

Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability

51 Introduction

The creation and use of linguistically annotated corpora for a wide variety of languages has been one of the most prominent developments in computational linguistics over the last fifteen years. While earlier attempts were largely restricted to morpho-syntactic annotation (part-of-speech tagging, morphological analysis, and lemmatization), more recent developments have concentrated on deeper levels of annotation including phrase structure, dependency structure, predicate argument structure, lexical semantics, information structure, and discourse structure. As the complexity of annotation increases and the same data are often annotated with regard to different levels of analyses, questions of quality assurance, reusability and sustainability have become key issues in the creation and maintenance of linguistically annotated corpora. This article surveys the current best practices in each of these three domains, introduces the underlying

research issues, and provides pointers to future research directions in these areas.

35 We will mainly be concerned with textual corpora. Sound recordings of language and other primary data in non-textual modalities are normally transcribed in some way or the other
40 before they are linguistically annotated. We, thus, take the textual representation of language as the starting point of our discussion. For issues of reusability and sustainability
45 of multi-modal corpora, see especially article 33 in this volume.

The remainder of this article is structured as follows: Section 2 reviews the different levels of analyses that
50 can be found in current linguistically annotated corpora and gives examples of reuse of data in the sense of multiple use of the same electronic text resource by different researchers or in different
55 research contexts. 'Reuse of text' in the sense of plagiarism is discussed in article 60. Section 3 focuses on the relationship of data reuse and annotation quality as well as
60 transparency of annotation. Section 4 considers corpora with distinct layers (or tiers) of annotation and focuses on issues of alignment and markup. Section 5 is concerned with metadata,
65 documentation and standardization

efforts. Section 6, finally, presents an integrated architecture for the sustainable representation of corpus data.

70

2 Corpus Resources and Reusability

The first linguistically annotated corpora that went beyond pure morpho-syntactic annotation were concerned with syntactic annotation in the form of phrase structure trees for English. One of the first efforts of this kind was the Gothenburg Corpus (Ellegård 1978), which was a hand-parsed section of the one million token Brown Corpus of American English (Francis/Kučera 1979). The Gothenburg Corpus comprised approximately 130,000 words and is a small resource by current standards. Its annotation was later reworked according to a very rich and detailed annotation scheme and then published as the SUSANNE Corpus (Sampson 1995) comprising also a small sample of speech from the London-Lund Corpus (Svartvik 1990). An early project involving automatic annotation and manual post-editing was the Lancaster Parsed Corpus (Garside/Leech/Váradi 1992, see also article 17). It is based on 140,000 words of the part-of-speech tagged Lancaster-Oslo/Bergen (LOB) Corpus of British English (Garside/Leech/Sampson 1987). Subsequently, the Penn Treebank

100project was launched at the University
of Pennsylvania, which resulted in the
creation of the Penn Treebank for
American English. The first preliminary
release of the treebank in 1992
105presented a corpus of more than 2.8
million tokens of skeleton-parsed text
including, among others, a one million
token subcorpus of Dow Jones Newswire
articles and the one million token Brown
110Corpus, which was parsed and completely
retagged using the Penn Treebank tagset
(Marcus/Santorini/Marcinkiewicz 1993).
In the following releases, the Dow Jones
primary data was slightly changed and
115has since been referred to as the *Wall
Street Journal Corpus*. It is fair to say
that the Penn Treebank has served as a
model of best practice for the creation
of treebanks for many other languages.
120We refer interested readers to article
17 for more detailed information.

This short survey of the first English
treebanks shows that the same primary
data have found their way into different
125annotation projects. For example, part
of the carefully sampled and balanced
Brown Corpus was syntactically annotated
in the Gothenburg Corpus as well as in
the SUSANNE Corpus. The Brown Corpus as
130a whole was first part-of-speech tagged
with the Brown tagset (Francis/Kučera
1982); later, it was retagged and parsed
as part of the Penn Treebank. Another

major textual resource of the Penn
135Treebank, the one million token Wall
Street Journal section, was originally
taken from the TIPSTER Information-
Retrieval Text Research Collection,
which comprises, among other texts, a
140three-year Wall Street Journal
collection.

