Computerlinguistik für Philologen

M. A. Moreaux / Inalco Paris

Es läßt sich nicht mehr bestreiten, daß Computerlinguistik ein weitgehend interdisziplinäres Fachgebiet darstellt. Unter den häufig aufgezählten Kenntnissen, die ein Computerlinguist beherrschen muß, werden Kenntnisse aus Linguistik und Informatik mit Recht als grundlegend betrachtet. Von entscheidender Bedeutung scheint jedoch ein drittes Gebiet, das selten oder nie sehr deutlich erwähnt wird, so, als ob man es für ganz selbstverständlich hielte: Hierbei handelt sich um alles, was Sprachfähigkeit und Sprachkenntnisse betrifft. Wir möchten deswegen ein besonderes Augenmerk auf den Stellenwert von Sprachkenntnissen in der LDV bzw. Computerlinguistik richten.

Wichtig zu bemerken ist, daß der Gegenstand der maschinellen Sprachverarbeitung nicht Sprache an sich ist, sondern eine Folge von in einer bestimmten Sprache ausgedrückten Äußerungen. Ein Sprachverarbeitungssystem muß also über ein sehr präzises und detailliertes Wissen über die Elemente und Strukturen der betroffenen Einzelsprache verfügen und dürfte um so robuster sein, desto mehr es in der Lage ist, aus seinen eigenen Sprachdaten Eigenschaften sprachlicher Einheiten zu errechnen. Daraus ergibt sich, daß eine maschinell interpretierbare Sprachdarstellung nicht nur völlig explizit formuliert werden muß, sondern auch bis ins einzelne alle Regelmäßigkeiten und Ausnahmen der zu behandelnden Phänomene beschreiben muß, auch die Regularitäten in den Irregularitäten.

Selbstverständlich müssen Sprachforschungen durch Korpora unterstützt werden. Deren Gebrauch dürfte jedoch Sprachkenntnisse nicht ersetzen, sondern von diesen Kenntnissen gesteuert werden, um eine Verfeinerung bzw. ein Ergänzen oder die Auswertung der Darstellung zu erlauben.

Maschinell interpretierbare Sprachdarstellungen sollen auch nicht bloße Aufzählungen von aus Texten und Reden direkt beobachtbaren Sprachfakten sein. Sie müssen theoretisch geleitet werden und implementierbar sein.

Die Erfordernisse der maschinellen Sprachverarbeitung zwingen dem Sprachwissenschaftler also eine strenge Methodik, eine sehr analytische Denkweise und eine ganz besondere Einstellung zur Sprache auf. Infolgedessen kann sich

eine Modellimplementierung nicht aus der Zusammenarbeit von Spezialisten mit unterschiedlichen Kompetenzen ergeben. Von einem Computerlinguisten wird unbedingt eine Mehrfachkompetenz erwartet. Er muß in der Lage sein, die vollständige Entwicklung eines computerorientierten Sprachmodells auszuführen, von der Problemstellung über den Entwurf einer linguistischen Lösung bis zur Implementierung.

Hierbei sollte man insbesondere darüber nachdenken, welcher Art die Sprachkenntnisse eines Computerlinguisten sein sollen. Die erwähnten Anforderungen setzen eine eingehende Kenntnis einer Einzelsprache voraus. Man benötigt darüber hinaus aber auch sprachwissenschaftliche Kenntnisse von der zu beschreibenden Sprache, auch wenn sie die Muttersprache ist, denn die als Muttersprachler erworbene Sprachfähigkeit scheint hier nicht ausreichend zu sein.

Äußerst wichtig ist, daß eine größere Zahl von Philologen Interesse an maschineller Sprachverarbeitung gewinnt und in die Computerlinguistik einsteigen möchte. In dieser Hinsicht ist das INALCO (Institut National des Langues et Civilisations Orientales*) in Paris eine sehr günstige Umgebung. Mit Ausnahme von westeuropäischen Sprachen können dort 80 Sprachen aus allen Erdteilen erlernt werden. Da diese Sprachen an höheren Schulen nicht unterrichtet werden, wird jeder am Institut immatrikulierte Student als Anfänger betrachtet. Je nach Sprache erstreckt sich das Anfangsstudiums über zwei bis drei Studienjahre† und wird mit einem Diplom (genannt DULCO) abgeschlossen.

