Der TUSNELDA-Standard: Ein Korpusannotierungsstandard zur Unterstützung linguistischer Forschung

Andreas Wagner[1] und Laura Kallmeyer[2]

[1] SFB 441, Universität Tübingen, Nauklerstr. 35, D-72074 Tübingen, Germany
wagner@sfs.nphil.uni-tuebingen.de

[2] UFRL, Université Paris 7, 2, place Jussieu, F-75251 Paris Cedex 05, France
laura.kallmeyer@linguist.jussieu.fr

Zusammenfassung Die Verwendung von Standards für die Annotierung größerer Sammlungen elektronischer Texte (Korpora) ist eine Voraussetzung für eine mögliche Wiederverwendung dieser Korpora. Dieser Artikel stellt einen Korpusannotierungsstandard vor, der die Anforderungen der Untersuchung unterschiedlichster linguistischer Phänomene berücksichtigt. Der Standard wurde im SFB 441 an der Universität Tübingen entwickelt. Er geht von bestehenden Standards, insbesondere CES und TEI, aus, die sich als teilweise zu ausführlich und zu wenig restriktiv, teilweise auch als nicht ausdrucksstark genug erweisen, um den Bedürfnissen korpusbasierter linguistischer Forschung gerecht zu werden.

1 Standards zur Korpusannotation

Die Annotation von Textkorpora mit klassifizierenden bzw. strukturierenden Informationen gewinnt für die linguistische Forschung zunehmend an Bedeutung. Die automatische Extraktion linguistisch interessanter Phänomene aus Korpora wird durch geeignete Annotationen wesentlich erleichtert und häufig erst ermöglicht.

Um die Wiederverwendbarkeit von Korpora zu garantieren und deren Zugänglichkeit zu erleichtern, ist es unumgänglich, sich bei der Annotation an gewisse Standards zu halten. Dies betrifft zum einen das abstrakte Format und zum anderen die inhaltlichen Prinzipien der Annotation. Die Verwendung eines standardisierten Kodierungsformates erlaubt es, einheitliche Tools für die Verarbeitung von Korpora (Annotation, Abfrage) einzusetzen. Eine Standardisierung der verwendeten Markierungen (Tags) dagegen erleichtert für den Benutzer die Arbeit mit verschiedenen Korpora, da eine Einarbeitung in ein neues Tagset deutlich vereinfacht, in manchen Fällen sogar unnötig wird.

Als Standardformat zur Kodierung von Korpora haben sich mittlerweile SGML (siehe [6], Teil 1, Kapitel 2 für eine Einführung in SGML) bzw. XML [1] etabliert, Metasprachen, die auf einer abstrakten Ebene Form und Struktur von Tags sowie Formate für Tag-Definitionen festlegen. Bezüglich der konkreten Tags, die bei der Annotation von Korpora verwendet werden, gibt es zwei Standards, den von der Text Encoding Initiative (TEI) entwickelten SGML-basierten

Standard [6] und den von der Expert Advisory Group on Language Engineering Standards (EAGLES) vorgeschlagenen Corpus Encoding Standard (CES), der zunächst in SGML [2] und anschließend auch in einer XML-Fassung [3] formuliert wurde.

Der in diesem Artikel vorgestellte Standard basiert auf TEI und vor allem auf CES. Daher soll im folgenden kurz auf diese beiden Standards eingegangen werden.

