

# Modellierung multilingualer Ressourcen

Gerhard HEYER Christian WOLFF

## 1. Zusammenfassung

Dieser Beitrag befaßt sich mit Problemen der Modellierung multilingualer Ressourcen, insbesondere multilingualer Lexika und Corpora. Er geht dabei von den im EU-Projekt MULTILEX entwickelten Entwurfsprinzipien (Lexikonbeschreibungssprache MLex<sub>D</sub>) aus und führt sie unter Berücksichtigung neuerer Kodierungsstandards weiter.

## 2. Wiederverwendbare linguistische Ressourcen

Der Begriff wiederverwendbarer Ressourcen weist ursprünglich auf das Lexikon als eine wesentliche, wenn nicht gar *die* wesentliche Schwachstelle einer verbreiteteren und kostengünstigen Nutzung natürlichsprachlicher Systeme hin.<sup>1</sup> In dem Maße, in dem die Portierung eines natürlichsprachlichen Systems auf eine andere Applikation bzw. eine andere Sprache oder eine andere Anwendungsdomäne von Ergänzungen oder sogar Umkodierungen des Lexikons abhängig ist, steigen auch die Kosten der Portierung. Da die Entwicklung von Sprachtechnologieprodukten wie elektronischen Lexika aber nur finanzierbar ist, wenn genügend viele Sprachen, Anwendungsdomänen oder Programme abgedeckt werden, nimmt die Entwicklung von Werkzeugen zur Lexikonerstellung und -portierung bzw. die Erstellung von wiederverwendbaren Lexika in der Sprachprodukttechnologie der letzten Jahre eine zentrale Stellung ein.<sup>2</sup> Unter Angabe der wesentlichen Kriterien zur begrifflichen Differenzierung lassen sich die folgenden drei Modelle wiederverwendbarer lexikalischer Ressourcen unterscheiden:<sup>3</sup>

---

<sup>1</sup> Vgl. HEYER (1993), 211ff.

<sup>2</sup> Vgl. HEYER (1995), 16ff.

<sup>3</sup> Vgl. HEYER & WALDHÖR (1995), 40ff.

### 1. Das *Datenpool*-Modell

- Wiederverwendbarkeit durch direkten Zugriff
- eine (ggf. sehr große) lexikalische Datenbank
- Erzeugung applikationsspezifischer Lexika durch Filter

### 2. Das *Datenaustausch*-Modell

- Wiederverwendbarkeit durch Kopieren
- Vielzahl applikationsspezifischer Lexika

### 3. *Datenkompilierungs*-Modell

- Wiederverwendbarkeit durch Kompilieren applikationsspezifischer Lexika
- Verwendung einer polytheoretischen und multifunktionalen lexikalischen Datenbank
- Erzeugung applikationsspezifischer Lexika mit Hilfe von Konvertern und Compilern

In der Praxis hat sich dabei besonders das *Datenkompilierungs-Modell* bewährt, das den Gedanken einer zentralen Informationsdatenbank mit dem Gedanken verbindet, applikationsspezifische Lexika mit Hilfe von Konvertern und Compilern zu erzeugen. Es stützt sich dabei nicht auf die Abbildung einer einzelnen linguistischen oder lexikologischen Theorie, sondern versucht, die Qualität lexikalischer Ressourcen auch durch Anwendung unterschiedlicher Optimierungsstrategien zu optimieren.

Unter Nutzung verschiedener Werkzeuge zur Konvertierung bestehender Quellen (u.a. Bücher, linguistisch aufbereiteter Textkorpora, lexikalischer Datenbanken und nicht zuletzt linguistischer Intuition) in eine vereinheitlichende multifunktionale und polytheoretische zentrale lexikalische (und gegebenenfalls andere linguistische Informationen umfassende) Datenbank<sup>4</sup> sowie die nachfolgende Erzeugung einzelner Anwendungen

---

<sup>4</sup> Unter *polytheoretisch* ist dabei die Annahme zu verstehen, daß eine solche multilinguale linguistische Ressource nicht nach einem einzelnen linguistischen bzw. lexikographischen Theoriemodell strukturiert ist, was neben praktischen Überlegungen (bei multilingualen Quellen erhöht sich die Vielzahl unterschiedlicher Strukturbeschreibungen) auch der Überzeugung Rechnung trägt, struktureller Heterogenität den Vorzug vor Informationsverlust bei Ausschluß bestimmter Formate zu geben.

durch Kompilieren der entsprechenden Anwendungsformate aus diesem zentralen Datenbestand heraus lassen sich lexikalische Ressourcen in beachtlichem Maße zumindest für *einen* Hersteller oder Nutzer linguistischer Programme wiederverwenden.