Das computerlinguistische Curriculum am INALCO wurde speziell für Sprachenstudierende entwickelt und setzt nach einem absolvierten sprachlichen Grundstudium von zwei bis drei Jahren ein. Der Studiengang, der somit erst Bestandteil des Hauptstudiums ist, umfaßt ein

1.) computerlinguistisches Grundlagenstudium, das zwei in sich abgeschlossene Teile enthält, die jeweils ein Studienjahr dauern: der erste Abschluß ist die „Licence“ und der zweite die „Maîtrise“.

2.) Promotionsstudium, das ebenfalls zweiteilig ist. Ein erster Teil, der ein Studienjahr dauert, führt zu dem als DEA bezeichneten

* Wortwörtlich wäre diese Bezeichnung als „Staatliches Institut für orientalische Sprachen und Kulturen“ zu übersetzen. „Orientalisch“ versteht sich hier aber viel breiter als üblich und betrifft nicht nur den geographischen Orient.

† In Frankreich dauert ein Studienjahr von Oktober bis Juni und umfaßt 27 Studienwochen.

Abschlußdiplom und ist als der erste Schritt in die Forschung zu betrachten. Der zweite Teil dehnt sich über einen Zeitraum von 3 bis 5 Jahren aus. Während dieser Zeit erarbeiten die Promovenden ihre Dissertation.

Alle Absolventen des Grundstudiums einer beliebigen Philologie* können sich im Studiengang „Computerlinguistik" einschreiben. Diese Studenten haben die jeweils studierte Sprache schon recht gut erlernt, kennen ihre Grammatik und sind infolgedessen in der Lage, über diese Sprache als Sprachsystem nachzudenken. Bemerkenswert ist, daß sie nicht selten mehrere Sprachen studiert haben. Das kann nur von Vorteil sein, denn damit werden sie darauf vorbereitet, Funktionsunveränderlichkeit an unterschiedlichen Äußerungsformen zu erkennen. Meistens haben sie aber noch kaum Einblicke in die (allgemeine) Linguistik bekommen. Es wird von den Studenten auch nicht erwartet, daß sie Kenntnisse aus dem Bereich der Informatik besitzen.

Angesichts des Wissens und der Fähigkeiten der in die Computerlinguistik einsteigenden Studenten werden die angebotenen Lehrveranstaltungen während der beiden ersten Studienjahre auf die Grundlagenausbildung gerichtet. Im allgemeinen sind die Studenten der philologischen Fächer weder an Interdisziplinarität, noch an die mathematisch präzise Denkweise gewohnt, die Computerlinguistik erfordert. Deswegen geht es darum, Grundlagen zu vermitteln, durch die der Wissensstand der Studierenden in bezug auf interdisziplinäres Wissen erweitert und sie, wenn man es so sagen darf, in eine andere Denkweise einführt. Der Schwerpunkt liegt auf den Fächern, die von einem Philologen als neu empfunden werden und ihm die größte Mühe bereiten: alles, was sich auf Formalisierung, Algorithmenbeschreibung und Programmierung bezieht.

Die Licence- und Maîtrise-Lehrpläne sind in vier Komponenten untergliedert. Eine betrifft die Sprache, denn jeder Student, der sich für ein Computerlinguistikstudium entscheidet, muß einen Teil seiner Lehrveranstaltungen im Bereich „Sprache" absolvieren. Jede der drei weiteren Komponenten entspricht der Vermittlung von Grundkonzepten, Methoden und Verfahren, die zu den beteiligten Disziplinen gehören und deren Kenntnis notwendig ist, um bestehende CL-Modelle verstehen und evaluieren zu können oder solche Modelle selbst zu entwik-

* Nicht nur Studierende einer „orientalischen" Sprache, sondern auch jeder westeuropäischen Sprache (Französisch, Deutsch, Englisch, …).

keln. Alle Lehrveranstaltungen sind Pflichtveranstaltungen und werden in Form von Vorlesungen und Übungen angeboten:

1.) Sprache (Licence-Studiengang: 100 Std.*; Maîtrise-Studiengang: 50 Std.): Dient der Vertiefung der Sprachkenntnisse.

2.) Sprachwissenschaft (Licence-Studiengang: 100 Std.; Maîtrise-Studiengang: 63 Std.): Einführung in die Grundlagen der Sprachwissenschaft. Die eingeführten Begriffe und Methodologien werden dann auf die Einheiten der verschiedenen sprachlichen Ebenen (Phonetik/Phonologie, Morphologie, Syntax, Semantik aber auch Lexikologie) angewandt. Nach einem kurzen Überblick über Anwendungsgebiete und Forschungsrichtungen der CL versucht eine der linguistischen Veranstaltungen, die Verhältnisse zwischen den verschiedenen Fächern zu skizzieren und hierbei die Studenten mit den Bedingungen einer computerorientierten Sprachbeschreibung vertraut zu machen.