Bei der Text Encoding Initiative (TEI) handelt es sich um ein 1988 begonnenes internationales Projekt, dessen Ziel die Entwicklung von Empfehlungen zur Annotation elektronischer Texte für Forschungszwecke war. Die Empfehlungen wurden 1994 veröffentlicht und sind in SGML formuliert. Die TEI-Empfehlungen wurden nicht speziell im Hinblick auf Korpora entwickelt, sondern sind wesentlich allgemeiner. TEI berücksichtigt verschiedene Textsorten, beispielsweise Prosa, Gedichte, Dramen, Wörterbücher, Briefe, gesprochene Texte usw. Die TEI-Empfehlungen sind sehr detailliert. Zum einen decken sie verschiedenste Aspekte von elektronischen Texten ab und zum anderen lassen sie häufig mehrere Möglichkeiten zu, ein bestimmtes Phänomen zu annotieren. Beide Punkte können bei der Kodierung von Korpora nachteilig sein. Ein großer Teil der von TEI vorgeschlagenen Kodierungsmöglichkeiten wird speziell für linguistische Zwecke nicht benötigt. Die Übernahme dieser Vorschläge brächte also keinen Nutzen, würde aber die Unübersichtlichkeit der Annotationsrichtlinien erheblich vergrößern. Der zweite Punkt ist noch problematischer: Um eine exhaustive Korpus-Suche nach bestimmten Phänomenen mit möglichst einfachen Suchanfragen zu ermöglichen, sollten gleiche Phänomene immer auf die gleiche Weise annotiert sein, d. h. alternative Kodierungen eines Phänomens sollten soweit wie möglich ausgeschlossen sein.

Der von EAGLES vorgeschlagene Corpus Encoding Standard (CES) wurde basierend auf den TEI-Empfehlungen entwickelt mit dem Ziel, diese Empfehlungen den speziellen Bedürfnissen der Annotation linguistischer Korpora anzupassen. CES sieht deutlich weniger Tags vor als TEI und ist wesentlich restriktiver. Insbesondere wird der Grundsatz verfolgt, dass jeweils nur eine mögliche Kodierung für ein bestimmtes Phänomen vorgesehen ist.

Das höchste Element in CES ist ein Korpus, bestehend aus einem Header gefolgt von entweder einem oder mehreren Dokumenten oder einem oder mehreren Teilkorpora. Ein Dokument wiederum besteht aus einem Header und einem Text.

Ein Header enthält

1. eine bibliographische Beschreibung des Korpus (`fileDesc`)
2. Information über die Beziehung zwischen elektronischem Text/Korpus und Originaltext (`encodingDesc`)
3. weitere Informationen über das Korpus (`profileDesc`) und
4. Information über Revisionen des Korpus (`revisionDesc`).

Ein Text besteht aus einem Body oder eine Gruppe von Bodies. Ein Body ist unterteilt in Paragraphen oder größere Abschnitte (Divisions). CES unterscheidet Elemente auf Paragraphenebene (Dialoge, Gedichte, Abbildungen, Absätze etc.) und Elemente, die innerhalb von Elementen auf Paragraphenebene auftreten können (Sätze, Referenzen, Korrekturen, Abkürzungen, Namen etc.).

2 Der TUSNELDA-Standard

Der zentrale Forschungsgegenstand des SFB 441 sind die empirischen Datenstrukturen, die die Grundlage linguistischer Theoriebildung darstellen. Ein Großteil der SFB-Projekte erstellt Korpora, die der empirischen Untersuchung verschiedener linguistischer Phänomene in verschiedenen Sprachen dienen (z. B. Modalverben im Deutschen, Anrede und Höflichkeit im Russischen, deiktische Ausdrücke im Serbischen/Kroatischen etc.). Diese Korpora werden zu einem Gesamtkorpus TUSNELDA (TUebinger Sammlung Nutzbarer Empirischer DAtenstrukturen) zusammengefasst.

Bei der Wahl des Annotationsschemas für TUSNELDA waren folgende Ziele zu berücksichtigen:

1. Erfassung der für den jeweiligen Forschungsgegenstand relevanten Phänomene
2. einheitliche Kodierung jedes Phänomens
3. Wiederverwendbarkeit der annotierten Korpora

Aus den Ausführungen in Abschnitt 1 folgt, dass CES den Punkten 2 und 3 gerecht wird. Deshalb wurde CES als Ausgangspunkt für die Entwicklung eines TUSNELDA-Annotationsstandards verwendet. TUSNELDA wurde zunächst in SGML kodiert, wobei die DTD (Document Type Definition) so konzipiert wurde, dass eine automatische Konvertierung in XML möglich ist.[1] Die meisten der in CES enthaltenen Definitionen konnten für TUSNELDA unverändert übernommen werden. Dazu zählt auch die am Ende von Abschnitt 1 skizzierte Grobstruktur eines (Teil-)Korpus und eines einzelnen Dokuments.

