DER AUFBAU MEHRSPRACHIGER LEXIKA MIT DEM LEXIKONSYSTEM "MULI"

EUROTRA-D-Teilprojekt am Institut für Kommunikationsforschung und Phonetik (IKP), Bonn

Seit 1985 wird vom Bundesministerium für Forschung und Technologie (BMFT) das EG-Projekt zur Entwicklung eines maschinellen Übersetzungssystems (EUROTRA) geördert. Als Teilprojekt des EUROTRA-D Projekts in Saarbrücken werden am Institut für Kommunikationsforschung und Phonetik (IKP Bonn) die Transferlexika zu den Sprachen: Dänisch, Griechisch, Niederländisch und Italienisch aufgebaut. Damit verbunden ist die Aufgabe, Software zu entwickeln und einzusetzen, die eine weitreichende Computerunterstützung der lexikographischen Arbeit leistet. Der folgende Beitrag soll die bei der lexikographischen Arbeit eingesetzte Software vorstellen und die Vorgehensweise bei der Lexikonerstellung erläutern.

Im Projektzeitraum (1.9.85-30.4.87) wurden die Transferlexika griechisch-deutsch, dänisch-deutsch und niederländisch-deutsch in einem Umfang von ca. 2.500 Worteinträgen bearbeitet. Als Ausgangsbasis diente hier der in alle EG-Sprachen übersetzte Text des ESPRIT-Programms der EG (kurz ESPRIT-Text). Für die Lexikonbearbeitung wurde das Lexikonsystem MULI (MUlti LIngual Dictionary System) entwickelt und zu einem rechnergestützten Arbeitsplatz für den Lexikographen ausgebaut.

Das Lexikonsystem MULI erlaubt es, die bilingualen Lexika in EUROTRA zu erstellen, zu vergleichen und zu verwalten (Abb. 1). Neben den Grundfunktionen Erfassen, Korrigieren und Löschen von Lexikoneinträgen zu einem ausgewählten Sprachpaar unterstützt MULI den Zugriff auf Lexikoneinträge anderer Sprachpaare und ggf. die Übernahme von Teilinformationen in das gerade bearbeitete Lexikon. Das Importieren und Exportieren lexikalischer Daten (von einer Datenbankdatei bzw. in eine ASCII-Datei) wird durch zusätzliche Module realisiert.

Die Erstellung verschiedener bilingualer Lexika mit gemeinsamen Sprachkomponenten wirft das Problem einer konsistenten Übersetzung der verschiedenen Lesarten eines Wortes auf. Die durch Lesartnummern gekennzeichneten Lesarten sollten im Hinblick auf die gemeinsame Nutzung der Lexika in einem Gesamtsystem durchgängig in der gleichen Wortbedeutung übersetzt sein. Das heißt z.B., daß [Bezug 1] im Dänischen nicht in der Wortbedeutung [Bezug einer Zeitung] und im Niederländischen im Sinn von [Bezug der Wohnung] übersetzt sein darf. Das Lexikonsystem bietet durch die Strukturierung der Datenbank und ihre Verknüpfung die Möglichkeit zur Konsistenzüberprüfung und -absicherung. Jeder Worteintrag einer Sprache wird nur einmal in der Datenbank des Lexikonsystems geführt. Übersetzungsäquivalente werden durch eine 'Datenbankrelation' miteinander verknüpft. Insbesondere werden importierte lexikalische Daten zunächst auf Inkonsistenzen mit den bereits vorhandenen Lexikoneinträgen überprüft.

Die Erweiterung des Lexikonsystems in Richtung auf einen lexikographischen Arbeitsplatz ist vor allem durch die Verbindung mit einem Konkordanzsystem erreicht worden. Das Konkordanzsystem erlaubt das direkte Aufsuchen von Textstellen mit bestimmten Charakteristika (Suchbedingungen). So kann z.B. in einem Text, wie dem ESPRIT-Korpus, ganz gezielt nach dem Vorkommen bestimmter Wortformen und Wortverbindungen gesucht werden oder auch nur ein schneller Überblick darüber gewonnen werden, welche Lesarten eines Wortes im Text auftreten. Verwendet wird das Konkordanzsystem BYU, bei dem es sich um ein Retrieval System mit umfangreichen Funktionen zur Textanalyse handelt. Da das Konkordanzsystem seinerseits auch den Zugriff auf ein angeschlossenes Lexikon erlaubt, kann bei der Analyse einer Textstelle ein direkter Vergleich mit den betroffenen Lexikoneinträgen vorgenommen werden.