### 3. Standardisierungsprobleme

Um das Datenkompilierungs-Modell jedoch nicht nur für *einen* Verlag oder Hersteller von Sprachprodukten erfolgreich einzusetzen, bedarf jede Repräsentationsebene einer weitergehenden Standardisierung. U.a. handelt es sich dabei um eine Standardisierung

1. der **Repräsentation** linguistischer Lexikoninhalte (*konzeptuelles Schema*)
2. der **Austauschformate** von Lexika (*externes Schema*)
3. der **Datenformate** von elektronischen Wörterbüchern (*internes Schema*)
4. der **Benutzerschnittstellen** von elektronischen Wörterbüchern.

In diesem Beitrag spielen nur die ersten drei Standardisierungen eine Rolle. Im Rahmen des von der EU geförderten Projekts MULTILEX, das unter Beteiligung von Hochschulen und Industriepartnern die Erarbeitung einer Lexikonbeschreibungssprache für die westeuropäischen Sprachen zum Ziel hatte, wurde hinsichtlich der Austauschformate, also dem *externen Schema* lexikalischer Datenbanken, der Industriestandard SGML empfohlen.<sup>5</sup>

### 4. Die Lexikonbeschreibungssprache MLex<sub>D</sub>

Die Lexikonbeschreibungssprache MLex<sub>D</sub> als Standardisierung des konzeptuellen Schemas multilingualer lexikalischer Ressourcen ist eine von bestimmten linguistischen Theorien oder Anwendungsformaten unabhängige Repräsentation linguistischer Sachverhalte, wie sie sich typischerweise für die linguistischen Ebenen Phonetik, Orthographie, Morphologie, Syntax, Semantik und Pragmatik formulieren lassen. Mit MLex<sub>D</sub> können monolinguale oder bilinguale Lexikoneinträge medienneutral und

---

<sup>5</sup> Vgl. AHMAD (1994).

anwendungsunabhängig, insbesondere hinsichtlich der Frage, ob das Lexikon von einem Menschen oder einer Maschine genutzt wird, erstellt werden. Unter einem *Lexikon* wird dabei eine Menge strukturierter und hierarchisch organisierter lexikalischer Einträge gemäß folgender Definition verstanden:

1. Ein *monolingualer* lexikalischer Eintrag ist ein Tupel  $\langle \text{Adresse}, \text{Angabe} \rangle$ , bestehend aus einer *Lexikonadresse* (Lemma-Namen) und einer Menge *lexikalischer Angaben*.
2. Die *Adresse* oder der Lemma-Name ist eine eindeutige Bezeichnung für die graphemisch-phonetisch-morphologischen Vorkommen eines Wortes (Wortformen).
3. Eine lexikalische *Angabe* ist eine Menge von (typisierten) Attribut-Wert-Paaren.
4. Ein *bilingualer* lexikalischer Eintrag ist ein Quadrupel  $\langle \text{Adresse}, \text{Angabe}, \text{Zielsprachenname}, \text{Zielsprachenangabe} \rangle$ , wobei ein monolingualer Eintrag um den Namen einer Zielsprache und die entsprechende zielsprachliche Angabe ergänzt ist.
5. Eine *lexikalische Regel* ist ein Tripel  $\langle \text{Regelname}, \text{Eingangsadresse}, \text{Ausgangsadresse} \rangle$ .

Als Grundlage von  $\text{MLex}_D$  dient eine Unterscheidung zwischen der Gestalt eines Ausdrucks, der GPMU (graphemic-phonological-morphological unit), und seiner Bedeutung in einer Sprache, der LU (lexical unit). Lexikalische Beschreibungen resultieren aus der Verknüpfung von GPMUs mit LUs, wobei nicht nur eine GPMU mit einer LU, sondern auch mehrere LUs mit ein und derselben GPMU (Homonymie) bzw. eine LU auch mit mehreren GPMUs (graphemische Varianten, Abkürzungen) verknüpft sein können.