Jedoch hat sich gezeigt, dass CES für die Annotierung einer Reihe von Phänomenen und Textmerkmalen, deren Erfassung für die Forschung im SFB 441 äußerst wichtig sind, keine geeigneten Mechanismen zur Verfügung stellt. Daher war es erforderlich, einige Definitionen zu modifizieren, sowie neue Elemente aufzunehmen. Für manche Phänomene bot TEI eine geeignete Kodierung, für andere Phänomene musste eine neue Kodierung entwickelt werden.

Außerdem wurden einige Definitionen restriktiver gefasst, z. B. in Bezug auf die Optionalität von Elementen oder Attributen. Damit soll sichergestellt werden, dass bestimmte, als wichtig erachtete Informationen in allen Teilkorpora kodiert sind.

Etwa 1/3 der Definitionen in TUSNELDA weichen von den entsprechenden CES-Definitionen ab bzw. haben keine Entsprechung in CES. Knapp 2/3 der TUSNELDA-Definitionen wurden unverändert von CES übernommen. So

[1] Teilweise weicht TUSNELDA hierbei bewusst von der XML-Version von CES (XCES) ab, vgl. Abschnitt 4.1.

ist ein Annotationsstandard entstanden, der ein starkes Gewicht auf Mechanismen für philologische empirische Forschung legt, jedoch überwiegend an einen bestehenden Standard angelehnt ist, was eine maximale Wiederverwendbarkeit sicherstellt.

Die TUSNELDA-DTD und Empfehlungen für die Benutzung von TUSNELDA finden sich in [5].

3 Vergleich mit CES und TEI

In diesem Abschnitt wird ein Gesamtüberblick über die Gemeinsamkeiten und Unterschiede zwischen CES und TUSNELDA gegeben. Ausgangspunkt für die TUSNELDA-DTD war die CES-DTD für primäre Daten (`cesDoc.dtd`) sowie die CES-Definitionen für den Header (`header.elt`).[2] 100 der 140 dort definierten Tags und Entities konnten unverändert für TUSNELDA übernommen werden. 34 CES-Tags bzw. -Entities wurden modifiziert. 21 Tags wurden neu hinzugefügt, von denen 6 Tags (in adaptierter Form) aus verschiedenen TEI-Modulen übernommen wurden.

Im Folgenden werden die einzelnen Tags und Entities in TUSNELDA, jeweils getrennt für den Header und den eigentlichen Text, in der Reihenfolge ihres Auftretens in der DTD aufgelistet. Bei den modifizierten bzw. neu hinzugefügten Tags werden zusätzliche Angaben (Art der Modifikation, ggfs. TEI-Modul) gemacht. Ein Teil der Modifikationen, die die Kodierung von Texten betreffen, werden in Abschnitt 4 ausführlich behandelt. In 3.1 und 3.2 werden die Änderungen kurz skizziert, auf die in 4 nicht eingegangen wird.

3.1 Header

Folgende Elemente und Entities wurden aus der CES-Header-DTD unverändert übernommen:

> %a.header %a.declarable h.title respType respName wordCount byteCount extNote distributor pubAddress telephone fax eAddress idno availability pubDate h.bibl biblStruct analytic monogr h.author edition imprint pubPlace publisher biblScope biblNote biblFull projectDesc samplingDecl quotation hyphenation transduction tagsDecl tagUsage annotations annotation refsDecl classDecl taxonomy category catDesc langUsage wsdUsage writingSystem textClass catRef h.keywords keyTerm translations translation translator revisionDesc change changeDate h.item

In Tabelle 1 sind die modifizierten, in Tabelle 2 die neu aufgenommenen Elemente aufgelistet.

Für `<extent>` wurden die Unterelemente `<tokenCount>` und `<characterCount>` neu eingeführt. Diese geben die Anzahl der Token (Wörter + Satzzeichen) bzw. die Anzahl der Zeichen (ohne Annotation) eines Textes bzw. Korpus

[2] Dementsprechend besteht die TUSNELDA-DTD aus den beiden Dateien `tusneldaDoc.dtd` und `tusneldaHeader.elt`.