The reuse of resources is not limited
to the primary data. The Wall Street
Journal section of the Penn Treebank is
145the most prominent example of how an
already annotated resource is reused for
further annotation. It has not only been
used as a reference source for a wide
variety of natural language processing
150applications but has also served as a
basis for further linguistic annotation
in the areas of semantics and discourse
analysis. The annotation of the second
release of the Penn Treebank (Penn
155Treebank II, Marcus et al. 1994),
comprises information on phrase
structure, basic predicate argument
structure, and some semantic
distinctions of adjuncts. Figure 1
160exemplifies a sentence in the Penn
Treebank II bracketing format, which is
reminiscent of the programming language
LISP. The locative adjunct is marked
with the label PP-LOC and the subject
165with NP-SBJ.

{insert figure 1 here}

The PropBank project created
170 Proposition Bank I
(Palmer/Gildea/Kingsbury 2005), which is
based on the Wall Street Journal section
of the Penn Treebank II. PropBank
annotates predicate-argument relations
175 in the sense of assigning coarse-grained
word senses to the predicates and
prototypical semantic roles to the
arguments. The annotation process made
use of the treebank's phrase structural
180 information: a rule-based argument
tagger preprocessed the corpus, which
was then manually post-edited. In some
cases, the PropBank argument structure
disagrees with the Penn Treebank
185 syntactic structure
(Palmer/Gildea/Kingsbury 2005: 81ff). In
Figure 2, the subject is marked with the
prototypical agent role ARG0, the direct
object with the prototypical theme role
190 ARG1, and the locative adjunct-like
argument with ARGM-LOC.

{insert figure 2 here}

195 The NomBank project (Meyers et al.
2004) again aims at creating a
supplementary annotation to the Wall
Street Journal section of the Penn
Treebank. In NomBank 1.0, nominal
200 predicates are marked, and their
arguments are annotated with

prototypical semantic roles in accordance with the PropBank annotation.

In the Penn Discourse TreeBank project
205(Miltsakaki et al. 2004), the Wall
Street Journal Corpus is annotated
according to a theory of low-level
discourse structure. It treats discourse
connectives, sometimes also referred to
210as *discourse markers*, such as *because*,
when, *but*, or *as a result* as a kind of
predicate between two arguments. In
addition to the explicit use of such
connectives in corpora, implicit
215relationships are also annotated by
invoking the same connectives. Instead
of building on the syntactic annotation
of the Penn Treebank, the project
decided to use the raw tokenized text
220for their annotation to avoid errors in
the Penn Treebank, and allow for cases
where discourse arguments do not align
with syntactic structures (Dinesh et al.
2005). The CCGbank (Hockenmaier/Steedman
2252005) is the result of an automatic
conversion of the whole Penn Treebank
into a corpus of Combinatory Categorical
Grammar derivations. It pairs syntactic
derivations with sets of word-to-word
230dependencies, which approximate the
underlying predicate-argument structure.
There are various other projects which
added annotation levels to parts of the
Wall Street Journal Corpus only. For
235example, the TimeBank Corpus

(Pustejovsky et al. 2003) comprises texts from various sources, among them articles from the Wall Street Journal Corpus, which are annotated with event
240 classes, temporal information, and aspectual information according to the specifications of TimeML (Pustejovsky et al. 2004). The Rhetorical Structure Discourse Treebank
245 (Carlson/Marcu/Okurowski 2003) contains, among other data, a selection of 385 Wall Street Journal articles from the Penn Treebank, which were annotated with discourse structure in the framework of
250 Rhetorical Structure Theory (Mann/Thompson 1988). Clauses and larger sequences are hierarchically ordered by a set of discourse relations such as background, elaboration and contrast.
255 For the message understanding competition MUC 6 (1995), a subcorpus of 318 Wall Street Journal articles was published, annotated with anaphora and coreference information. The PARC 700
260 Dependency Bank (King et al. 2003) consists of 700 sentences, which were randomly extracted from one section of the Wall Street Journal Corpus, parsed with an LFG grammar, and given gold-
265 standard annotations of grammatical dependency relations by manual correction and extension. The FrameNet project (Bake/Fillmore/Lowe 1998), which itself does not aim at creating an