Für den computerunterstützen Lexikonaufbau stellt das BYU-System schon bei der Materialerstellung eine große Hilfe dar. Die Lexikonerstellung ist rechnergestützt in den folgenden Schritten erfolgt.

1. Mit dem BYU-System wird ein Textindex aufgebaut, der sämtliche Wortformen eines Textes mit Angabe der Fundstellen, der absoluten Häufigkeit und einem Kontextausschnitt enthält. Eine solche Wortliste, generiert aus dem dänischen, niederländischen und griechischen ESPRIT-Text, ist die ideale Ausgangsbasis für die weitere lexikographische Bearbeitung. Der Lexikograph kann sich so auf die mitgegebene Kontextangabe stützen oder falls notwendig die gesamte Textpassage heraussuchen.
2. In einer ersten manuellen Bearbeitungsphase wird die Wortliste aller 'types' lemmatisiert und so die Grundlage für das zum Text gehörende Grundformenlexikon geschaffen. Für das Deutsche ist inzwischen das System zur automatischen Lemmatisierung von G. Willée auch auf Personal Computer lauffähig.
3. In der weiteren Bearbeitung werden Informationseinheiten wie Wortartenspezifizierung und die deutsche Übersetzung ergänzt. Durch diese Vorgehensweise kann der Lexikograph zumindest zu einem Teil von der Erfassungsarbeit befreit werden.

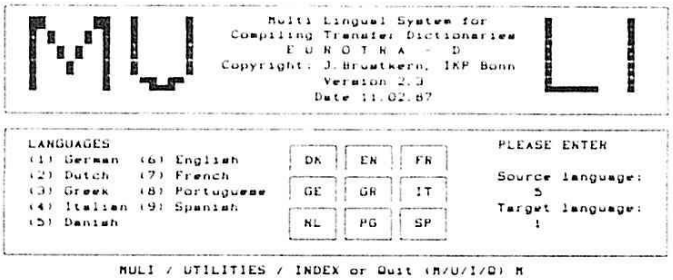

Abb. 1: Auswahl eines Sprachpaars im Lexikonsystem MULI

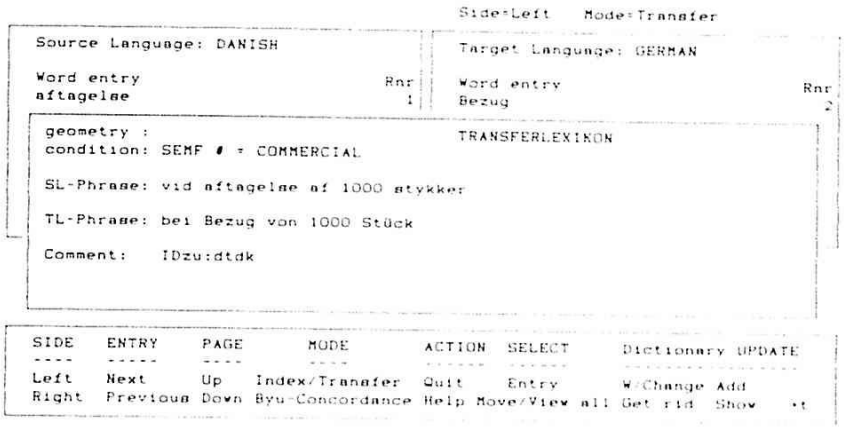

Abb. 2: Benutzeroberfläche zur Lexikonbearbeitung

4. Die so gewonnenen lexikalischen Grunddaten werden in das Lexikonsystem MULI übernommen. Für Bearbeitung, Zugriff und Vergleich der lexikalischen Daten stehen damit zahlreiche Funktionen zur Verfügung (Abb. 2).

In einer späteren Phase sollen die Transferlexika inhaltlich ergänzt werden. Dies betrifft vor allem die Aufnahme weiterer Lesarten und semantischer Spezifikationen. Die lexikalische Arbeit wird außerdem auf das Italienische ausgedehnt. Die im Projekt aufgebaute Softwareumgebung für den computergestützten lexikographischen Arbeitsplatz wird weiter entwickelt. Für diese Arbeiten ist eine Verlängerung des Projekts bis zum 30.6.1988 beantragt.

Jan Brustkern, EUROTRA-D, Institut für Kommunikationsforschung und Phonetik, Bonn, Universität Bonn