Tabelle 1. Modifizierte CES-Definitionen im TUSNELDA-Header

Tag/Entity	Art der Modifikation
tusneldaHeader	Umbenennung (cesCorpus), Attribute weggelassen, Attribute obligatorisch gemacht
fileDesc	Content restringiert
titleStmt	Content restringiert
respStmt	Content restringiert
editionStmt	Attribut obligatorisch gemacht
extent	Content Struktur geändert
publicationStmt	Content restringiert
sourceDesc	Content erweitert
encodingDesc	Content erweitert
editorialDecl	Content restringiert
correction	Attribut Default geändert
segmentation	Content Struktur geändert
normalization	Attribut Default geändert
profileDesc	Content restringiert
creation	Attribute hinzugefügt, Attribut weggelassen
language	Attribut hinzugefügt

an. In `<sourceDesc>` wurde für bibliographische Angaben von Ton- und Video-Aufzeichnungen ein neues Unterelement `<recordingStmt>` mit Unterelementen `<recording>`, `<recNote>`, `<equipment>` und `<broadcast>` von TEI übernommen. Für `<language>` wurde ein Attribut `ethnologue` eingeführt, in dem die verwendete Sprache, neben der aus CES übernommenen Klassifizierung nach ISO 639, gemäß [4] klassifiziert werden kann.

In `<editorialDecl>` wurde das Unterelement `<conformance>` weggelassen, da für TUSNELDA (noch) keine "conformance levels" wie für CES definiert wurden. Die CES-Definition von `<respStmt>` wurde modifiziert, um ihre Konvertierbarkeit in XML zu ermöglichen. Die Änderungen in `<tusneldaHeader>`, `<fileDesc>`, `<titleStmt>`, `<publicationStmt>`, `<correction>` und `<normalization>` hängen mit stärkeren Restringierungen in TUSNELDA im Vergleich zu CES zusammen. Die Modifikationen von `<segmentation>` (neue Unterelemente `<tag>`, `<segmMethod>` und `<segmNote>`) und `<creation>` sowie die Platzierung von `<annotations>` unter `<encodingDesc>` statt unter `<profileDesc>` ermöglichen eine besser strukturierte Repräsentation bestimmter Informationen.

3.2 Text

Folgende Elemente und Entities in TUSNELDA wurden aus der CES-Document-DTD unverändert übernommen:

> %a.global %a.text %x.token %m.token %base.seq %phrase.seq %par.seq text body group div opener head keywords byline docAuthor dateline

Tabelle 2. Neue Definitionen im TUSNELDA-Header

Tag/Entity	TEI-Modul	Art der Modifikation
tokenCount	—	—
characterCount	—	—
recordingStmt	teihdr2	Content restringiert
recording	teihdr2	Content erweitert, Content restringiert
recNote	—	—
equipment	teihdr2	Content Struktur geändert
broadcast	teihdr2	Content restringiert
tag	—	—
segmMethod	—	—
segmNote	—	—

address closer p list item label note bibl author poem lg l figDesc table row cell caption speaker stage date name term time title num ptr ref

In Tabelle 3 sind die modifizierten, in Tabelle 4 die neu aufgenommenen Elemente aufgelistet.

