Thorsten Trippel, Thierry Declerck und Ulrich Heid

Sprachressourcen in der Standardisierung

Wir berichten über internationale Normungsarbeit im Bereich von Sprachressourcen. Die Normen werden von internationalen Arbeitsgruppen im Rahmen der *International Organization for Standardization* (ISO) entwickelt und jeweils national von entsprechenden Gruppen, in Deutschland koordiniert vom Deutschen Institut für Normung (DIN), begleitet und diskutiert. Für die automatische Sprachverarbeitung besteht seit Jahren zunehmend Bedarf an elektronischen Ressourcen: Lexika, Korpora, Grammatiken, Annotationskonventionen, Sprachdatensammlungen, usw. Damit solche Ressourcen über einen einzelnen Anwendungskontext hinaus wiederverwertbar sind und zwischen Arbeitsgruppen ausgetauscht werden können, wird an einer Normung ihrer Repräsentationsformate und der zur Beschreibung von Ressourceninhalten benutzbaren Vokabularien gearbeitet (Datenkategorien). Waren in der Vergangenheit Standardisierungsbemühungen auf bestimmte Ausschnitte aus dem Spektrum der linguistischen Beschreibungen von Ressourcen beschränkt (z.B. die EU-Projekte SAM im Bereich gesprochener Sprache, EAGLES und ISLE im Bereich von Morphosyntax, Syntax, lexikalischer Semantik in Texten und Lexika und Sprachtechnologie), so ist die Zielsetzung der 2002 und 2003 gegründeten ISO (TC 37 SC 4) bzw. DIN (NAT AA 6) Arbeitsgruppen breiter: es geht um Metarichtlinien für die Repräsentation und Annotation von Texten ebenso wie um Datenkategorien für Lexika, morphologische und morphosyntaktische Analyse, usw. Wir beschreiben den aktuellen Stand der Normungsdiskussion.

1 Einführung in die Normierung für Sprachressourcen: Historischer Kontext

Sprachressourcen sind eine Klasse heterogener Informationen, die Gegenstand von Linguistik und Sprachtechnologie sind, aber auch in Anwendungskontexten wie Übersetzung und Lexikonentwicklung gefragt sind. Dazu gehören Textkorpora, Lexika, Daten gesprochener Sprache, aber auch Annotationsrichtlinien und -verfahren. Die verschiedenen Ressourcen unterscheiden sich dabei häufig sowohl durch ihre Form, die verwendeten Datenstrukturen und Anwendungskontexte. Ebenso variieren die Annotationsstrukturen sehr oft von Projekt zu Projekt oder zwischen verschiedenen Anbietern von Ressourcen. In der Praxis läuft dies in der Regel auf idiosynkratische Bestimmungen von Datenformaten und Verarbeitungsverfahren hinaus; Konsistenzprüfungen werden, wenn überhaupt, ad hoc von Experten oder von applikationsspezifischen Parsern durchgeführt. Diese idiosynkratischen Strukturen erlauben es aber nicht, Daten zwischen Applikationen oder Nutzern auszutauschen, ohne vorher eine detaillierte Analyse und Transformation

durchzuführen, selbst wenn Teilstrukturen und Informationsgehalt direkt vergleichbar wären.

Bei vielen Verfahren, insbesondere für statistische Ansätze, besteht ein erheblicher und stetig größer werdender Bedarf an hochwertigen Ressourcen. Daher wird eine Standardisierung von Formaten und Formalismen über Grenzen der linguistischen Theoriebildung hinweg angestrebt, um zumindest existierende Werkzeuge und Verfahren verwenden und austauschen zu können, aber auch um Ressourcen selbst als Grundlage auch anderer als der ursprünglich intendierten Verwendungen benutzen zu können.

Der vorliegende Beitrag beruht auf den Normentwürfen und Vorschlägen aus dem Standardisierungsgremium der International Organization for Standardization (ISO) im Rahmen des technischen Komitees 37, Arbeitsausschuss 4 (*Language Resources*) ISO TC 37/SC 4. Das Komitee wurde im Sommer 2002 gegründet und ist international besetzt; dabei entsenden interessierte nationale Normungsinstitutionen Experten, die an Normentwürfen mitarbeiten. In Deutschland, wie in anderen Ländern, gibt es seit Mai 2003 eine nationale Arbeitsgruppe, die deutsche Beiträge zur internationalen Normung koordiniert. Sie wird organisatorisch vom Deutschen Institut für Normung (DIN) im Rahmen seines Normungsausschusses Terminologie betreut. Die Autoren sind Mitglieder dieser Arbeitsgruppe.

