A Flexible Framework for Integrating Annotations from Different Tools and Tagsets

Christian Chiarcos
Universität Potsdam
chiarcos@ling.uni-potsdam.de

Stefanie Dipper
Ruhr-Universität Bochum
dipper@linguistics.rub.de

Michael Götze
Universität Potsdam
goetze@ling.uni-potsdam.de

Julia Ritz
Universität Potsdam
julia@ling.uni-potsdam.de

Manfred Stede
Universität Potsdam
stede@ling.uni-potsdam.de

Abstract

Tools for linguistic annotation employ different data models and accompanying visualization metaphors, depending on the particular type of annotation envisaged. When a corpus is to be annotated on multiple layers, and the annotations are to be related to one another, the output formats of the annotation tools need to be unified. We describe an implemented framework for this step: reading the output of a variety of tools into a single database, where the data can be visualized, queried and evaluated across the layers. Then, besides the integration of resources at *format* level, we also seek compatibility between annotation *tagsets*: We describe how ontologies can be used to mediate between competing tagsets intended to cover the same class of linguistic phenomena.

1 Introduction

Manual linguistic annotation is labour-intensive and expensive. It is therefore of utmost importance to provide software environments that ensure the efficiency of the overall process. This can be done in three different ways: (1) by careful selection of the data to be annotated; (2) by providing (partial) automatic analyses that only need to be confirmed or changed by the human annotator; or (3) by tailoring the data models and the look-and-feel of annotation tools as good as possible to the kind of annotation performed.

In this paper, we are focusing on the last option. Nowadays, a variety of annotation tools are freely available, which support different styles of annotation for different purposes, such as layer-based transcription or labelling of words/phrases, coreference links, syntax trees, or discourse trees.

Another trend that has emerged in recent years is the availability of corpora annotated simultaneously on various levels, so that inter-relationships between the annotations can be explored. When such multilayer corpora are to be created with existing dedicated annotation tools, a new problem arises: Output formats of the annotation tools can differ considerably, and annotations need to be aligned in order to be useful for purposes such as those mentioned above. To solve these problems, we have developed a software framework involving a generic standoff representation format; conversion from tool output to the generic format; aligning the annotations in a database that allows for visualization (which is not covered in this paper), retrieval, and statistical analyses of the data. By integrating an ontology in the query mechanism, resources based on differing annotation schemes can be queried simultaneously.

2 A Generic Standoff Format for Integrating Annotations

2.1 Representational standards

Nowadays the need for standardized annotation schemes and representation formats is widely recognized. Language resources must be well-documented and annotations be easy to interpret if they are to be beneficial for users other than the cor-

pus developers themselves. Standardization of representation formats concerns both the *physical* and *logical* data structures (see, e.g., (Schmidt, 2004)).

The logical data structure refers to the *data models* that are used to model the linguistic phenomena and their properties. We distinguish three types of data structures: (i) "annotation graphs": labeled directed acyclic graphs (LDAGs) whose nodes refer to a time line; annotation graphs are typically used for modeling time-aligned information (Bird and Liberman, 2001); (ii) structural annotations: DALGs whose nodes refer to other nodes; usually used for syntactic and other tree-like annotations; (iii) feature structures, used, e.g., for syntactic analyses in frameworks such as HPSG and LFG, but rarely used in the context of corpus annotation.

The division between the paradigms of time-aligned annotation graphs and hierarchical structures has weakened in recent years. For instance, the data model of annotation graphs has been generalized, resulting in the format ATLAS (Laprun et al., 2002), which supports both annotation graphs and hierarchical structures. Similarly, the NITE Object Model (Carletta et al., 2003b) and the general-purpose Linguistic Annotation Framework (Ide et al., 2003) serve both camps.