Die Erweiterungen von `%m.inter` und `%m.phrase` ergeben sich aus der Einführung der neuen Elemente `<linkGrp>` auf Paragraphenebene und `<marked>` und `<unclear>` auf Satzebene. In TUSNELDA kann `<quote>` auch Tabellen enthalten, was in CES nicht zugelassen ist. Die beiden Elemente `<s>` und `<q>` haben jeweils ein Attribut `nested` bekommen, mit dem man kodieren kann, ob ein Satz/Zitat in einen anderen Satz / ein anderes Zitat eingebettet ist oder nicht. Bei `<gap>`, `<reg>`, `<corr>` und `<abbr>` wurde der Wert des Attributs `cert` auf `sure`, `probable` oder `presumable` beschränkt, um einigermaßen aussagekräftige Werte zu erhalten. Bei `<hi>`, `<foreign>`, `<distinct>` und `<mentioned>` wurde die nicht XML-konforme *exclusion exception* entfernt. Bei `<measure>` sind in TUSNELDA keine konkreten Werte für `type` vorgegeben, weil dies zu restriktiv ist. Das Element `<unclear>` zum Kodieren von unklaren (z. B. unleserlichen) Textstellen wurde aus TEI übernommen. `<linkGrp>` fasst `<link>`-Elemente zusammen. `<xptr>`, ein mögliches Unterelement von `<linkGrp>`, kann dokumentübergreifende Pointer enthalten, was zur Zeit in TUSNELDA zwar noch nicht benötigt wird, aber später vielleicht wichtig wird.

4 Ausgewählte Modifikationen von CES in TUSNELDA

4.1 Annotation von Dialogen

Dieser Abschnitt behandelt eine der technisch motivierten Änderungen, die aufgrund der für TUSNELDA verlangten XML-Konvertierbarkeit gegenüber CES vorgenommen werden mussten.

Zur Annotation von Dialogen, z. B. in Dramen, sieht CES ein Element `<sp>` (speech) mit einem Unterelement `<stage>` für Regieanweisungen vor. `<stage>`

Tabelle 3. Modifizierte Definitionen im TUSNELDA-Document

Tag/Entity	Art der Modifikation
%m.inter	Content erweitert
%m.phrase	Content erweitert
tusneldaCorpus	Umbenennung (cesCorpus), Attribut weggelassen
tusneldaDoc	Umbenennung (cesDoc), Attribut weggelassen
quote	Content erweitert
figure	Content erweitert
sp	Content Struktur geändert
s	Attribut hinzugefügt
q	Attribut hinzugefügt
gap	Attribut Wertebereich restringiert
reg	Attribut Wertebereich restringiert
corr	Attribut Wertebereich restringiert
hi	Content erweitert
foreign	Content erweitert
distinct	Content erweitert
mentioned	Content erweitert
measure	Attribut Wertebereich erweitert
abbr	Attribut Wertebereich restringiert

kann beliebig tief in `<sp>` eingebettet sein, also als direktes Unterelement oder als Unterelement eines Unterelementes von `<sp>` usw. auftreten. In SGML wird dies mit einer *inclusion exception* ausgedrückt. Dieses Konstrukt ist in XML jedoch nicht zulässig und konnte deshalb nicht in TUSNELDA übernommen werden.

XCES (CES in XML) sieht `<stage>` nur als direktes Unterelement von `<sp>` vor. Die XCES-Lösung ist zwar XML-konform, hat aber zwei Nachteile: Erstens ist die Zusammenfassung eines Sprechers und seines Textes in `<sp>` (wie in CES) nicht mehr gegeben, denn `<sp>` erlaubt beliebig viele Elemente `<speaker>`, `<p>` und `<stage>` in beliebiger Reihenfolge. Zweitens lassen sich Auftreten von Regieanweisungen innerhalb von Paragraphen nicht adäquat annotieren.

Für TUSNELDA wurde daher ein neues Element `<spokenPar>` definiert, das wie `<p>` aufgebaut ist, aber zusätzlich `<stage>`-Elemente enthalten kann. `<sp>` enthält in TUSNELDA beliebig viele Elemente `<speaker>` gefolgt von beliebig vielen Elementen `<stage>` oder `<spokenPar>`. Dies garantiert, dass mit jeder neuen Sprecherangabe ein neues Element `<sp>` beginnen muss. Außerdem kann `<stage>` sowohl auf Paragraphenebene (als Unterelement von `<sp>`) als auch als Teil eines Paragraphen (als Unterelement von `<spokenPar>`) auftreten. Beispiel:

```
<sp who="Lady Windermere">
    <speaker>Lady Windermere.</speaker>
    <spokenPar>That will do !</spokenPar>
</sp>
<sp><stage>Exit Parker C .</stage></sp>
<sp who="Lady Windermere">
```