Die Arbeitsgruppe ISO TC37/SC4 führt auf internationaler Ebene Versuche der Normierung fort, die in den 1990er Jahren in einzelnen Projekten der Sprachverarbeitung begonnen worden waren. Beispiele solcher früheren Versuche für Vereinheitlichung von Annotationen und Annotationsverfahren sind etwa die EU-Projekte SAM für die Sprachsignalannotation, EAGLES (Expert Advisory Groups on Linguistic Engineering Standards) für morphosyntaktische und syntaktische Annotation von Textkorpora (URL: `http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html` und `http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html`) oder ISLE (International Standards for Language Engineering), z.B. zur Repräsentation von Wörterbucheinträgen in NLP-Systemen (vgl. MILE). Diese Vorhaben waren auf eine Harmonisierung bzw. Standardisierung der Annotationen selbst ausgerichtet; die Harmonisierung sollte durch einen Konsensus-Prozeß erreicht werden: eine Art minimaler, allgemein akzeptierter Basisvorschlag für die jeweilige Annotation wurde erarbeitet.

Demgegenüber ist es das Kennzeichen eines Teils der hier diskutierten Normungsvorhaben, dass die Standardisierung *eine Ebene höher* ansetzt, auf der Schicht von Meta-Annotationen, von Frameworks für die Erstellung und den Austausch von Annotationen, Datenstrukturen und Ressourcen, oder bei Prozeduren für die Erstellung von Inventaren für Datenkategorien. Ein Teil der Normung ist also nicht mehr auf Harmonisierung der Ressourcen durch gemeinsame Formate, sondern auf Interoperabilität durch gemeinsame Meta-Formate, Austauschformate, Herangehensweisen usw. gerichtet. Jeder, der Ressourcen produziert, soll seine Daten in ein solches Format abbilden können; jeder der fremde Ressourcen nutzt, soll die Gewähr haben, eine interpretierbare *Übersetzung* oder Transformation leisten zu können.

2 Einsatzbereite Standards und Standardentwürfe

Im Rahmen der ISO Standardisierung gibt es verschiedene Phasen, bevor eine Norm als verbindlich anzusehen ist, nämlich *Work Item*, *Committee Draft* (CD), *Draft International Standard* (DIS), *Final Draft International Standard* (FDIS). Das Work Item ist dabei nur eine Beschreibung eines Normierungsvorhabens, CD der erste in den Normenausschüssen zur Diskussion stehende Entwurf einer Norm, DIS ist eine entsprechend fortgeschrittene Version, die auch interessierten Kreisen außerhalb der Standardisierungsgremien zugänglich gemacht werden kann, und FDIS ist eine fast endgültige Version, die sich schon zur Implementierung in Testumgebungen eignet.

Derzeit sind im Bereich der Sprachressourcen einige Standards in der Entwicklung; diese betreffen die folgenden Fragestellungen:

allgemeine (linguistische) Annotation: Grundlagen für die Kodierung linguistischer Informationen

Wörterbuchbeschreibungen: Beschreibung und Austausch von Wörterbucheinträgen und ganzen Wörterbüchern

Wortsegmentierung: sprachübergreifende Kriterien zur Beschreibung von Wortgrenzen

morphosyntaktische Annotation: einheitliche Annotation von morphosyntaktischer Information

syntaktischen Annotation: einheitliche Annotation von syntaktischer Information, ein neues Work Item

Merkmalsstrukturen: Kodierung von Merkmalsstrukturen verschiedener linguistischer Theorien

Datenkategorien: Definitionen und Beschreibung der Relation verschiedener linguistischer Datenkategorien.

Diese Normen werden im Folgenden kurz charakterisiert und diskutiert. Die Standards, die heute schon einsetzbar sind, werden durch Beispiele exemplifiziert.

2.1 Grundlagen linguistischer Annotation: Linguistic Annotation Framework (LAF)

Durch die Verwendung existierender Konventionen aus dem World Wide Web Consortium (W3C) wie XML (Bray et al., 2004), RDF (Beckett, 2004), OWL (McGuinness und van Harmelen, 2004), etc. versucht das Linguistic Annotation Framework eine einheitliche Grundlage für die Annotation von linguistischen Daten zu legen. Dabei liegt ein Schwerpunkt auf höheren Annotationsebenen, etwa morphosyntaktische, syntaktische und semantische Annotation, die auf tieferen Ebenen aufsetzen, ohne dabei gegenüber anderen Bereichen abgeschlossen zu sein.

Die auf der Grundlage verschiedener Bedürfnisse entwickelten Annotationsstandards, z.B. die Ergebnisse von EAGLES (Calzolari und McNaught, 1996; Leech und Wilson, 1996), ISLE (Atkins et al., 2002, 2003), etc. zu Morphosyntax, Syntax, Semantik und Lexikon haben zu einer Vielzahl von Inkompatibilitäten geführt. Um eine gemeinsame Basis existierender Annotation zu finden, wird daher mit Hilfe einer allgemeinen Merkmalsstruktur auf Grundlage von Datenkategorien, die ebenfalls zu standardisieren sind, ein generisches Datenformat definiert. Bestehende Annotationen sind daher in dieses Format transformierbar. Ziel ist also ein Metaformat, das es erlauben soll, linguistische Annotationen auszudrücken und auszutauschen.

