Die integrierte Repräsentation linguistischer Daten

Andreas MENGEL

1. Einleitung

In den letzten Jahren ist die Anzahl elektronisch verfügbarer Datenbestände für die Sprachverarbeitung stark gestiegen. Diese Tatsache ist in doppelter Hinsicht eine Herausforderung für die Sprachverarbeitung. Erstens bietet die größere Menge an Daten mehr Möglichkeiten, hypothesengeleitete Fragestellungen mit Hilfe großer Datenmengen zu bearbeiten. Die zweite Herausforderung besteht in der Tatsache, daß die prinzipielle Verfügbarkeit elektronisch gespeicherter Daten, die nicht automatisch einen problemlosen Austausch garantieren kann, ihr Pendant in der tatsächlichen Verwendbarkeit durch entsprechende Maßnahmen auf der Kodierungsebene finden muß. Solche Maßnahmen sind Thema dieses Beitrags.

2. Problem

Bei elektronisch verfügbaren linguistischen Ressourcen kann es sich um digitalisierte Sprachaufnahmen, deren Transkripte, geschriebene Texte, linguistische Annotationen, Lexika etc. handeln. In vielen Fällen ist es wünschenswert, existierende linguistische Daten zu übernehmen, anstatt sie selbst zu erstellen, da die Akquisition und Annotation von Sprachdaten sehr aufwendig ist. Für das Ablegen dieser Art von Daten existiert z.Zt. kein Standard. Zunächst finden sich unterschiedliche Formate für die Speicherung von Sprachsignaldaten. Ebenso gibt es eine Vielzahl verwendeter Repräsentationsformalismen für Annotations- oder Lexikondaten (für eine Übersicht s. KLEIN et al. 1998). Dieser Mangel an Standardisierung hat zur Folge, daß der Austausch der Daten erschwert wird: Für die Übernahme von Daten, denen ein anderer Standard zugrunde liegt, muß die zur Verarbeitung nötige Software angepaßt oder neu entwickelt werden. Prinzipiell ist dabei der für die Anpassung der Datenformate an die eigenen Bedürfnisse nötige Aufwand zwar geringer als eine eigene Erhe-

bung und Annotation der Daten selbst. Die fehlende Standardisierung erfordert aber nicht nur einen erhöhten Aufarbeitungsaufwand, sondern birgt darüber hinaus die Gefahr, daß das spezielle Format der Kodierung falsch interpretiert wird. Nicht zuletzt muß auch der eigene Standard, in den zu konvertieren ist, ebenfalls entwickelt werden. Höherer Aufwand entsteht also durch Entwicklungskosten, da an verschiedenen Stellen für die gleichen Korpora Konvertierungssoftware, Repräsentationsstandards und Zugriffssoftware entwickelt werden müssen (s. Abbildung 1). Zusätzlich besteht die Gefahr der Fehlinterpretation der Kodierung.