Tabelle 4. Neue Definitionen im TUSNELDA-Document

Tag/Entity	TEI-Modul	Art der Modifikation
figTrans	—	—
display	—	—
spokenPar	—	—
displayedPar	—	—
situation	—	—
unclear	teicore2	Attribut Wertebereich restringiert, Attribute weggelassen
marked	—	—
anchor	teilink2	Attribute weggelassen
xptr	—	aus cesAlign.dtd übernommen
linkGrp	—	aus cesAlign.dtd übernommen
link	—	aus cesAlign.dtd übernommen

```
    <spokenPar><stage>Speaking to Lord Windermere</stage>
       Arthur , if that woman comes here - I warn you -
    </spokenPar>
</sp>
```

4.2 Markierung spezifischer linguistischer Einheiten

In diesem und den folgenden Abschnitten geht es um linguistisch motivierte Änderungen, die aufgrund der in den SFB-Projekten gewünschten Annotationen gegenüber CES vorgenommen werden mussten.

Wie bereits erwähnt, beschäftigen sich die verschiedenen SFB-Projekte mit unterschiedlichen linguistischen Objekten (Modalverben, deiktische Ausdrücke, Temporaladverbien etc.). Diese Objekte sollen in den jeweiligen Teilkorpora von TUSNELDA automatisch auffindbar sein. Um dies auf einfache Weise zu ermöglichen, wurde das Element `<marked>` zur Kennzeichnung der zu untersuchenden linguistischen Einheiten eingeführt. `<marked>` hat das Attribut `type`, das die Kategorie der betreffenden Einheit angibt, z. B. `<marked type="adv-tmp">morgen </marked>`. Bestimmte linguistische Einheiten ließen sich nach einer elaborierten linguistischen Annotation (POS-Tagging, Parsing) auch ohne zusätzliche Markierung automatisch auffinden. Jedoch ist eine derartige Annotation häufig (z. B. für die Suche nach Höflichkeitsformen) nicht ausreichend.

4.3 Transkription gesprochener Dialoge

Eines der Projekte erstellt Transkriptionen von Video- und Audio-Aufnahmen. Ein Problem hierbei ist die Modellierung paralleler bzw. überlappender Äußerungen. CES bietet keinen Mechanismus zur Alignierung von Äußerungen innerhalb eines Textes. TEI enthält dagegen mehrere geeignete Alignierungsmechanismen. TUSNELDA lässt nur einen dieser Mechanismen zu: Ein Element `<anchor>` erlaubt es, Referenzpunkte in Äußerungen zu markieren; mit Hilfe eines Elements

<link> können Referenzpunkte in verschiedenen Äußerungen miteinander verbunden werden. Auf diese Weise lassen sich simultane Zeitpunkte in verschiedenen Äußerungen kennzeichnen.

4.4 Transkription von Comics

Ein anderes Projekt untersucht Comics, um Zusammenhänge zwischen dem Gebrauch von lokalen Deiktika und Zeigegesten zu ermitteln. Die Transkription eines Bildes sollte daher die Äußerung eines Sprechers und seine Zeigegeste als zusammengehörend kodieren. Außerdem sollte unterschieden werden zwischen im Bild (gewöhnlich in Kästchen am Bildrand) auftretenden "Metatexten" einerseits und vom Annotator hinzugefügten Situationsbeschreibungen andererseits. Das für die Kodierung von Bildern in CES und TEI vorgesehene Element <figure> bietet keine Möglichkeit, diese Differenzierungen explizit auszudrücken. Deshalb wurde in TUSNELDA ein neues Element <figTrans> als Unterelement von <figure> eingeführt. <figTrans> besteht aus einem oder mehreren <sp>-Elementen (vgl. 4.1), die jeweils die Äußerung eines Sprechers und die zugehörige Zeigegeste umfassen. Gesten werden in einem (ebenfalls neu eingeführten) Element <situation> kodiert und mit Hilfe von Schlüsselwörtern klassifiziert. Im Bild vorhandene "Metatexte" werden mit <stage> annotiert. Beispiel:

```
<figure id="s17b7">
  <figTrans>
    <sp who="Gutemine">
      <spokenPar>Also gut ! Ich komme gerade vom Seher . Er ist
        <marked type="deik-lok">dahinten</marked> im Wald . . .
        Aber sag ’ s nicht weiter !<ptr target="s17b7n">
      </spokenPar>
      <situation>
        <keywords>
          <term>Daumen</term>
          <term>gebeugt</term>
        </keywords>
      </situation>
    </sp>
    <sp>
      <stage id="s17b7n">Gallische Redensart</stage>
    </sp>
  </figTrans>
</figure>
```

