fazit, das aus dieser Diskussion gezogen werden kann, wird trefflich von Crocker auf den Punkt gebracht: "Understanding the human capacity for language requires that cognitive science grapple with the scale and complexity of language as addressed in computational linguistics, while many technologies [will] only become successful once they truly reflect the fact that language is ultimately a cognitive process" (Crocker, S. 50f).

Allerdings ist es aus meiner Sicht unangemessen, in der nächsten Zeit auf wesentliche Ergebnisse der Kognitionswissenschaft zu hoffen. So ist beispielsweise die (für kognitionswissenschaftliche Modelle sicher spannende) Gehirnschau noch nicht geeignet, den aktuellen Problemen der CL abzuhelfen. Außerdem existieren kaum interdisziplinär ausgebildete Fachleute, die aus einem entsprechend breit angelegten Blick auf Sprache und Kognition relevante Erkenntnisse produzieren. Vor allem Deutschland weist hier ein Defizit auf und ist nach einem Memorandum der Gesellschaft für Kognitionswissenschaft (s. www.gk-ev.de/memorandum.pdf) ca. 15 Jahre hinter dem Stand der USA.

6 Weiße Flecken

Trotz aller Vielfalt enthält das Buch einige thematische Lücken. Dies ist zum einen auf das Fehlen entsprechend ausgerichteter Beitragender zurückzuführen (z.B. ist der sicherlich perspektivenreiche Bereich des Computer Assisted Language Learning (CALL) nicht vertreten). Zum anderen treffen bestimmte Themenbereiche bzw. -ebenen wie z.B. grammatiktheoretische und implementatorische Fragestellungen schlicht nicht den Zeitgeist. Erstere wären interessant für eine post-revolutionäre Zeit, in der theoretische Erörterungen bzgl. der Beschreibung syntaktischer und semantischer Strukturen wieder eine stärkere Rolle spielen: was wären die Theorien und Formalismen, auf die man sich dann einigen könnte/sollte, d.h., welche von ihnen werden "überleben"?

Im Hinblick auf Fragen, die die Implementationsebene betreffen, wäre ein "Nebukadnezar´s razor" einführenswert: schaffen wir (bzw. schafft die ständig steigende Rechnerperformanz) es, unseren Studenten und AbsolventInnen eine zunehmende babylonische Programmiersprachenvielfalt (C-Varianten, Java, Perl etc.) zu ersparen, so dass man sich langfristig wieder auf der IT-Ära angepassste PROLOG- oder LISP-Varianten o.Ä. zurückziehen kann (auch wenn ein Natural Language Toolkit in Python (nltk.sourceforge.net) seinen Reiz hat)? Oder

wird es auch in Zukunft eine Trennung von Forschungs- und Praxis- Implementationsgepflogenheiten und -standards geben (bzw. zumindest eine aufgabenspezifische Hybridizität in der Verwendung von Programmiersprachen)? Dies würde für einen CL-Absolventen, der neben multidisziplinärer theoretischer und methodischer Kompetenz auch über implementatorische Fertigkeit verfügen soll, einem Spagat mit Schwergewichten gleichkommen. Auch hier, befürchte ich, fehlen die für eine unvoreingenommene Beantwortung dieser Fragen notwendigen kompetenten Fachpersonen.

Als weitere Lücke empfindet man/empfinde ich das Fehlen weitreichender Prognosen, sozusagen die Perspektive in den Perspektiven. Die Beitragenden bleiben meist eher im Nahbereich des Ist-Zustands, obwohl der Titel des Buches doch gerade auch auf langfristigere, mutigere Prognosen hoffen lässt. Zwar berichtet Haenelt über die Ergebnisse einer Umfrage zur Zukunft mit der Sprachtechnologie, jedoch nur Paprotté, Portele und Schmitz wagen eine breit angelegte Prognose (Zimmermann beschränkt seine immerhin bis 2020 reichende auf die MÜ). Dabei geht Schmitz am weitesten, wenn er behauptet, "dass es 'Computerlinguistik' in fünfzig Jahren nicht mehr geben wird, jedenfalls nicht unter diesem Namen und nicht als eigenständige Disziplin" (S. 251). Ist, wie er behauptet, der Terminus "Computerlinguistik" wirklich im Prinzip jetzt schon überholt (wegen angeblicher Referenz auf das bis dato veraltete Werkzeug Computer)? Man schaue hierzu nur auf die englische Bezeichnung der Disziplin (*computational linguistics*), die ein anderes, abstrakteres und somit beständigeres Verständnis suggeriert bzw. impliziert.