Dabei stellt die Definition der Datenkategorien genauso ein Problem dar wie die Verwendung verschiedener Merkmalsstrukturhierarchien, die auf unterschiedlichen theoretischen Annahmen herrühren können. Das Problem der Definition von Datenkategorien soll dabei durch ein offenes Datenkategorien-Repository gelöst werden (siehe Abschnitt 2.7), wodurch eine maximale Unabhängigkeit von spezifischen Theorien möglich wird. Interessierte Kreise sollen die Möglichkeit erhalten, Vorschläge für Datenkategorien zu machen. Alle von einem dafür benannten Gremium akzeptierten Datenkategorien werden samt Beschreibung und Beispielen zentral gesammelt und jedem Benutzer zur Verfügung gestellt. Die Merkmalsstrukturhierarchie ist dagegen nicht als linguistische Theorie per se zu betrachten, auch wenn sie unter Umständen eine bestimmte linguistische Theorie abbildet, sondern nur als Austauschformat. Ob diese Trennung zwischen Theorie und Austauschformat allerdings vollständig erreicht werden kann, ist noch nicht beschrieben worden.

2.2 Grundlagen lexikalischen Markups: Lexical Markup Framework (LMF)

Im Bereich der Terminologiedatenbanken gibt es die Bestrebungen, für den Austausch ein Framework zu definieren, das als Grundlage für die Überarbeitung des *Machine Readable Terminology Interchange Format* (MARTIF, ISO 12200) dienen soll. Analog dazu werden derzeit Standards für die Beschreibung lexikalischer Datenbanken, insbesondere Wörterbücher mit definitorischen Inhalten, entwickelt, in denen formale Oberflächenmerkmale und Semantik voneinander getrennt strukturiert werden, und in denen lexikalische Strukturen eindeutig modelliert werden.

Verschiedene Applikationen, die Lexika verwenden, legen unterschiedliche Datenmodelle zugrunde. Um zu einer Vereinheitlichung dieser Datenmodelle für semantische Lexika zu kommen, beziehen sich die in der Entwicklung befindlichen Normen auf das Lexical Markup Framework (LMF), das eine Repräsentation lexikalischer Information in einem einheitlichen Modell darstellt, durch welches zunächst zumindest die Inhalte allgemeiner einsprachiger Definitionswörterbücher und ähnlich strukturierter Lexikondatenbanken repräsentiert werden können.

Derzeit ist bei der Erstellung des *LMF* nicht beschrieben, wo Grenzen liegen, wodurch nicht klar ist, welche lexikalischen Ressourcen mit seiner Hilfe abgebildet werden können. Da dieser Standard aber analog zu terminologischen Ressourcen im Terminology Markup Framework (Neufassung von ISO 12200:1999 (1999)) definiert wird, ist an die

Beschreibung semantischer Lexika im Sinne der lexikalischen Semantik zu denken. Eine Behandlung der Grenzen sollte jedoch Gegenstand von weiteren Normentwürfen sein, bevor dieser Standard verabschiedet werden kann.

2.3 Die Segmentierung von geschriebenen Wörtern für die Informationsverarbeitung

Als Grundlage für die automatische Verarbeitung von Sprachressourcen wird in der Regel davon ausgegangen, dass man verschiedenen Wörter im elektronisch verfügbaren Text voneinander unterscheiden kann. Dies setzt voraus, dass man eine Möglichkeit hat, Grenzen zwischen Wörtern zu ziehen. In westlichen Sprachen mit lateinischer Schrift wird diese Segmentierung von Wörtern durch Leerzeichen als typographische Konvention deutlich gemacht, was allerdings nicht für andere Schriftsysteme gelten muss, und in der Tat nicht allgemeingültig ist. Die Schreibung von Chinesisch, Japanisch und Koreanisch kommt z.B. traditionell ohne Leerstellen zwischen Wörtern aus (der aktuelle Normungsvorschlag ist übrigens von chinesischen Wissenschaftlern als *New Work Item* auf den Weg gebracht worden). Die Differenzierung von Wörtern und Phrasen aufgrund von linguistischen Kriterien ist auch nicht so allgemein, wie es Programmiersprachen antizipieren, die Wortgrenzen bei Interpunktionszeichen oder Leerzeichen als erreicht ansehen. Ein Beispiel dafür ist die nicht notwendigerweise eindeutige Segmentierung von Komposita, etwa durch verschiedene Möglichkeiten der Zusammen- oder Getrenntschreibung, die durch die Schreibung mit Binde-Strich (wie in diesem Wort) auch ein Hybrid kennen. In einem Handbuch eines deutschen Industrieunternehmens finden sich beispielsweise eine Vielzahl von Varianten für Begriffe wie *Gasmeßgerät*, *Gas-Meßgerät* und *Gasmeß-Gerät* nebeneinander. Ähnliches gilt z.B. in Paralleltexten, wenn etwa im Deutschen Komposita zusammen geschrieben werden, werden in relativ nah verwandten Sprachen wie Englisch dagegen als Kombination von zwei oder mehr typographischen Wörtern dargestellt.