3.) Sprachverarbeitung bzw. Computerlinguistik (Licence-Studiengang: 75 Std.; Maîtrise-Studiengang: 125 Std.): Den thematischen Schwerpunkt bildet hier die Behandlung formaler Modelle und formaler Darstellungsverfahren. Die Licence-Veranstaltungen führen in die den Philologen meistens fehlenden Grundlagen der Mathematik und der Logik ein (moderne Logik, Mengentheorie und Relationskalkül). Diese Grundkenntnisse werden dann in den Maîtrise-Veranstaltungen vertieft und durch die Darstellung der nicht-klassischen Logiken, der Theorie formaler Sprachen und der Automatentheorie erweitert. Zum Schluß werden auch Parsing-Strategien behandelt.

4.) Informatik und Programmierung (Licence-Studiengang: 88 Std.; Maîtrise-Studiengang: 150 Std.): Nach einer kurzen Einführung in den Aufbau und das Funktionieren eines Computers wird das Hauptgewicht auf Entwurfsprinzipien von Algorithmen und Datenstrukturen gelegt. Die Grundlagen der strukturierten Programmierung werden durch das Erlernen einer ersten

* Stunden pro Studienjahr

Programmiersprache vermittelt. Dabei ist C die hier gewählte Programmiersprache. Darüber hinaus werden dann im Rahmen des Maîtrise-Studiengangs auch die Konzepte der logik- und objektorientierten Programmierung dargestellt und Prolog und C++ erlernt. Die Studenten erhalten die Aufgabe, Basisalgorithmen der Sprachverarbeitung zu erarbeiten oder eigene Lösungen zu erstellen, die sie dann in einer Programmiersprache umsetzen müssen.

Die DEA-Veranstaltungen sind auf speziellere Kenntnisse bezogen, wie z.B. auf formale Theorien und Beschreibungsformalismen in der Computerlinguistik, die theoretisch und mathematisch komplex sind und das früher vermittelte Wissen unbedingt voraussetzen. Ein Computerlinguist muß natürlich in der Lage sein, abschätzen zu können, ob eine Grammatiktheorie und ein Beschreibungsformalismus zur Abbildung und Erklärung der zu behandelnden Phänomene besser als andere geeignet sind. Die Beteiligung von Forschern anderer französischer Hochschulen (Grenoble und Nizza), aber auch aus mehreren europäischen Ländern (Deutschland, Tschechische Republik, Großbritannien, Belgien) ermöglicht den Studenten, einen weitreichenden Überblick über zahlreiche Bereiche der Computeringuistik zu gewinnen.

Der DEA-Lehrplan umfaßt vier Lehrveranstaltungsgruppen. Jede ist einem besonderen Thema gewidmet und wird mit einer Prüfung abgeschlossen:

1.) formale Beschreibungsmodelle (100 Std.)

2.) Methoden zur morphologischen, syntaktischen und semantischen Analyse (100 Std.)

3.) Anwendung von KI-Methoden und Verfahren im Bereich der Sprachverarbeitung (50 Std.)

4.) Anwendungsgebiete der Computerlinguistik und Systeme (125 Std.)

Der Studierende wird in eine Forschungsgruppe integriert, in der er an selbständiges wissenschaftliches Arbeiten gewöhnt wird. Zum Abschluß muß er sich inhaltlich und technisch mit einer bestimmten Problemstellung auseinandersetzen. Dabei erarbeitet er ein begrenztes aber nicht-triviales Phänomen der studierten Sprache, muß seine Ergebnisse in einer längeren schriftlichen Hausarbeit

darstellen und seine Lösung in ein Programm umsetzen. Zu den derzeit im Rahmen solcher Arbeiten behandelten Sprachen gehören Französisch, Englisch, Deutsch, Arabisch, Tschechisch, Italienisch, Malaiisch und Japanisch.

Das skizzierte Ausbildungsprofil wurde 1979 nach einem Besuch von Patrice Pognan, dem Leiter des CERTAL (Centre d'Etudes et de Recherche en grammaire et Traîtement Automatique des Langues*), in Hamburg entwickelt und ist in vielerlei Hinsicht Ergebnis der Diskussionen mit Walther von Hahn. Es ist in den nachfolgenden Jahren durch zahlreiche eigene Forschungs- und Lehrerfahrungen ergänzt und den Bedürfnissen der Studenten des INALCO angepaßt worden.

*eine Forschungsgruppe, die ihren Sitz im INALCO hat