The physical data structure, on the other hand, refers to the "external" representation of the data. Here the de-facto standard is XML for serializing the data. Often, a standoff-architecture is used, which stores primary data and its annotations in different files (Sperberg and Bernard, 1994; Dybkjær et al., 1998). For the serialization of structural annotations, a natural way to represent trees is by using XML embedding structures. If structural annotations contain non-tree-like structures (e.g. crossing branches for discontinuous constituents), extra means like `xlink` attributes have to be employed (König and Lezius, 2000). Such representational means are less perspicuous and harder to interpret than the straightforward representation via XML embedding.

2.2 Integration of multiple annotations

Whereas these data models and formats might in principle host multi-level, heterogeneous annotation, projects that actually deal with data annotated at more than two levels (like MULI (Baumann et al., 2004)) tend to develop task-specific formats. Only recently, researchers started to integrate and merge annotations from different sources into one format: (Witt et al., 2005) merge multiple XML annotations of the same primary data into one XML format, leaving the original annotations intact as far as possible. For the representation of structurally-conflicting markup, elements have to be broken up and transformed into milestones. In contrast, (Ide and Suderman, 2007) propose one common pivot standard format, "GrAF", which all annotations have to be mapped onto. The format makes use of generic XML element names such as `node` and `edge` and encodes feature-value annotations by generic XML attributes `name` (e.g. "cat") and `value` (e.g. "NN").

In our approach, we pursue the same strategy as (Ide and Suderman, 2007). Our representation format, which we describe in the next section, is quite similar to the GrAF format. It serves as the "neutral" interchange format between different types of annotation structures and, at the same time, as the common import format to the linguistic database ANNIS (see Sct. 4). It supports querying and visualizing the data and its multi-level annotation, and includes ontology-based query evaluation which allows for searching data annotated with different tagsets. This integrated architecture probably distinguishes our approach from the above-mentioned ones.

2.3 Our representation format

Our representation format PAULA[1] (a German acronym for 'Potsdam interchange format for linguistic annotation') focuses on the integration of different annotation structures. We assume that corpus developers apply specialized annotation tools which are tailored to the specific annotation tasks. For instance, *annotate* (Brants and Plaehn, 2000) is frequently used for syntactic annotations; *Palinka* (Orasan, 2003) or *MMAX2*[2] for discourse-level annotations such as co-reference; *Exmaralda* (Schmidt, 2004) is applied for dialog transcription and various layer-based annotations. For these tools, we provide scripts that map the tool output to our representation format. The scripts are publicly available via the Internet: users can upload their data and

[1] http://www.sfb632.uni-potsdam.de/projects/d1/paula/doc/
[2] http://mmax2.sourceforge.net/

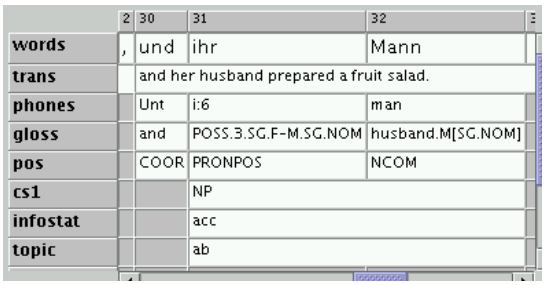

Figure 1: Annotation example (screenshot of the tool *Exmaralda*).

annotations (we currently provide converters for Exmaralda, MMAX2, Tiger XML (König and Lezius, 2000), URML (Reitter and Stede, 2003), Palinka, and a generic importer for annotations using inline-XML markup). The data is converted automatically to PAULA, and the user can copy it to the database ANNIS or perform statistical analyses with our WEKA-based application, see Sct. 4.3).

The mappings from the tool outputs to our format are defined such that they only transfer the annotations from one format into another without *interpreting* them or adding any kinds of information.