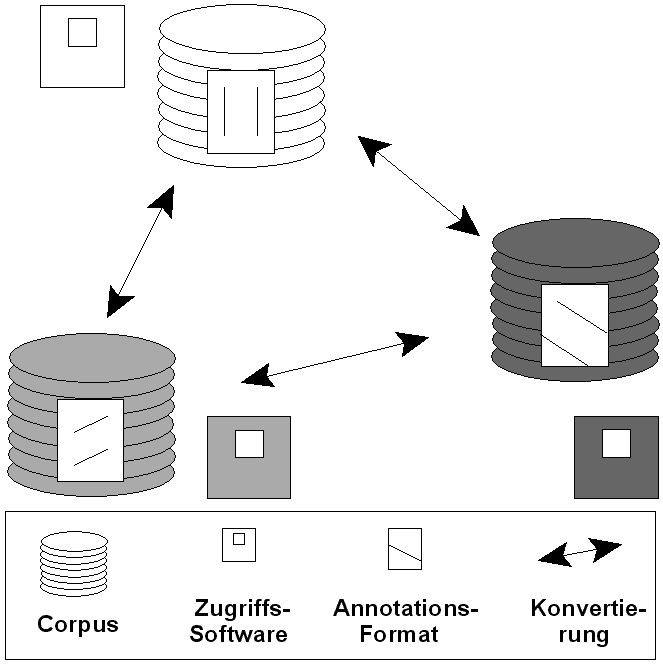

Abb. 1: Zusätzlicher Aufwand durch fehlende Standards

Eine Standardisierung der Kodierung von Korpusdaten kann aber nicht nur helfen, den Akquisitionsaufwand zu reduzieren, sondern erhöht auf lange Sicht die Anzahl von Standardkorpora, die für die Messung und den Vergleich der Qualität und Performanz linguistischer Theorien und Software zur Verfügung stehen.

3. Standardisierungbemühungen

Das EU-Projekt MATE (`http://mate.nis.sdu.dk`) hat die Entwicklung von Standards für die Repräsentation und Verarbeitung linguistischer Beschreibung von Sprachdaten zum Inhalt. Dafür werden Vorschläge zur theorie- und sprachenunabhängigen Kodierung dieser Information und zur Entwicklung einer Softwareumgebung, die die Verarbeitung entsprechender Daten unterstützt, gemacht. Die Zielsetzung dieser Standardisierungsvorschläge ist dabei nicht nur die Verbesserung des Austausches linguistischer Daten, sondern die Vereinheitlichung und Integration linguistischer Daten überhaupt. Neben der Beschreibung von Einheiten (Lauten, Morphen, Wörtern etc.) ist dabei eine Fülle anderer Daten zu berücksichtigen, die im folgenden genannt werden.

Sprachverhaltensdaten: Sprachverhaltensdaten sind zunächst die eigentlichen Primärdaten, die meist visuelles (Bilder, Filme) oder akustisches (Klangaufnahmen) meßbares Verhalten speichern.

Kategorisierungsdaten: Kategorisierungsdaten sind diejenigen Informationen, die über Segmente (Laute, Wörter) und ihre physikalischen Eigenschaften (Spektrum, Dauer) oder über deren Kategorisierung (Lautklasse, Wortart) Auskunft geben. Diese Daten bilden den Kernbereich dessen, was unter *Korpusdaten* verstanden wird.

Wissensbasen: Wissensbasen bilden Informationssammlungen über Eigenschaften von Korpustoken, die als konstant über einen zeitlichen oder räumlichen Bereich oder eine Sammlung von Korpora hinweg angesehen werden (Lexika, Thesauri, Situationsbeschreibungen).

Annotationsdaten: Annotationsdaten enthalten solche Informationen, die über die für die Kodierung zugrunde gelegten Theorien und Annotationsanweisungen, Anmerkungen zur Kodierung, Kodierungsalternativen, und Kodierungsverlaufsprotokolle Auskunft geben.

Angesichts der Unterschiedlichkeit und Anzahl dieser Daten gilt es, möglichst einheitliche Verfahren zu ihrer Repräsentation zu entwickeln. Wünschenswerte Objekte der Standardisierungsbemühungen sind dabei mindestens die im folgenden beschriebenen Aspekte.

3.1 Kodierung einzelner Phänomene

Hauptinteresse der linguistischen Beschreibung sind Einheiten, die auf verschiedenen Beschreibungsebenen klassifiziert werden, z.B. Laute, Morphe, Wörter, Phrasen. Für einen einheitlichen Zugriff auf Beschreibungsdaten ist ein Standard nötig, der unabhängig von dem jeweiligen Beschreibungsgegenstand und der verwendeten Theorie ist. Die diesem Standard zugrundeliegenden Prinzipien müssen demnach auch transparent sein, damit jede neue Beschreibung diesem Standard gemäß repräsentiert werden kann. Im Projekt MATE wird hierfür XML (*extensible markup language*) verwendet. XML-Dokumente können eine DTD (*document type definition*) enthalten, die die Struktur und die Attribute der Elemente eines Dokumentes beschreiben. Dadurch ist nur ein minimaler Anteil der Beschreibung vorbestimmt, nämlich daß es Einheiten (*elements*) gibt, die in Beziehung zu anderen Einheiten stehen und daß diese Einheiten Eigenschaftsdimensionen (*attributes*) besitzen können, die bezüglich ihrer Ausprägungen (*values*) näher spezifiziert werden. Ein einfaches Beispiel wäre die Kodierung des Wortes *Haus*, das den Wortartwert (*pos: part of speech*) *NN* hat:

```
<word pos="NN">Haus<word>
```