Und wird Computerlinguistik wirklich von den IT-gereiften Sprachwissenschaften absorbiert werden, wie Schmitz behauptet? Ich bewundere das hier zutage tretende Vertrauen in die Wandelbarkeit disziplinärer Strukturen, das mir allerdings nicht gerechtfertigt zu sein scheint. Meine Alternativprognose: alles bleibt mehr oder weniger so, wie es ist. Sicher, es wird Veränderungen geben. Disziplinäre Diversifikation mit terminologisch indizierter Grüppchenbildung (von Linguistischer Datenverarbeitung über linguistische Informatik bis zur Texttechnologie) gab es jedoch immer, eine Unbeständigkeit, die sich fortsetzen wird. Mein Tipp: nach Sprach- und Texttechnologie kommen noch *Dialogtechnologie, Diskurs- und Dialogtechnologie* und/oder *Intelligente Kommunikationstechnologie* hinzu. Und *computational semiotics, computational teaching* o.Ä. Sicherlich existiert das eine oder andere davon sogar schon irgendwo.

Möglicherweise wird es auch Fusionen geben. Z.B. die der Computerlinguistik mit der Künstlichen Intelligenz, wenn inhaltsrelevante Fragen (Ontologien, Semantik, Inferenzen) stärker in den Vordergrund kommen und zu einer Wende führen. Oder mit der Kognitionswissenschaft. Viele Claims sind jedoch schon abgesteckt und benannt, und die Änderungen werden eher (kleine) Umordnungen sein. Beispielsweise existiert bereits eine Schnittmenge von Computerlinguistik und Kognitionswissenschaft bzw. KI als *natural language processing*, wenn man den Terminus so verstehen will. Sprachwissenschaftler werden vielleicht doch eher Sprachwissenschaftler im engeren Sinne bleiben, die sich nachrangig für kognitive und verarbeitende Aspekte interessieren, auch wenn sie zunehmend computerlinguistische Ressourcen und Tools verwenden werden.

ISSN 1615-3014

Ein weiterer Bereich, in dem es an langfristiger Perspektive mangelt, ist der der Lehre. Hiermit meine ich nicht CALL, nicht CL-E-Learning (Handke), und nicht die aktuelle Diskussion der CL-Ausbildung (Hellwig). Vielmehr geht es zum einen darum, die *CL innerhalb der IT* stärker herauszuarbeiten als die Disziplin, die für den Bereich Sprache, Text und Kommunikation verantwortlich zeichnet. Aufgrund der Prominenz dieses Bereichs lässt sich prinzipiell ein allgemeiner und somit sehr viel umfangreicherer Anspruch auf Anteile an der Lehre in IT-Disziplinen einklagen, als zur Zeit überhaupt angedacht wird (s. aber z.B. Hellwig und Dale et al. 2003). Und noch weiter gedacht: Könnte es nicht sein, dass die Unkenntnis bestimmter CL-Inhalte aufgrund ihres Einsatzes in allgemein verwendeter Sprachtechnologie irgendwann auch schon in der Schule als Analphabetentum angesehen wird? Dann müssen die Lehrer auf breiter Basis entsprechend ausgebildet werden...

Zum anderen geht es darum, die *Anwendungsmöglichkeit der CL im Bereich Ausbildung und Lehre* zu erkennen: Was heute noch als (web-basierte) Multimedia-Applikation zur Vermittlung von Inhalten daherkommt, stellt sich in 20 Jahren vielleicht schon als flexibles Lehrtool dar, das, wenn auch eingeschränkt, in Dialoge mit Lernenden tritt, und zwar nicht nur im Bereich des Sprache-Lernens, sondern auch z.B. im Rahmen eines Jura-, Zahnmedizin- oder Ökotrophologie-Studiums sowie in der Schule.