As an example, consider the original annotation of a short text fragment, annotated with the tool *Exmaralda*. Fig. 1 shows selected annotation levels, as displayed by the annotation tool[3]. Exmaralda's XML representation format implements annotation graphs, i.e., the primary data and all annotations refer to a common timeline, marked by timeline items (`tli`), whose IDs serve as anchors for the annotations. Annotations are called `event`s, they are anchored to the timeline via `start`/`end` attributes. The `tier` element specifies the type of annotation (e.g. `pos`), the `event` tags contain the actual annotation (`PRONPOS` for possessive pronoun, `NCOM` for common noun). The following fragment displays the primary data `ihr Mann` ('her husband') and their POS annotations.

```
<tli id="T18"/>
<tli id="T19"/>
<tli id="T44"/>
...
<tier id="TIE1" category="words">
<event start="T18" end="T19">ihr</event>
<event start="T19" end="T44">Mann</event>
<...
<tier id="TIE13" category="pos">
<event start="T18" end="T19">PRONPOS</event>
<event start="T19" end="T44">NCOM</event>
```

The corresponding representation of our pivot format PAULA presents the primary data in a `body` element. It defines `mark`ables for segments that receive annotations. A first layer of `markables` points to text regions in the `body` element, by means of `XPointer` expressions (see the markables with IDs `tok_20/21`). These markables can be thought of as tokens. Another layer of markables is added on top of the token markables (see the markables with IDs `pos_15/16`); they point to the tokens by means of `xlink:href` attributes. The actual annotations "PRONPOS" and "NCOM" are encoded by `feat` elements ("features"), which are anchored to the second layer of markables.

```
<body>... ihr Mann ...</body>
...
<mark id="tok_20" xlink:href="#xpointer(
   string-range(//body,'',97,3))"/>
   <!-- ihr -->
<mark id="tok_21" xlink:href="#xpointer(
   string-range(//body,'',101,4))"/>
   <!-- Mann -->
...
<mark id="pos_15" xlink:href="#tok_20"/>
<mark id="pos_16" xlink:href="#tok_21"/>
...
<feat xlink:href="#pos_15" value="PRONPOS"/>
<feat xlink:href="#pos_16" value="NCOM"/>
...
```

The reason for introducing an extra layer of `markables` is that annotations in Exmaralda can refer to *spans* of token markables. In this case, there are two choices: Either anchor `feat` elements to a sequence of markables, similar to token markables, which are anchored to sequences of characters. Or introduce another layer of markables that are anchored to sequences of other markables, and `feat` elements then refer to this extra layer.

In principle our format could host both alternatives. However, we opted for the second alternative because we aim at rather rigid mapping "rules" so that the resulting pivot representation is as uniform as possible. This facilitates further processing and interpretation of the data.

[3] Layers (from top): the primary-data layer; translation to English; phonetic transcription according to Sampa; morpheme glosses, parts of speech, basic syntactic constituents ("cs1"), and information-structural annotation ("infostat", "topic") according to (Dipper et al., 2007).

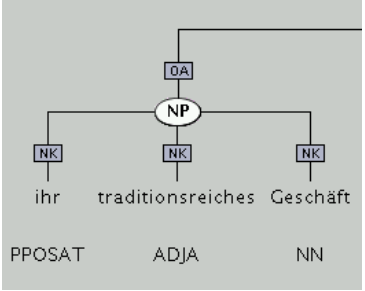

Figure 2: Annotation example (screenshot of the tool *TIGERSearch*).

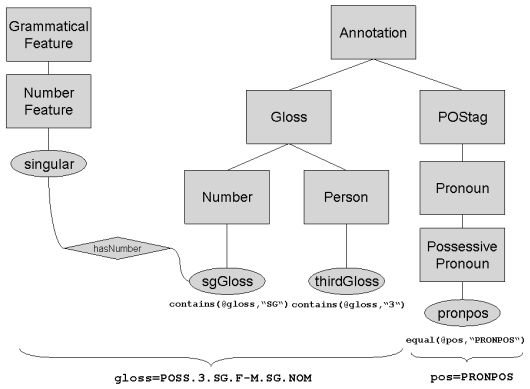

Figure 3: Fragment of Dipper et al.'s (2007) annotation model.

Another annotation example is shown in Fig. 2[4]. The annotations follow the STTS (Schiller et al., 1999) and TIGER (Brants et al., 2004) schemes.