Durch diese Bedingungen und die Offenheit von XML kann die Beschreibung von Sprachdaten theorie- und phänomenunabhängig mit einem einheitlichen Formalismus kodiert werden.

Der Umstand der weiten Verbreitung von XML und der damit einhergehenden guten Verfügbarkeit von verarbeitender Software wirkt sich darüber hinaus begünstigend aus.

3.2 Verteilung der Informationen auf Dateien

Der Vorgang der Annotation von Sprachdaten bezieht sich im allgemeinen auf unterschiedliche Sprachebenen und besteht u.U. in einer gewissen hierarchischen Abhängigkeit dieser Ebenen untereinander (z.B. Laute, Wörter, Sätze). Daten werden für bestimmte Zwecke erhoben, und die bearbeiteten Annotationsebenen sind am aktuellen Forschungsinteresse

innerhalb der Institution, die diese Annotation erstellt, ausgerichtet. So werden unterschiedliche Beschreibungsebenen von verschiedenen Annotatoren bearbeitet. Weiterhin zerfällt der Prozeß der Annotation innerhalb einer Beschreibungsebene in mehrere Analyseschritte; so steht vor der Silbenbeschreibung die Segmentierung auf Lautebene, und die syntaktische Beschreibung setzt idealerweise eine Wortsegmentierung voraus. Dadurch bietet sich die Standardisierung einer organisierten Verteilung dieser verschiedenen Informationen auf Dateien und die Möglichkeit der Bezugnahme der unterschiedlichen Beschreibungseinheiten untereinander an.

3.3 Struktur der Bezugnahme auf andere Einheiten

Die Verteilung der Kodierung der Annotation der Einheiten unterschiedlicher Beschreibungsebenen und ihrer gegenseitigen Bezüge hat weitere Implikationen: Weil diese Bezüge sich auf für die so verbundenen Einheiten konstitutive Eigenschaften gründen — so sind Anfangs- und Endzeitpunkt eines Wortes auch Anfangs- und Endzeitpunkte des ersten bzw. letzten Lautes des Wortes — liegt es nahe, die Kodierung der Eigenschaften (in XML: *values*) ebenfalls zum Objekt der Standardisierung zu machen. Die Verteilung, Berechnung und Vererbung von Eigenschaften garantiert so eine höhere Konsistenz und einfachere Pflege der Daten, etwa wenn Zeitinformation nur an einer Stelle geändert werden muß.

3.4 Beschreibung der zugrundeliegenden Theorie

Die einheitliche und theorieunabhängige Kodierbarkeit ist ein wichtiges Ziel für die standardisierte Repräsentation von Informationen und die hierdurch erleichterte Verarbeitung wegen einer geringeren Anzahl dafür benötigter Softwarekomponenten. Diese syntaktische Standardisierung garantiert aber nur eine bessere Verarbeitbarkeit auf der Ebene der Software. Die Bedeutung dessen, was kodiert wurde, muß denjenigen, die die Annotationen verwenden wollen, aber auch vermittelbar sein. Deshalb sind hier zwei weitere Arten von Informationen sowie ihre Verbindung mit den Annotationen von Bedeutung: Erstens muß der Prozeß der Segmentierung und Klassifizierung, also die Bestimmung der Einheiten und ihrer Eigenschaften beschrieben werden, zweitens ist die genaue – textuelle – Be-

schreibung und Definition der verwendeten Elementnamen, Attribut- und Attributwertebezeichnungen nötig, da nur über diese die genaue Bedeutung und Qualität der Kodierung ermittelt werden kann.