## 5. Weiterentwicklung von $\text{MLex}_D$ für die Kodierung multilingualer Corpora

In Weiterführung des  $\text{MLex}_D$ -Formats erarbeiten wir eine XML-basierte Spezifikation für ein multilinguales Lexikon.<sup>6</sup> Im Unterschied zu

---

<sup>6</sup> Vgl. QUASTHOFF & WOLFF (1999); die dort erarbeitete Spezifikation definiert ausge-

dem ursprünglichen MLex<sub>D</sub>-Ansatz, der auf der Verwendung einer standardisierten SGML-DTD (*document type definition*) beruht, erfolgt hier die Kodierung mit Hilfe von XML (Extensible Markup Language) bzw. mit XML-basierten Spezialstandards wie RDF (Resource Description Framework).<sup>7</sup> Dabei wird von folgenden Annahmen ausgegangen:

1. Die kleinsten Einheiten des linguistischen Modells sind die Attribut-Wert-Paare, die sich durch Elemente einer XML-DTD darstellen lassen, zusammengehörige Merkmale können durch Einbettung von Elementen hierarchisch geordnet werden. Die Operatoren der SGML-Syntax sind ausreichend mächtig, um die benötigten Kardinalitäten von Merkmalen ausdrücken zu können.
2. Die gegenüber SGML vereinfachte Syntax von XML erleichtert die Lexikon-Analyse. Gleichzeitig unterstützt der modulare Aufbau von XML die Zusammenführung heterogener Ausgangsmaterialien: Durch Einführung verschiedener Namensräume (XML name spaces) besteht keine unmittelbare Notwendigkeit, alle Daten *sofort* auf eine einheitliche Master-DTD abzubilden.<sup>8</sup>
3. Umgekehrt sind mit der Anwendung deklarativen Markups unterschiedlich komplexe Sichten auf den Datenbestand in Abhängigkeit von der jeweiligen Anwendung möglich: Für die Verarbeitung von Massendaten kann es notwendig sein, die Merkmalsmenge des Lexikons für eine effiziente Verarbeitung auf ein Minimum zu beschränken, während für Anwendungen aus dem Bereich der Wissensverarbeitung eine Beschreibung mit maximalem Detaillierungsgrad sinnvoll sein kann.
4. Die getrennte Kodierung von Daten (XML) und Metadaten (resource description framework) erleichtert die anwendungsorientier-

---

hend von MLex<sub>D</sub> ein Kodierungsformat für ein multilinguales Lexikon.

<sup>7</sup> Vgl. LASSILA & SWICK (1999); BRICKLEY & GUHA (1999).

<sup>8</sup> XML Namespaces sind ein geeignetes Werkzeug, um unterschiedliche Kodierungsformate innerhalb eines XML-Dokuments bzw. eines Lexikoneintrags zulassen zu können. Sie stellen eine alternative Modularisierungsmöglichkeit zur Verwendung von SGML-Subdokumenten dar, die jeweils unterschiedliche DTDs verwenden können, vgl. MALER & EL ANDALOUSSI (1996), 217f.

te Wiederverwendung von Ressourcen. Dies betrifft sowohl den “klassischen” Bereich der bibliographischen Beschreibung von Information durch das Dublin Core-Tagset<sup>9</sup> als auch anwendungsbezogene Metainformation.

Wir gehen bei der Spezifikation von einer zweischichtigen Architektur aus: Ein Basisinventar an Beschreibungsmerkmalen bildet als *kanonisches Modell* die Grundlage und Zielstellung für die Kodierung multilingualer Ressourcen. Es führt die Struktur von MLex<sub>D</sub> weiter und ergänzt sie um

1. eine zentrale Eintragsstruktur für multilinguale Einträge bzw. die Relationierung der einzelsprachlichen Einträge;
2. Metainformation zu bibliographischen und technischen Angaben und
3. eine differenzierte Struktur für semantische Information (Zuordnung von Klassifikationen und Ontologien, terminologische Einordnung).<sup>10</sup>