3 An Ontology of Linguistic Annotations

So far, we have described aspects of the technical integration of multi-layered annotations from different sources and their representation. However, the integration of data from different sources (and partially from different languages) not only involves the integration of technical formats but also the conceptual integration. It is well-known that tag identifiers can differ widely and quite often involve idiosyncratic abbreviations. As an example, consider the great variety of tags assigned to *her* as a possessive determiner in different tag sets for English, which at a first glance seem to be fairly arbitrarily chosen at least in parts: `PP$` (Brown, (Greene and Rubin, 1981)), `TB` (London-Lund Corpus, (Eeg-Olofsson, 1991)), `PRP$` (Penn, (Santorini, 1990)), `DD` (POW, (Souter, 1989)), `PRON(poss,sing)` (ICE, (Greenbaum, 1992)), `APPGf` (Susanne, (Sampson, 1995)).

Here, we present a structured, modular ontology that is capable of both the conceptual integration of different annotation schemes by specifying a terminological reference, and the lossless representation of specific annotations.

This structured ontology involves two primary modules, a set of ANNOTATION MODELS which are representations of one annotation scheme, each, and a REFERENCE MODEL which represents a generalization over different annotation models, and thus, a common terminological reference.

A given annotation model is constructed solely on the basis of available annotation documentation, mostly guidelines if available, and annotated examples. Hence, it is a formalization of the annotation documentation, exhaustive with respect to the available documentation, but without any additional interpretation in terms of generally assumed linguistic categories, etc.

The partial ontological representation of the `pos` and `gloss` annotations of *ihr*, the German equivalent to the possessive pronoun 'her' (cf. Fig. 1) in terms of Dipper et al.'s (2007) model is given in Fig. 3. In the same way, annotations of the STTS tagset are represented in a separate annotation model.

While an annotation model is specific to one particular language, community, or purpose, the reference model is a general terminological resource, and consequently based on a broad range of resources, including specific annotation models, grammatical references, textbooks, but also existing terminological references such as the EAGLES recommendations for Morpho-Syntax (Leech and Wilson, 1996), and the GOLD ontology (Farrar and Langendoen, 2003). In case of divergent conceptualizations, e.g. the classification of attributive possessive pronouns

[4]The phrase *ihr traditionsreiches Geschäft* 'her traditional business' is annotated as an NP which functions as an accussative object ("OA"). Terminal nodes are labeled by POS tags according to the STTS tagset: "PPOSAT" (possessive pronoun), "ADJA" (attributive adjective), "NN" (common noun)

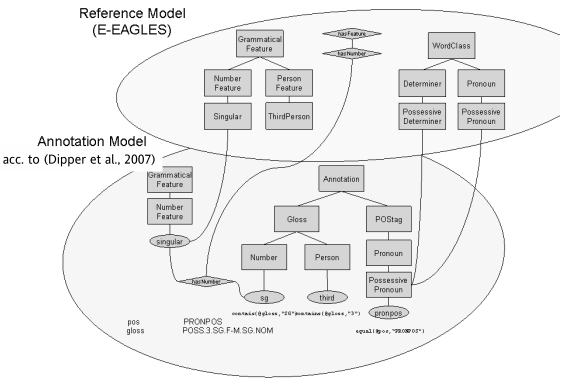

Figure 4: Fragment of E-EAGLES reference model and its linking with Dipper et al.'s (2007) annotation model.

as either Pronouns or Determiners, the EAGLES taxonomy was taken as an orientation, hence, the reference model is also referred to as *E(xtended)-EAGLES ontology*.

Annotation models and the reference model represent self-contained ontologies on their own. The conceptual integration of annotation models is then performed by means of a declarative LINKING between both the reference model and a specific annotation model. In the linking, every concept (class) of the annotation model is assigned a superclass from the reference model – including complex superclasses composed with the set operators $\cup, \cap$, or $\setminus$.

For the annotation model fragment in Fig. 3, the corresponding linking of concepts and the property `hasNumber` with their respective counterparts in the reference model is illustrated in Fig. 4.