3.5 Beschreibung der Annotation

Weitere Informationen, die die Annotation selbst betreffen, sind alternative Kodierungsmöglichkeiten, die Einzelfälle betreffen, und eine Beschreibung des tatsächlichen Annotationsverlaufes. Bei diesen Fällen handelt es sich um Meta-Annotationen, die ebenfalls für die Interpretation und die Verarbeitung der Daten wichtig sein können.

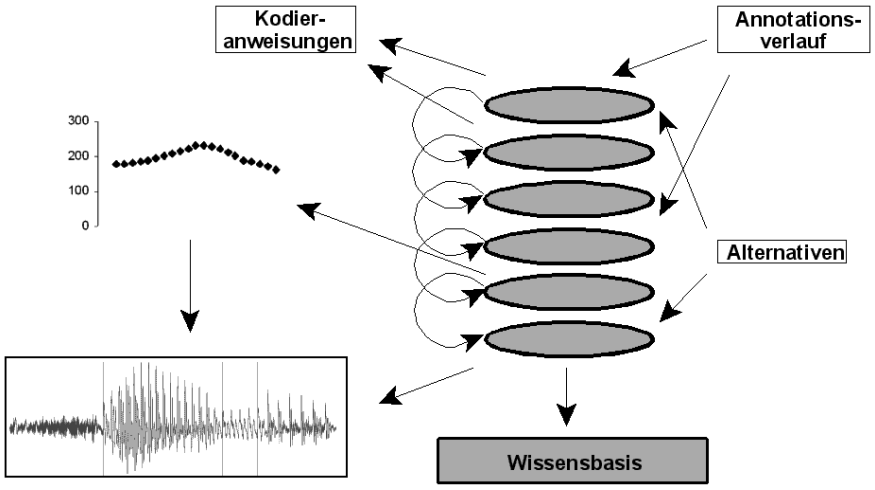

Abb. 2: Unterschiedliche Annotationsinformationen und ihre Beziehungen untereinander

Abbildung 2 gibt einen Überblick über die beschriebenen Arten von Informationen und deren Bezüge untereinander, die hier durch Pfeile gekennzeichnet sind. Die Pfeilrichtung gibt dabei Hinweise darauf, von welchen Dokumenten und Einheiten (XML: *elements*) auf welche anderen (durch XML-*href-Attribute*) verwiesen werden kann.

4. Zusammenfassung

Dieser Beitrag hat verschiedene Informationsarten von Korpora beschrieben und für die Notwendigkeit ihrer Berücksichtigung und gegenseitige Bezugnahme argumentiert. Neben einer Vereinheitlichung der Kodierung von Sprachdatenkorpora auf der Formatebene stellt dieser Aspekt einen weiteren wichtigen Schritt zur generellen Verbesserung der allgemeinen Verfügbarkeit und des Austausches von Ressourcen dar. Weitere und konkrete Vorschläge hierzu finden sich bei MENGEL et al. (1999).

Literatur

KLEIN et al. (1998): M. K., N.O. BERNSEN, S. DAVIES, L. DYBKJÆR, J. GARRIDO, H. KASCH, A. MENGEL, V. PIRRELLI, M. POESIO, S. QUAZZA and S. SORIA, Supported Coding Schemes. MATE Deliverable D1.1, July 1998.

MENGEL et al. (1999): A. M., L. DYBKJÆR, J.M. GARRIDO, U. HEID, M. KLEIN, V. PIRRELLI, M. POESIO, S. QUAZZA, A. SCHIFFRIN and C. SORIA, MATE Dialogue Annotation Guidelines (M-DAG). MATE Deliverable D2.1, August 1999.

XML: W3C "*Extensible Markup Language*": `http://wwww.w3.org/XML` .