Ziel des Normungsvorhabens ist eine Vereinheitlichung, um z.B. Benchmarks zu Evaluationszwecken von sprachverarbeitenden Systemen definieren zu können, primär mit dem Ziel einer Wortdefinition für nicht lateinische Schriftsysteme ohne Wortgrenzenkonventionen. Mittelfristig können die Ergebnisse von Wortsegmentierungsverfahren in der Informationsrecherche und der Terminologieextraktion unmittelbar eingesetzt werden.

Basierend auf linguistischen Regeln, Häufigkeit und Stabilität von Zeichenkombinationen wird die *Wortheit* von Mehrwort-Ausdrücken auf der Grundlage von Wortlisten aus Korpora bestimmt, und es wird ein Metamodell für die Segmentierung von Wörtern definiert.

Ein wesentliches Problem besteht in den unterschiedlichen Auffassungen zu Wortgrenzen und in der Verwendung von Worteinheiten in existierenden Systemen. Dieses Problem ist etwa aus der Praxis des Übersetzungswesens hinreichend bekannt, in dem Übersetzungsumfänge nach Wortzahl bewertet werden. Dies führt bereits bei der Bewertung von agglutinierenden Sprachen zu Problemen. Ein einheitliches Vorgehen wäre also auf Grund von praktischen Erwägungen sinnvoll, um eine einheitliche Bezugsgröße

definieren zu können. Das Vorhaben ist Ende 2005 auf dem Stand eines zur Normierung vorgeschlagenen Work Items.

2.4 Grundlagen für Morphosyntaktische Annotation: Morphosyntactic Annotation Framework (MAF)

Das Ziel des *Morphosyntactic Annotation Framework* ist eine einheitliche Kodierung von morphosyntaktischen Informationen, die in Datenströmen enthalten sind, also sowohl im Bereich der textuellen Daten als auch zur Signalannotation.

Der Entwurf des Morphosyntaktischen Annotations Frameworks (MAF) besteht aus zwei Teilen:

1. Die Segmentierung, also die Bestimmung der Wörter, die Behandlung von Ambiguitäten und die formale Beschreibung von internen Strukturen mit Hilfe von Merkmalsstrukturen (Attribut-Werte Paaren)

2. Eine inhaltliche Beschreibung der morphosyntaktischen Annotation, also eine Angabe zur Einbettung strukturierter Informationen. Dies schließt auch die Möglichkeit der multiplen Annotation mit ein, etwa für Numerus, Genus, Tempus, etc., weil ja viele Formen synkretistisch sind (*Hunde:* `nom/gen/acc plural`).

Diese Norm bezieht sich dabei unmittelbar auf relevante Datenkategorien zur Beschreibung der morphosyntaktischen Annotation. Ferner gibt es auch Anknüpfungen zur Segmentierungproblematik, da zu klären wäre, wie etwa für Deutsch Komposita zu behandeln sind, als mehrere Wörter oder als lexikalische Einheit. Die Arbeiten sind Ende 2005 bis zu einem Committee Draft gediehen.

2.5 Syntactic Annotation Framework (SynAF)

Innerhalb des eContent Projekts *LIRICS* (siehe auch `http;//lirics.loria.fr` und Sektion 3 unten), wird an einem Normvorhaben für syntaktische Annotationen gearbeitet. Ein entsprechendes Work Item wurde dazu dem ISO Commitee TC 37/SC4 vorgelegt und bereits akzeptiert.

Das *Syntactic Annotation Framework* (SynAF) verfolgt primär zwei Ziele:

1. Die Definition eines Metamodells für syntaktische Annotationen (ähnlich wie für die Segmentierung oder die Morphosyntax, wie weiter oben beschrieben).

2. Die Aufstellung einer Liste von Datenkategorien (s. Ide und Romary (2004)) als Grundlage für eine einheitliche syntaktische Annotation.

Die Standardisierungsarbeit von SynAF basiert auf den neuesten Entwicklungen im Bereich der syntaktischen Annotation, sowohl als Ausgabe von Parsern, die oft dem Zweck der Theorievalidierung dienen, als auch als bestehende Baumbanken, die primär als Trainingsdaten für Analysesysteme dienen. Das Metamodell und die Datenkategorien, die in

SynAF definiert werden, sollen dann die Interoperabilität und die Wiederverwendbarkeit von diesen zwei Typen von Ressourcen unterstützen.