In consequence of the linking, the concise annotation of *ihr* ('her', an example from Figures 1 and 2) can be rephrased in terms of the reference model. Consequently, an ontological description such as `PossessivePronoun and hasNumber(Singular) ...` naturally expands (by means of $\subseteq$ and $\in$) into a disjunction of several specific annotations according to different annotation models, e.g. matching both the scheme A tag `PRONPOS` and the scheme B tag `PPOSAT`.

The advantage of this structured account is that it avoids the plain identification of categories from different annotation schemes with standard categories, as it was required in older standardization approaches, e.g. (Leech and Wilson, 1996). Instead, the relations of high complexity can be specified and the necessary *interpretation* of categories in the annotation scheme is represented in an explicit, transparent, and modifiable way.

This tripartite structure of annotation models, reference model, and the linking in between can be augmented by the optional linking of the reference model with additional EXTERNAL REFERENCE MODELs, ontological formalizations of community- or language-specific terminological systems. Currently, we provide a linking with two external reference models, GOLD, the General Ontology of Linguistic Description (Farrar and Langendoen, 2003), developed in the context of language documentation, and the OntoTag ontologies (de Cea et al., 2004) developed in the context of Semantic Web applications, but so far specific to Romance languages.

We claim that this modular approach is more flexible as it allows alternative specifications of linking and the inclusion of alternative upper models as well as additional domain models. In contemporary annotation practice, its technological counterpart is the standoff paradigm (see Section 2).

4 A Database for Multiply Annotated Corpora

Having discussed both technical and conceptual issues of data integration, we now turn to the task of accessing integrated, multi-level corpora.

ANNIS[5] is a linguistic database that can be accessed as a server with standard web browsers via the internet, or installed on a local computer. The "local" ANNIS is a Java servlet application without a database backend: Operations for querying and visualizing are conducted on the data in main memory. This eases installation, but obviously limits the amount of data to be handled. The server version is currently being extended by a relational database. The overall goal of ANNIS is to provide access to heterogeneous multi-level annotations by providing suitable means both for visualization (which we do not address in this paper) and for querying.

[5] http://www.sfb632.uni-potsdam.de/annis/

Target user groups are linguists from different linguistic communities with basic computer skills, to whom developing or adapting existing query and visualization toolkits such as GATE (Cunningham et al., 2002) or the NITE XML Toolkit (Carletta et al., 2003a) would be too advanced or time-consuming. By providing import facilities for the PAULA pivot format described in Section 2, ANNIS supports the idea of distributed annotation with specialized ready-to-use tools. At present, our usage scenarios include the development and analysis of historical corpora, construction of a typological database with data from 16 different languages (Götze et al., 2005), and the creation of a text corpus with rich discourse-related annotations (Stede, 2004).

In the following sections, we focus on the facilities for querying and analyzing cross-layer phenomena that our system provides. The resources we chose for illustration in this paper are listed in Table 1. Corpus A is transcribed speech (maptask dialogues and question–answer pairs); corpus B is newspaper text, partially annotated by two annotators. The annotation layers given here are Information Structure (IS), Part-of-Speech (PoS) and Syntax; tools/formats are given in subscripts.

Corpus A	Corpus B
$\text{IS}_{EXMARaLDA}$	IS_{MMAX}
$\text{PoS, Syntax}_{EXMARaLDA}$	$\text{PoS, Syntax}_{TIGER}$

Table 1: Resources in our Database.

4.1 Annotation-based Querying

The query language implemented with ANNIS builds upon existing query languages and offers typical relations like dominance, inclusion ('_i_'), and overlap. Specifically, the language provides operators both for hierarchical and temporal relations. The latter are of particular relevance for querying multi-level annotations, since time often constitutes the only relation between annotations of different annotation levels. Moreover, the query language allows accessing different annotations of the same corpus, so that, for instance, competing analyses indicating disagreements between annotators can be found, as in (1) wrt. to the givenness of an item:[6]

(1) `ann1::givenness=new & ann2::givenness=giv & #1 _=_ #2`

(2) `aboutness=ref & !givenness=* & #1 _=_ #2`

The negation operator '!' allows us to formulate queries that check for the completeness of annotations. This is illustrated with (2), which checks (across layers) whether all referring expressions are annotated for the feature *givenness*.