5 Zusammenfassung und Ausblick

In diesem Artikel wurde ein Korpusannotierungsstandard vorgestellt, der sich weitgehend an bestehende Standards anlehnt, jedoch spezifische Anforderungen vielfältiger sprachwissenschaftlicher Fragestellungen berücksichtigt. Damit garantiert der Standard einerseits Wiederverwendbarkeit und einheitliche Verarbeitung der kodierten Korpora und weist andererseits die nötigen Mechanismen für differenzierte empirische linguistische Untersuchungen auf.

Die Entwicklung des TUSNELDA-Standards ist insofern ein fortlaufender Prozess, als sich die Anforderungen der Projekte an die zu untersuchenden Daten verändern bzw. weiterentwickeln. Der Standard wird entsprechend diesen Erfordernissen angepasst werden. Geplant ist beispielsweise die Integration eines Schemas für morphologische und syntaktische Annotation.

Danksagung

Die Entwicklung des hier vorgestellten Korpusannotierungsstandards wäre ohne die Mitarbeit der verschiedenen Projekte des SFB 441 nicht möglich gewesen. Ausgangspunkt für die Arbeit an dem TUSNELDA Standard war ein mit Mitarbeitern des SFB und anderer korpuslinguistischer Projekte in Tübingen veranstalteter Workshop, der eine kritische Begutachtung von CES in Hinblick auf die Arbeit des SFB zum Ziel hatte. Dieser Workshop führte zu einem ersten Entwurf von TUSNELDA. Weitere Änderungen ergaben sich aus der Arbeit einzelner SFB-Projekte mit der TUSNELDA-DTD. Für viele wertvolle Anregungen und Hinweise bei der Entwicklung von TUSNELDA möchten wir den Teilnehmern des ersten Workshops und denjenigen, die ihre Arbeit mit TUSNELDA mit uns diskutiert haben, danken. Dies sind Michael Betsch, Bernhard Brehmer, Hervé Dejean, Sam Featherston, Gabriela Fulir, Alda Galsterer, Stefanie Herrmann, Konstanze Jungbluth, Sandra Kübler, Lothar Lemnitzer, Jürgen Mellinger, Detmar Meurers, Frank Henrik Müller, Reimar Müller, Slavica Stevanović, Tylman Ule und Lars Wigger.

Literatur

1. Bray, T., Paoli, J., Sperberg-McQueen C. M., Maler, E. Hrsg.: Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation. W3C, 2000 `http://www.w3.org/TR/2000/REC-xml-20001006`.
2. Expert Advisory Group on Language Engineering Standards (EAGLES): Corpus Encoding Standard – Document CES 1. Version 1.5, 20 March 2000. `http://www.cs.vassar.edu/CES/`.
3. Expert Advisory Group on Language Engineering Standards (EAGLES): XCES Corpus Encoding Standard for XML. XML version of the CES DTDs. Document XCES 0.2., 16 August 2000. `http://www.cs.vassar.edu/XCES/`.
4. Grimes, B. F. Hrsg.: Ethnologue: Languages of the World, Thirteenth Edition. SIL Publications, 1996.
5. Kallmeyer, L., Meyer, R., Wagner, A.: Guidelines for the TUSNELDA Corpus Annotation Standard. To appear. `http://www.sfb441.uni-tuebingen.de/c1/tusnelda_guidelines.html`.
6. Sperberg-McQueen, C. M., Burnard, L., Hrsg.: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative, Chicago and Oxford, 1994. `http://etext.virginia.edu/TEI.html`.