Ein letzter Perspektivenmangel, der hier angemahnt werden soll, betrifft die *Ziele der CL*. Offenbar übertrifft man sich zur Zeit in der *community* damit, alles vermutlich und vermeintlich Unmachbare vor die Tür des Hauses zu setzen. Prominentes Beispiel ist die MÜ: Fully Automatic High Quality Translation (FAHQT) ist unter Berufung auf shifts, turns und revolutions nicht einmal mehr als Problem salonfähig. Wenngleich dies angesichts des gegenwärtig Vorstellbaren nachvollziehbar ist, frage ich mich, ob die CL wirklich in Richtung Fachhochschulniveau einer besseren IT-Support-Technologie abdriften will oder ob sie sich einige hehre Ziele leistet und zu diesen steht, weitab von dummen Behauptungen, dass diese Ziele bald erreichbar sind, und unter gleichzeitigem Verzicht auf realitätsferne Curricula.

7 Zusammenfassung

Das Zauberwort für die nächsten Jahre lautet *Standardisierung*. Ohne gemeinsame Bemühungen, Dinge auf einen gemeinsamen Nenner zu bringen und Gleiches beim selben Namen zu nennen, wird die Zukunft der CL von Verzettelung und Redundanz geprägt sein, und die Lösung zentraler Probleme wird weiter auf sich warten lassen. Dies gilt für alle relevanten Ebenen (Theorie, Methode, Implementation).

Die Disziplin CL sollte stärker durch ein *sowohl-als auch* geleitet sein: sowohl Statistik wie Regeln sind sinnvoll, sowohl Theorie/Kognition als auch Praxis haben ihren berechtigten Platz. Einseitige Schwerpunktsetzung wird irgendwann von der Realität überholt und führt in eine Sackgasse.

Fragen der *Semantik* (d.h. der Beziehung sprachlicher und konzeptueller Strukturen) werden an Bedeutung zunehmen. Hierzu sollte die CL zwecks Vermeidung eines Dornröschenschlafs nicht einseitig auf die Ergebnisse der formalen/logischen Semantik bauen, sondern deren Resultate dankend zur Kenntnis nehmen und unter Berücksichtigung der Erkenntnisse

weiterer Disziplinen (Informatik, Linguistik, KI, Kognitionswissenschaft) moderne, einsetzbare Antworten produzieren.

Neben vielen Anwendungsgebieten einer reifen Sprachtechnologie ist *Lehre* ein Bereich, der für den Einsatz intelligent interagierender tutorieller Werkzeuge besonders breite Perspektiven eröffnet.

Ein wichtiges Element in der weiteren Entwicklung der CL wird die *Ausbildung ihrer AbsolventInnen* sein. Ideal ist ein eigenständiger Diplom-/Master-Studiengang mit Bachelor-Vorstufe, der als unabhängige *interdisziplinäre* Einrichtung außerhalb der Linguistik oder Informatik angesiedelt ist, der aber Lehrleistung sowohl importiert als auch exportiert. Praxisorientierte Aspekte mit entsprechenden Ausbildungsanteilen (z.B. Software-Engineering) sollten eine wichtige, aber mit theoretischen und methodischen Anteilen nur gleichwertige Rolle spielen.

Selbstverständlich ist dies eine sehr subjektiv gefärbte Zusammenfassung, die mit keiner der im Buch geäußerten Meinungen vollständig übereinstimmen muss. Der geneigte Leser mag sich selbst davon überzeugen.

Fazit: Dieses Buch regt zum Vor- und Nachdenken an, was mehr ist, als man generell von einem Buch erwarten kann. Ich empfehle es keinem, der sichere Erkenntnisse in seinem Regal stehen haben möchte, aber jedem, der sich für die Disziplin CL interessiert.

Literaturangabe

Dale, Robert/Mollá Aliod, Diego/Schwitter, Rolf (2003): "Natural Language Processing in the Undergraduate Curriculum". In: Greening, Tony/Lister, Raymond (eds.): *Proceedings of the Fifth Australasian Computing Education Conference (ACE2003)*. Adelaide, South Australia: 9-13.