Die multilinguale Eintragsstruktur besteht aus einer Liste von Referenzen auf einzelsprachliche Einträge sowie aus Zusatzinformation, die sich auf den Eintrag als Gesamtheit beziehen (Kommentare; Hinweise auf Direktionalität z.B. für die Übersetzung). In Weiterführung der konzeptuellen Trennung von GPMU und LU in MLex<sub>D</sub> (vgl. oben Kap. 4) repräsentiert eine solche Struktur jeweils nur *eine* Bedeutungsvariante; Polysemien werden durch unterschiedliche LUs ausgedrückt, die aber jeweils auf identische syntaktische und morphologische Beschreibungen verweisen können. Dabei sind die einzelsprachlichen Beschreibungsmerkmale auf die in Tabelle 1 aufgeführten Kategorien aufgeteilt.

Auf einer zweiten Ebene (*ressourcenbezogenes Modell*) sind externe Kodierungsschemata aus unterschiedlichen Quellen in die Lexikonbeschreibung integriert, d. h. sie werden ohne deklaratives Markup aus den

---

<sup>9</sup> Vgl. Dublin Core Metadata Initiative 1997; das Dublin Core Tagset umfaßt 15 Felder (wie etwa Subject, Title, Author, Publisher etc.).

<sup>10</sup> Dazu wurden unterschiedliche terminologische Standards wie TIF (terminology interchange format, vgl. TIF 1994) ausgewertet und für den Aufbau semantischer und pragmatischer Beschreibungskategorien verwendet, vgl. auch SCHMITZ (1999), 111ff.

Quellen übernommen und den entsprechenden Grobkategorien des kanonischen Modells zugeordnet.<sup>11</sup> Hier wird (zunächst) auf ein hypothetisches “universelles” Beschreibungsmodell verzichtet. Das zwei-Ebenen-Modell geht von einer evolutionären Weiterentwicklung von Sprachdatencorpora aus, bei der unter Einsatz unterschiedlicher Heuristiken eine Migration der Datenbestände von der zweiten auf die erste Ebene stattfindet. Darunter ist vor allem das Abschöpfen heterogener Informationsquellen zu verstehen, u.a. durch

- die Generalisierung gesicherten sprachlichen Wissens aus hochwertigen Ressourcen und
- die Einbeziehung von statistischen und musterorientierten Verfahren für die Ableitung linguistischer Merkmalsbeschreibungen.

<b>Kategorie</b>	<b>Inhalt</b>
Haupteintrag	führt die verschiedenen Beschreibungsbereiche zusammen und enthält allgemeine Information (u.a. standardisierte Schreibweise, Eintragstyp [z.B. Term, Phrase, Kompositum], Sprache, Kategorie)
syntaktische Beschreibung	syntaktische Beschreibung, untergliedert nach Wortarten
morphologische Beschreibung	Grundform, morphologische Zerlegung, Angaben zur Derivation, Liste flektierter Formen
orthographische Beschreibung	Varianten, Typ (z.B. Akronym)
semantische und pragmatische Beschreibung	Definition(en), Einordnung in Ontologien und Klassifikationsschemata, Hinweise zur Verwendung
Homographen	Liste und Typisierung von Homographen
Metadaten	Bibliographische Angaben (Quelle, Autor) sowie Verwaltungsangaben

Tabelle 1: Übersicht zu den Beschreibungskategorien

---

<sup>11</sup> Technisch ist dies durch entsprechende Definition von mixed content models in den XML-DTDs möglich, d.h. auf der Makroebene (z.B. Angaben zur Pragmatik) können die Inhalte entweder mit detailliertem deklarativem Markup enthalten sein, oder sie sind ohne weitere Analyse im Ausgangsformat übernommen.