Das Metamodell von SynAF muss flexibel genug sein, um die zwei Haupttypen von syntaktischen Annotationen abzudecken: Konstituentenstrukturen und Dependenzstrukturen.

Als Eingabematerial für SynAF werden folgende Ressourcen verwendet:

- Zum einen so genannte *legacy data*, die in Baumbanken zu finden sind, wie zum Beispiel innerhalb der *Penn Treebank*.
- Zum anderen bestehende Grammatiken, welche die syntaktischen Strukturen für verschiedene Sprachen abdecken.

Die Ausgangsbasis für SynAF besteht demnach in Korpora, die syntaktische Konstituenz und Dependenz kombinieren, wie TIGER (Uszkoreit, 2003) für das Deutsche, oder ISST (Montemagni et al., 2002) für das Italienische, aber auch Korpora zu nicht-europäischen Sprachen (siehe auch Abeillé et al. (2003)). Ebenfalls werden Ausgaben von syntaktischen Parsern berücksichtigt, die in verschiedenen Kontexten und Anwendungen entwickelt wurden (zum Beispiel HPSG, Pollard und Sag (1987), LS-GRAM (siehe auch LS-GRAM (2005) in der Bibliographie) und LFG Grammatiken (siehe hierzu die Referenz zum LFG Pargram Projekt (2005)) oder flache und robuste Grammatiken).

SynAF wird sich auch dem Thema der syntaktischen Ambiguitäten widmen (aufbauend hier auf bereits existierenden Vorschlägen, die in MAF gemacht worden sind, siehe Clément und de la Clergerie (2005)). Auch das Thema der mehrschichtigen Annotationen wird von SynAF angesprochen (zum Beispiel für die parallele Beschreibung von flachen vs. tiefen Analysen). Hier wird SynAF sich in das Linguistic Annotation Framework (LAF) einfügen. Dies spielt für die Beschreibung sogenannter langer Abhängigkeiten eine Rolle, die häufig eine eigene Annotationsschicht (*layer*) brauchen, um repräsentiert zu werden.

Ferner wird diskutiert, ob SynAF auch Information über syntaktischen Operationen beschreiben können soll, d.h. ob die Annotation auch Angaben über die beteiligten Prozesse auf dem Weg zum Analyseergebnis aufnehmen soll.

2.6 Repräsentation von Merkmalsstrukturen: Feature Structure Representation

Merkmalsstrukturen sind übliche Formalismen zur Beschreibung von Strukturen in vielen linguistischen Theorien und Ansätzen, etwa in der HPSG, LFG, im generativen Lexikon, etc, wo hierarchisierte Merkmalsstrukturen Verwendung finden. Abbildung 1 zeigt so eine Merkmalsstruktur für das englische Wort *dog* im HPSG-Paradigma.

Merkmalsstrukturen weisen einen hohen Informationsgehalt auf, sind sehr stark formalisiert und bieten sich daher für die automatisierte Verarbeitung an. Für die Verarbeitung im Bereich automatisierter Systeme werden dafür komplexe Strukturen und Programme eingesetzt, wobei nicht zuletzt die Komplexität den Bedarf nach einem einheitlichen Formalismus für den Austausch von Merkmalsstrukturen über einzelne

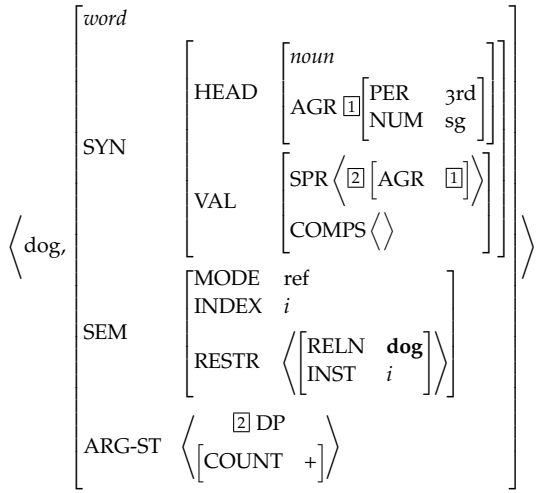

Abbildung 1: HPSG-Lexikoneintrag in Merkmalsstrukturform für das englische Wort *dog*, aus Sag et al. (2003), S. 254.

Systeme hinaus begründet. Hierzu kann man auf Vorarbeiten der Text Encoding Initiative (TEI, siehe Sperberg-McQueen und Burnard (2004)) zurückgreifen, die eine Syntax für die Beschreibung von Merkmalsstrukturen entwickelt hat.