4.2 Concept-based Querying

For cases where users are searching for instances of a certain annotation concept (see Section 3) rather than of a concrete tag, we provide for more abstract queries. A query preprocessor retrieves all tag descriptions that correspond to an ontological description and translates them into a disjunction of specific annotation values. If multiple annotation schemes (domain models) are considered, such a description may be expanded into a disjunction of tags from different tagsets and/or tiers.

Ontology-sensitive sub-queries are composed according to the following context-free grammar[7]:

```
ONTOQUERY := {CUE in ONTOEXP}
ONTOEXP   := ONTOCONCEPT |
             (ONTOEXP ONTOOP ONTOEXP) |
             ONTOPROPERTY (ONTOFEATURE)
ONTOOP    := and | or | without
```

Consequently, multiple queries for PoS tags from different annotation schemes can be replaced by one single ontology-sensitive corpus query. The query for possessive pronouns can be abbreviated as in (3).

(3) `pos in {PossessivePronoun}`

As opposed to the choice of regular expressions, this ontology-driven tag expansion allows a user to generalize over the specific form of annotations and tag names; it merely requires conceptual understanding.

[6] The queries (1) and (2) specify constraints over the annotations (givenness=new), their annotation set (ann1), and their relation ('_=_' states that both arguments refer to the same primary data). As (2) shows, wildcards can be used.

[7] ONTOCONCEPT, ONTOPROPERTY and ONTOFEATURE correspond to word classes, properties and grammatical features specified in the reference model. ONTOQUERYs can be embedded in arbitrary code which remains untouched during query expansion.

4.3 Cross-resource cross-layer analysis: Use cases

In addition to interactive queries, ANNIS provides for carrying out a range of statistical analyses.

Hypothesis testing. Suppose we want to investigate how givenness[8] of NPs is linked to their type. We enter a suite of concept-based queries similar to query (4) and receive a contingency table like Table 2.

(4) `givenness=giv & pos in {PossessivePronoun} & #1 _i_ #2`

	giv	acc	new
possNP	5	13	15
dem/defNP	133	91	135
indefNP	151	245	240
name	81	68	163
pers/demPron	109	22	8

Table 2: Contingency table: givenness vs. NP type

As a null hypothesis, we could assume stochastic independence between the features *givenness* and *NP type*. Using Pearson's (1900) χ^2, however, we can discard this hypothesis with high significance ($\chi^2 = 200.51, df = 8, p < .0005$).

Annotation Mining. We also exploit these merged resources to "re-feed" the annotation process by training classifiers on them. For this purpose, we built a component that exports data from the pivot format (see Sct. 2) to the Attribute Relation File Format (ARFF) used in WEKA (Witten and Frank, 2005), a common, ready-to-use data mining environment. As to the export, the user can specify the basic entity to be used, e.g. tokens, noun phrases, etc. Then, these entities are extracted (along with the features they are annotated with), forming one dataset per entity. In WEKA, experiments with different classifiers (SVMs, HMMs, decision trees) can be carried out. Currently, a reimport of the classification results to our pivot format is under construction. Thus, the automatic annotation can be presented to human annotators for correction.

[8] (Dipper et al., 2007) use the values `giv`(en) for previously mentioned discourse referents, `acc`(essible) for referents that can be inferred from the context via 'bridging', `new` for referents new to recipient and discourse; non-referring NPs are not annotated.

5 Summary

We gave an overview of our software environment for producing multi-layer annotated corpora: a pivot format serving as "interlingua" between annotation tools, an ontology-based approach for mapping between tagsets, and a database that integrates the various annotations, and allows for querying the data (either by posing simple queries or by using the ontology) and for statistical analyses. Our conversion tools (to and from the pivot format) and the ANNIS database are freely available for research purposes. At present, we are adding a relational database to the server version of ANNIS. Future work will focus on improving our visualization of query results.