Die Anwendung des MLex<sub>D</sub>-basierten Formats und der oben skizzierten Entwicklungsmethode erfolgt im Kontext der Erweiterung des Leipziger Corpus- und Lexikonprojekts (vgl. QUASTHOFF 1998, LÄUTER & QUASTHOFF 1999 und <http://www.wortschatz.uni-leipzig.de>) zu einem multilingualen Corpus. Dort werden derzeit zunächst parallel umfangreiche einzelsprachliche Corpora aufgebaut, die durch Integration multilingualer Lexika und ihre Aufbereitung im oben beschriebenen Format zusammengeführt werden sollen. Entscheidend ist dabei einerseits die Offenheit für unterschiedlich strukturierte Information, andererseits die kontinuierliche Optimierung der vorhandenen Ressourcen durch "linguistische Qualitätssicherung" (z.B. der Abgleich neuer Lexikoneinträge mit abgesichert hochqualitativen Substanzen). Das hier skizzierte zweistufige Kodierungsformat soll bei diesem bottom-up-Ansatz für den Aufbau eines multilingualen Korpus sowohl den Austausch erleichtern als auch durch die Verwendung von deklarativem Markup die Weiter- und Wiederverwendung der multilingualen Daten für unterschiedliche Anwendungen der Sprachprodukttechnologie vereinfachen.

## Literatur

- AHMAD, K. (ed., 1994): Multilex: Final Report. Guildford: University of Surrey.
- BOGURAEV, Bryan and BRISCOE, Ted (eds., 1989): Computational Lexicography for Natural Language Processing. London / New York: Longman.
- BRAY et al. (1998): Tim B., Dave HOLLANDER and Andrew LAYMAN, Namespaces in XML. World Wide Web Consortium Recommendation, Januar 1999. <http://www.w3.org/TR/REC-xml-names>.
- BRICKLEY, Dan and GUHA, R.V. (eds., 1999): Resource Description Framework (RDF) Schema Specification. World Wide Web Consortium Proposed Recommendation, März 1999. <http://w3.org/TR/WD-rdf-schema>.
- Dublin Core Metadata Initiative (1997): Dublin Core Metadata Element Set: Reference Description. [http://www.purl.org/dc/about/element\\_set.htm](http://www.purl.org/dc/about/element_set.htm).
- HEYER, Gerhard (1992): On the Role of the Dictionary and Dictionary-Based Approaches in Language Products Technology. In: Frank BECKMANN und Gerhard HEYER, (eds.), Theorie und Praxis des Lexikons, 207-217. Berlin / New York: de Gruyter.

- HEYER, Gerhard (1995): Elements of a Natural Language Processing Technology. In: G. HEYER and H. HAUGENEDER (eds.), *Language Engineering. Essays in Theory and Practice of Applied Natural Language Computing*, 15-32. Wiesbaden: Vieweg.
- HEYER, Gerhard and WALDHÖR, Klemens (1995): General Language Resources: Lexica. In: Marianne KUGLER, Kurshid AHMAD and Gregor THURMAIR (eds.), *Translator's Workbench. Tools and Terminology for Translation and Text Processing (ESPRIT Research Reports Pr. 2315 · TWB · Vol. 1)*, 40-48. Berlin u.a.: Springer.
- HEYER, Gerhard und WOLFF, Christian (eds., 1998): *Linguistik und neue Medien*. Wiesbaden: Deutscher Universitätsverlag.
- LASSILA, Ora and SWICK, Ralph R. (eds., 1999): *Resource Description Framework (RDF) Model and Syntax*. World Wide Web Consortium Recommendation, Februar 1999. <http://w3.org/TR/REC-rdf-syntax>.
- LÄUTER, Martin and QUASTHOFF, Uwe (1999): *Kollokationen und semantisches Clustering* (in diesem Band).
- MALER, Eve and EL ANDALOUSSI, Jeanne (1996): *Developing SGML DTDs. From Text to Model to Markup*. Upper Saddle River, NJ u.a.: Prentice Hall PTR.
- QUASTHOFF, Uwe (1998): *Projekt Der Deutsche Wortschatz*. In: HEYER & WOLFF 1998, 93-99.
- SCHMITZ, Klaus-Dirk (1999): *MARTIF – ein SGML-basiertes Austauschformat für terminologische Daten*. In: Wiebke MÖHR und Ingrid SCHMIDT (eds.), *SGML und XML. Anwendungen und Perspektiven*, 109-121. Berlin u.a.: Springer.
- TIF (1994): *Computational aids in terminology – Terminology Interchange Format (TIF) – An SGML Application*. ISO/DIS 12200. ISO, 1994-11-03.
- WOLFF, Christian and QUASTHOFF, Uwe (1999): *LEMDIC. Leipzig Multilingual Dictionary Design*. Draft Technical Report, Leipzig University, CS Inst., NLP Dept., July 1999, Version 0.7.