Eine mögliche Repräsentation der in Abbildung 1 angegebenen Struktur gemäß einem Normentwurf ISO CD 24610-1:2003 (2003) stellt der Ausschnitt aus einem XML-Baum in Tabelle (1 – am Ende dieses Beitrags) dar.

Bei der Beschreibung dieser Merkmalsstrukturen gibt es natürlich Anknüpfungspunkte an die Problematiken anderer Standardentwürfe. So ist mit der Festlegung der Beschreibung von Merkmalsstrukturen zunächst einmal nicht festgelegt, inwieweit linguistische Theorien von ihren jeweiligen Repräsentationsformalismen, in diesem Fall also den getypten Merkmalsstrukturen, abzukoppeln sind. Das Problem der Trennung zwischen Repräsentationsformat und Theorie wird in dem Moment offensichtlicher, in dem Datenkategorien verschiedener linguistischer Theorien, entweder mit unterschiedlichem Namen und unterschiedlicher Semantik, mit unterschiedlichem Namen aber gleicher Bedeutung, oder gar mit gleichem Namen aber unterschiedlicher Bedeutung auftreten.

Zur Validierung der Angemessenheit des als Norm vorgeschlagenen Beschreibungsverfahrens sollten daher verschiedene linguistische Analysesysteme als Referenz implementiert und auf Interoperabilität zwischen den Merkmalshierarchien untersucht werden. Die Benennung von Datenkategorien ist dabei Gegenstand eigenständiger Normierungsbestrebungen. Nichtsdestotrotz ist die Beschreibung von Merkmalsstrukturen weit fortgeschritten und die Verabschiedung als internationaler Standard ISO 24610-1 darf für die nähere Zukunft erwartet werden.

2.7 Datenkategorien für elektronische lexikalische Ressourcen: Terminology and other language resources – Data categories

In der Sektion über die Merkmalsstrukturen wurde kurz auf das Problem der Bedeutung-Namens-Paaren in verschiedenen linguistischen Theorien hingewiesen. Ziel bei der Standardisierung von Datenkategorien für Sprachressourcen ist es, Datenkategorien für die Verwendung in lexikalischen Datenbanken zu definieren. Dabei wird von einer korrespondierenden, in Überarbeitung befindlichen Norm für Datenkategorien aus dem Bereich der Terminologie ausgegangen (ISO 12620:1999, 1999). Da erste Studien gezeigt haben, dass es erhebliche Überschneidungen zwischen Datenkategorien in Terminologie und Lexikographie gibt, wurde ein Vorschlag zur Trennung in verschiedene Standards mit einer allgemeinen Methodenspezifikation und separaten Kategoriebeschreibungen für Terminologie und Lexikographie verworfen.

Ausgehend von zentralen Datenkategorien werden bei der Standardisierung von Datenkategorien Mechanismen normiert, die der Erweiterung der Datenkategoriebasis dienen sollen (siehe auch Ide und Romary (2004)). Dies soll gewährleisten, dass Datenkategorien interoperabel sind und auch bei neueren oder verbesserten Theorien standardisierte Datenkategorien Verwendung finden können (einschließlich *Rückwärtskompatibilität*). Daher wurde zwar eine lange Liste von Datenkategorien erstellt, die teilweise mit Subkategorien versehen sind, zusammen mit einer Definition der Kategorie, aber der Schwerpunkt liegt auf der Definition eines Datenkategorie-Registers.

Eine besondere Komplexität erhält dieses *Data Category Repository* dadurch, dass es Offenheit gegenüber neuen Entwicklungen verlangt, was dazu führt, dass man Mechanismen zur Aufnahme von neuen Datenkategorien definieren muss, die ebenfalls unabhängig von Vorlieben und theoretischen Annahmen sind. Allerdings muss sichergestellt werden, dass die Offenheit nicht dazu führt, dass äquivalente Datenkategorien unabhängig voneinander definiert werden, was dem Grundsatz der Austauschbarkeit diametral entgegenstünde.

Diese Diskussion zeigt, dass die Normierung auf der einen Seite weit fortgeschritten ist, indem bereits ein Grundgerüst an Datenkategorien existiert, aber die Erweiterungsfunktionalität des Datenkategorieregisters und die eindeutige Beschreibung der Modalitäten, wie es ergänzt werden soll, sich noch in der Entwicklung befindet.

3 Einsatz von Standards in der Praxis

In vielen Feldern ist unmittelbar klar, dass es einen Bedarf an Standards für linguistische Annotationen gibt. Ein Beispiel dafür ist die transferbasierte maschinelle Übersetzung: Wenn standardisierte syntaktische Annotationen für Quell- und Zielsprache vorliegen, ist zu erwarten, dass Übersetzungssysteme mit geringerem Aufwand auf Seiten der Trainingsdaten erstellt werden können.