References

S. Baumann, C. Brinckmann, S. Hansen-Schirra, G. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich. 2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *Proc. of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, Boston.

S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

T. Brants and O. Plaehn. 2000. Interactive corpus annotation. In *Proceedings of LREC 2000*, Athens, Greece.

S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003a. The NITE XML Toolkit. *Behavior Research Methods, Instruments, and Computers*, 35(3).

J. Carletta, J. Kilgour, T. O'Donnell, S. Evert, and H. Voormann. 2003b. The NITE Object Model library for handling structured linguistic annotation on multimodal data sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (NLPXML-2003)*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Meeting of the ACL*.

G. Aguado de Cea, A. Gómez-Pérez, I. Álvarez de Mon, and A. Pareja-Lora. 2004. Ontotag's linguistic ontologies: Improving semantic web annotations for a better language understanding in machines. In *ITCC '04: Proc. of the Int'l Conference on Information Technology: Coding and Computing*, Washington, DC, USA. IEEE Computer Society.

S. Dipper, M. Götze, and S. Skopeteas, editors. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, volume 7 of *Interdisciplinary Studies on Information Structure (ISIS)*. Universitätsverlag Potsdam, Potsdam, Germany.

L. Dybkjær, N. Bernsen, H. Dybkjær, D. McKelvie, and A. Mengel. 1998. The MATE markup framework. MATE Deliverable D1.2.

M. Eeg-Olofsson. 1991. *Word-class tagging: Some computational tools*. Ph.D. thesis, Department of Linguistics and Phonetics, University of Lund, Sweden.

S. Farrar and D. T. Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100.

M. Götze, S. Skopeteas, T. Roloff, and R. Stoel. 2005. Towards a cross-linguistic production data archive: Structure and exploration. In Balder ten Cate and Henk Zeevat, editors, *TbiLLC*, volume 4363 of *Lecture Notes in Computer Science*, pages 127–138. Springer.

S. Greenbaum, 1992. *The ICE tagset manual*. University College London.

B. Greene and G. Rubin, 1981. *Automatic grammatical tagging of English*. Department of Linguistics, Brown University, Providence, R.I.

N. Ide and K. Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proc. of The Linguistic Annotation Workshop (LAW) 2007*, Prague.

N. Ide, L. Romary, and E. de la Clergerie. 2003. International Standard for a Linguistic Annotation Framework. In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*.

E. König and W. Lezius. 2000. A description language for syntactically annotated corpora. In *Proc. of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 1056–1060, Saarbrücken.

C. Laprun, J. Fiscus, J. Garofolo, and S. Pajot. 2002. A practical introduction to ATLAS. In *Proceedings of LREC 2002*, Las Palmas,Spain.

G. Leech and A. Wilson, 1996. *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora*. Istituto di Linguistica Computazionale, Pisa.

C. Orasan. 2003. Palinka: a highly customisable tool for discourse annotation. In *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo.

K. Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50(Series 5)*.

D. Reitter and M. Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In *Proc. of the 4th Int'l Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest, Hungary.

G. Sampson. 1995. *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Clarendon, Oxford.

B. Santorini, 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science, University of Pennsylvania. Technical report MS-CIS-90-47.

A. Schiller, S. Teufel, C. Stöckert, and C. Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart and University of Tübingen.

T. Schmidt. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris. ELRA.

C. Souter. 1989. A Short Handbook to the Polytechnic of Wales Corpus. Technical report, ICAME, Norwegian Computing Centre for the Humanities, Bergen University, Norway.

C. M. Sperberg and L. Bernard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Text Encoding Initiative, Chicago, Oxford.

M. Stede. 2004. The Potsdam Commentary Corpus. In *Proc. of the ACL Workshop on Discourse Annotation*, Barcelona.

A. Witt, D. Goecke, F. Sasaki, and H. Lüngen. 2005. Unification of XML Documents with Concurrent markup. *Literary and Linguistic Computing 2005*, 20:103–116.

I. Witten and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufman, San Francisco, 2nd edition.