Rezensionen

Sinnentstellender Fehler in letztem LDV-Forum: Rezension zu „Lexikon und Text“ (Feldweg/Hinrichs) sowie „Lexikonimport, Lexikonexport“ (Hoetker/Ludewig)

In der letzten Ausgabe ist auf schwer nachvollziehbare Weise in der Sammelrezension der beiden Niemeyer-Bände über Lexika (Feldweg/Hinrichs: Lexikon und Text sowie Hoetker/Ludewig: Lexikonimport, Lexikonexport) ein schwerer, systematischer Fehler unterlaufen, für den Autor und Redaktion alle Beteiligten und alle Leser um Entschuldigung bitten. Die Kürzel, mit denen auf die beiden Werke verwiesen wurde, waren nämlich grundsätzlich vertauscht verwendet worden. Um jegliche Unklarkeit auszuschließen, folgt nun direkt die korrekte Fassung der Rezension als Nachdruck. Außerdem liegt dem Heft ein Korrekturblatt bei, der in die Ausgabe LDV-Forum 1998/1 einsortiert werden sollte.

Feldweg, H. und Erhard W. Hinrichs (Hrsg.):

Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschliessung des Deutschen. *[LSM]*

Lexicographica Series Maior 73. Tübingen 1996, Niemeyer Verlag

Hoetker, W. und Petra Ludewig (Hrsg.):

Lexikonimport, Lexikonexport. Studien zur Wiederverwendbarkeit lexikalischer Informationen. *[SI]*

Sprache und Information, Bd. 31. Tübingen 1996, Niemeyer Verlag

Rezensiert von Johann Haller, Saarbrücken (e-mail: hans@iai.uni-sb.de).

Gleich zwei Bände mit fast gleichem (Unter-)titel zur gleichen Zeit im gleichen Verlag: Das muß doch ein wichtiges Thema sein – oder hat da einer bei Niemeyer nicht aufgepaßt? Aber bei näherem Hinsehen ergibt sich, daß nur ein Artikel gleichzeitig in beiden Bänden zu finden ist, und immerhin einmal in Englisch und einmal in Deutsch.

Aus welchen theoretischen und praktischen Gründen das Thema der Wiederverwendbarkeit so wichtig ist, erfährt man in sehr übersichtlicher und systematischer Weise in der Einleitung von SI, das neun längere und in drei nach Forschergruppen getrennte Aufsätze enthält. Zu einem theoretischen Umschwung (mehr lexikalistisch begründete Sprachtheorien) kommt die Zunahme der praktischen Relevanz lexikalischer Ressourcen, die sich auch in der steigenden Anzahl von EU-Projekten zu diesem Thema zeigt. Einige davon (EAGLES, DELIS, MULTEXT, ET10-51) sind zwar dem Spezialisten bekannt, es ist aber sehr verdienstvoll, daß gerade die Ergebnisse dieser Projekte ausführlich und im Zusammenhang in deutscher Sprache zugänglich gemacht werden. Gerade die mangelnde Sichtbarkeit dieser und ähnlicher Projekte in der Öffentlichkeit war mehrmals kritisiert worden. Das aktuelle mehrsprachige Korpus- und Lexikonprojekt der EU (PAROLE) sollte eigentlich die Fortsetzung sein; daß die Fortwirkungen dieser Projekte sich besonders im deutschen Teilprojekt (PAROLE-D) in Grenzen halten, mag zunächst darauf zurückzuführen sein, daß sich die genannten EU-Projekte hauptsächlich mit der englischen Sprache beschäftigten. Hier und da taucht mal ein deutsches Beispiel zwischen den „be"-Formen und den syntaktischen Verwendungsmöglichkeiten von „taste" auf. Außerdem sind zwei Jahre in der Computerlinguistik ein langer Zeitraum (die entsprechenden Workshops fanden 1994 statt), in dem Ergebnisse schnell veralten. Beide Bände (SI und LSM) machen eigentlich neugierig auf die weiteren Fortschritte; das studentische Beispielprojekt in einem Osnabrücker SI-Artikel, das sich mit der Anwendung des mehrfach verwendbaren Lexikons in Sprachlernsystemen befaßt, hofft darauf, daß die (eigentlich altbekannten) Schwierigkeiten bald überwunden werden: Disjunktions- und Negations-Operatoren für den Formalismus, Anbindung an eine große lexikalische Datenbank für partielle Lexikoneinträge etc. Ein wichtiger Schritt in diese Richtung ist sicher mit der in den drei Bochumer Beiträgen beschriebenen Bearbeitung des Cobuild-Lexikons geschehen; dieser wird von Schnelle auf der Hypothese der Bedeutung als Implikation theoretisch fundiert, und von den beiden anderen Autoren am Beispiel des HPSG-Formalismus ausgearbeitet und technisch umgesetzt. Ein Teil des Ergebnisses ist (noch kosten-

frei) im WWW zu besichtigen (http://ww.linguistics.ruhr-uni-bochum.de/ccsd/).