Die Anwendungen gehen jedoch weit über die Sprachverarbeitung hinaus, insbesondere in den Bereich des *Semantic Web* (Berners-Lee et al., 2001), einer Erweiterung des World Wide Webs. Damit das Semantic Web tatsächlich funktionieren kann, müssen Webseiten

semantisch annotiert werden. Im Bereich des *Semantic Web* versteht man unter semantischer Annotation dabei eine Verarbeitung, die einen Text mit Informationen anreichert, die aus Wissensbasen stammen, also aus Datenbanken, Taxonomien, Ontologien, etc. Es gibt aber wenige Werkzeuge, die diese Arbeit unterstützen, und selbst die existierenden Werkzeuge können nicht darüber hinweg täuschen, dass die semantische Auszeichnung extrem zeitintensiv ist. Daher gibt es Bemühungen, diese Art der Annotation zu automatisieren, und zwar auf der Grundlage von sprachverarbeitenden Werkzeugen, die den Webdokumenten eine (linguistische) syntaktische Struktur verleihen, bevor sie dann auf die Wissensbasen abgebildet werden, um semantische Annotationen zu generieren. Standardisierte linguistische Annotationen würden diesen Abbildungsprozess erheblich erleichtern (Buitelaar und Declerck, 2003).

Im gleichen Kontext wird nach Möglichkeiten gesucht, Wissensbasen automatisch aus größeren Dokumentmengen zu extrahieren, wobei auch maschinelles Lernen eingesetzt wird. Dieses Verfahren verlangt eine größere Menge von linguistischen Annotationen, die speziell auch Dependenz-Relationen aufweisen, damit sogenannte RDF-Tripel erzeugt werden können. Diese RDF-Tripel kann man sich als Subjekt, Objekt und Prädikat vorstellen, d.h. einem Gegenstand wird mittels eines Verbs eine Eigenschaft zugewiesen.

Um verfügbare linguistische Annotationen z.B. in Form von Baumbanken oder Ergebnissen von Parsern für Werkzeuge im Semantic-Web-Kontext verfügbar zu machen, müssen diese Ressourcen auf standardisierte Annotationen abgebildet werden, da es wesentlich einfacher ist aus einer großen Menge von standardisierten Annotationen Wissen zu akquirieren, als aus heterogenen oder gar idiosynkratisch annotierten Dokumenten.

Die Mitarbeit an Normungsaktivitäten und die Entwicklung von Standards basiert auf den fachlichen Interessen und Bedürfnis in verschiedenen Projekten, Vorhaben und Unternehmen. Auf Initiative einer französischen Forschungsorganisation (LORIA) wurde zusätzlich ein europäisches Projekt im Rahmen des *eContent*-Programms eingereicht und bewilligt, das sich schwerpunktmäßig auf die Erforschung von Ressourcen, sowie auf die benötigte standardisierte Infrastruktur konzentriert.

Das Projekt *LIRICS* (Linguistic Infrastructure for Interoperable Resources and Systems) läuft seit dem 1. Januar 2005 und treibt auf europäischer Ebene einige der oben vorgestellten Themen in enger Kooperation mit der ISO voran. So wurde zum Beispiel das ISO-Vorhaben *SynAF* (vgl. Abschnitt 2.5 oben) innerhalb von *LIRICS* initiiert. *LIRICS* entwickelt auch eine open-source Implementierung, Webservices und Testsuites für neun Sprachen, welche die Implementierung einiger der oben besprochenen Standards unterstützen und validieren. Die Notwendigkeit der Standardisierung von Sprachressourcen und ihre wirtschaftliche Relevanz für die Generierung von digitalen Inhalten ist dadurch auch auf europäischer Ebene dokumentiert und wird aktiv unterstützt.

Literatur

Abeillé, A., S. Hansen-Schirra und H. Uszkoreit (Hrsg.) (2003). *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest.

Atkins, S., N. Bel, F. Bertagna, P. Bouillon, N. Calzolari, C. Fellbaum, R. Grishman, A. Lenci, C. MacLeod, M. Palmer, G. Thurmair, M. Villegas und A. Zampolli (2002). From resources to applications. Designing the multilingual ISLE lexical entry. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, S. 687–693.

Atkins, S., N. Bel, P. Bouillon, T. Charoenporn, D. Gibbon, R. Grishman, C.-R. Huang, A. Kawtrakul, N. Ide, H.-Y. Lee, P. J. K. Li, J. McNaught, J. Odijk, M. Palmer, V. Quochi, R. Reeves, D. M. Sharma, V. Sornlertlamvanich, T. Tokunaga, G. Thurmair, M. Villegas, A. Zampolli und E. Zeiton (2003). Standards and best practice for multilingual computational lexicons and MILE (the multilingual ISLE lexical entry). Deliverable D2.2-D3.2 ISLE computational lexicon working group, International Standards for Language Engineering (ISLE), Pisa. Entstehungsjahr anhand von Dateimetadaten verifiziert.