Einen diffuseren Eindruck macht zunächst LSM, das eine Reihe kürzerer Artikel enthält; die in der kurzen Einleitung unternommene Gliederung in sechs Bereiche hilft da nur wenig weiter: Lexikon-Text, Wörterbücher, Wissensrepräsentation, Korpora, Linguistische Annotation, Statistische Disambiguierung. Zur lakonischen Kürze kommt auch noch ein Druck- bzw. Grammatikfehler in der ersten Zeile. Um die formale Kritik gleich auf einmal loszuwerden: Druckfehler, leider auch sinnentstellende, gibt es eine ganze Menge. Daß mit „Beispielgruppenabgabe" eigentlich die entsprechende „Angabe" gemeint ist, bedarf schon einiger Überlegung, bei „müßen" mit scharfem S war wohl die neue Rechtschreibung irgendwie schuld. Und schließlich meinen noch ein paar Autoren, sie müßten ihre Lateinkenntnisse an den Mann bringen. Auch hier kann man „computativ" noch erschließen, aber bei „Fehlende Lexika sind ein Desiderat" hätte man es auch einfacher haben können. Vom Inhalt her findet man in manchen Artikeln Interessantes: immer noch die Bonner Wortdatenbank (eine Pionierleistung, wenn auch inzwischen ein bißchen betagt), endlich ein paar Lichtblicke aus dem Institut für deutsche Sprache in Mannheim (jetzt mit dem Projekt COSMAS auch im Internet zu erproben: http://www.ids-mannheim.de/ldv/cosmas/), von dem man schon seit vielen Jahren computerlinguistische Aktivitäten erhofft hätte, Information über das eindrucksvolle CISLEX in München (die deutsche Entsprechung zum LADL-Lexikon von Maurice Gross in Frankreich), Information zu einem großen Korpus-Projekt der EU (MULTEXT) und für Programmierfans eine Reihe technischer Beschreibungen, wie man zu einsatzfähigen Programmen kommt, die auch mit großen Text- und Wortmengen umgehen können. Schwedisch-Fans werden auch bedient; für diese Sprache – und dann auch für Deutsch – werden Vorschläge für Tagsets gemacht, über die aber die Zeit vermutlich schon wieder hinweggegangen ist. Daß doch gelegentlich auch auf alte Materialien zurückgegriffen wird, zeigen die vereinzelten Verweise auf alte Saarbrücker Wörterbücher, Analyse- und Lemmatisierungsprogramme, die bei der Konfektion umfangreicher Vorlagen zur Anwendung kommen. Solche werden bei den Versuchen des stochastischen Tagging benötigt, deren Beschreibung den Abschluß von LSM bildet. LSM hat am Schluß einen lieblos gemachten (und damit ein bißchen nutzlosen) Index, in dem ‚Wiederverwertbarkeit' gerade einmal auftaucht, dafür „leere" Wörter wie ‚Zeichen', ‚ZEIT' und einige Varianten linguistischer Begriffe wie ‚Wortartklassifikation', ‚Wortartenklassifikation' etc. Hier wäre es sicher angebracht, sich einmal auf eine Schreibweise festzulegen und diese dann beizu-

behalten. Da hätte man es lieber wie SI machen sollen und den Index ganz weglassen.

Insgesamt betrachtet, kann der fortgeschrittene Student der Computerlinguistik oder der Forscher, der etwas über den Stand der maschinellen Lexika in Deutschland wissen will, durch die Lektüre der beiden Bände einen recht guten und vollständigen Überblick bekommen; es fragt sich natürlich, ob in Zeiten der elektronischen Publikation wirklich fast zwei Jahre vergehen müssen, bis die Ergebnisse eines schnell fortschreitenden Gebietes auf Papier vorliegen. Der Rezensent schwankt noch, was ihm mehr Spaß macht: die schnelle und aktuelle Info via Internet oder die bequeme Lektüre der Bücher. Nach der Lektüre kann man noch ein bißchen träumen: Wie wäre es, wenn COSMAS eine (syntaktische) Kontextsuche à la DELIS hätte, wenn es ein standardisiertes Tagset für das Deutsche gäbe, das auch in der von Bochum ins Internet gestellten DUDEN-Version enthalten wäre und so weiter ...