Beckett, D. (2004). RDF/XML syntax specification (revised). URL: `http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/`.

Berners-Lee, T., J. Hendler und O. Lassila (2001). The Semantic Web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*.

Bray, T., J. Paoli, C. M. Sperberg-McQueen, E. Maler und F. Yergeau (2004). Extensible Markup Language (XML) 1.0 (third edition). URL: `http://www.w3.org/TR/2004/REC-xml-20040204/`.

Buitelaar, P. und T. Declerck (2003). Linguistic annotation for the Semantic Web. In S. Handschuh und S. Staab (Hrsg.), *Annotation for the Semantic Web*. Amsterdam: IOS Press.

Calzolari, N. und J. McNaught (1996). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: A common proposal and applications to european languages. Technical report, Expert Advisory Group on Language Engineering Standards (EAGLES).

Clément, L. und É. de la Clergerie (2005). MAF: A morphosyntactic annotation framework. In *Proceedings of the 2nd Language and Technology Conference (LT'05)*, Poznan, S. 90–94.

Ide, N. und L. Romary (2004). A registry of standard data categories for linguistic annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

ISO 12200:1999 (1999). Computer applications in terminology – machine-readable terminology interchange format (MARTIF) – negotiated interchange. Technical report, ISO.

ISO 12620:1999 (1999). Computer applications in terminology – data categories. Technical report, ISO.

ISO CD 24610-1:2003 (2003). Language resource management – feature structures – part 1: Feature structure representation. Technical report, ISO. Committee Draft.

Leech, G. und A. Wilson (1996). Recommendations for the morphosyntactic annotation of corpora. Technical report, Expert Advisory Group on Language Engineering Standards (EAGLES).

LFG Pargram Projekt (2005). LFG parallel grammar project. URL: `http://www2.parc.com/istl/groups/nltt/pargram/` – Stand: 6. Dezember 2005.

LS-GRAM (2005). LS-GRAM project. URL: `http://www.iai.uni-sb.de/iaide/en/ls_gram.htm` – Stand: 6. Dezember 2005.

McGuinness, D. L. und F. van Harmelen (2004). OWL web ontology language overview. URL: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, A. Lenci, O. Corazzari, A. Zampolli, F. Fanciulli, M. Massetani, R. Basili, R. Raffaelli, M. Pazienza, D. Saracino, F. Zanzotto, F. Pianesi, N. Mana und R. Delmonte (2002). Building the Italian syntactic-semantic treebank. In A. Abeillé (Hrsg.), *Building and Using syntactically annotated corpora*, S. 189–210. Dordrecht: Kluwer.

Pollard, C. und I. Sag (1987). *Head-Driven Phrase Structure Grammar*. CSLI and University of Chicago Press.

Sag, I. A., T. Wasow und E. M. Bender (2003). *Syntactic Theory* (2. Aufl.). Stanford: CSLI Publications.

Sperberg-McQueen, C. M. und L. Burnard (2004). TEI P4 guidelines for electronic text encoding and interchange XML-compatible edition. URL: http://www.tei-c.org/P4X/.

Uszkoreit, H. (2003). TIGER project. URL: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/ – Stand: 6. Dezember 2005.

```
<fs>
  <f org="list" name="dog">
     <fs>
       <f name="orth"><str>dog</str></f>
       <f name="word"/>
       <f name="syn">
         <fs>
             <f name="head">
                <fs>
                   <f name="noun"/>
                   <f name="agr">
                      <fs id="one">
                          <f name="per"><str>3rd</str></f>
                          <f name="num"><str>sg</str></f>
                      </fs>
                   </f>
                </fs>
             </f>
              <f name="val">
                <fs>
                   <f name="spr" org="list">
                      <fs>
                         <f name="null" fVal="two"/>
                         <f name="agr" fVal="one"/>
                      </fs>
                   </f>
                   <f name="comps" org="list"/>
                </fs>
              </f>
         </fs>
       </f>
       <f name="sem">
          <fs>
              <f name="mode"><str>ref</str></f>
              <f name="index"><str>i</str></f>
              <f name="restr">
                <fs>
                    <f name="reln"><str>dog</str></f>
                    <f name="inst"><str>i</str></f>
                </fs>
              </f>
         </fs>
       </f>
       <f name="arg-st" org="list">
         <fs id="two"><f name="dp"/></fs>
         <fs>
              <f name="count"><plus/></f>
          </fs>
       </f>
     </fs>
  </f>
</fs>
```

Tabelle 1: Eine mögliche XML-basierte Repräsentation der in Abbildung 1 angegebenen Struktur.