270 annotated corpus resource but a lexical
database, provides five texts of the
Wall Street Journal Corpus annotated
with FrameNet semantic roles and word
senses for evaluation of the relation
275 between its own semantic annotation and
the one of PropBank.

There are various examples of multiply
annotated data in languages other than
English. One of the earliest is the
280 300,000 token Swedish Talbanken (Teleman
1974) which was manually annotated with
morpho-syntactic and syntactic
information (see also article 17,
section 1). Recently, Talbanken was
285 revived, its format updated and the
annotations to some extent automatically
re-encoded (see, e.g., Saxena/Borin 2002
and Nilsson/Hall/Nivre 2006). A major
non-English resource is the Prague
290 Dependency Treebank for Czech (Böhmová
et al. 2000), which comprises annotation
of morpho-syntactic information,
surface-oriented dependency structure,
and a non-isomorphic tectogrammatical
295 structure including, among others,
annotation of semantic roles in the
framework of Functional Generative
Description (Sgall et al. 1986), topic
focus articulation, coreference, and
300 information structure. The German TiGer
Treebank (Brants et al. 2004) consists
in its second release of 50,000
sentences which are annotated with

morpho-syntactic information, part-of-
305speech tags, phrase structure, and
functional dependencies. The SALSA
project (Burchardt et al. 2006) enriches
the TiGer Treebank with FrameNet
relations. The TiGer Dependency Bank
310(Forst et al. 2004) is a sample of 2,000
sentences from the treebank, which are
automatically converted and subsequently
extended and corrected in correspondence
to the English PARC 700 Dependency Bank.
315The second major treebank of German,
TüBa-D/Z (Hinrichs et al. 2004),
consists of about 27,000 sentences in
its third release and comprises
information on morpho-syntax, phrase
320structure, topological structure,
functional structure, named entities, as
well as anaphora and coreference
annotation. Another smaller resource is
The Potsdam Commentary Corpus (Stede
3252004). In addition to morpho-syntax,
phrase structure and functional
information, it is augmented with
discourse relations in the framework of
Rhetorical Structure Theory
330(Mann/Thompson 1988), information
structure and anaphoric relations.

These examples show that different
aspects of linguistic description come
into play in linguistically annotated
335corpora. In an ideal world, the
different annotation levels could be
interpreted as distinct analyses of the

same data. In the real world, however,
they are often maintained as separate
340resources which are largely disparate.

This leads to the question of how to
integrate different levels of annotation
into a comprehensive corpus resource. It
would be desirable to use a combined
345representation of all levels of
information, to be able to search a
complex database and to specify
restrictions on all levels of
annotation. In the context of
350sustainability an integrated
representation is desirable too, since
it would allow the specification of
general tools and format conversions
which reduce the risk of losing one
355resource or the other due to obsolete
formats or software. The representation
of different levels of annotation,
however, especially if they are created
in different projects, places great
360demands on the data format.

3 Quality and Consistency of Annotation

Reuse of data is one of the
motivations for creating annotated
365resources in the first place (Leech
1997). The very same corpus data can
then be used and interpreted by
different researchers potentially from
different fields pursuing diverse
370research questions. Linguistic
annotations of corpora can be regarded

as useful resources if they are well-formed and consistent. For the annotation a data format is defined and used, for example, the brackets and the position of the labels in the Penn Bracketing Format. A corpus annotation is well-formed if it conforms to this defined format. General markup languages like XML define well-formedness constraints which can be checked by software tools. Therefore, XML-based linguistic annotation may be carried out by means of general or specialized XML tools (see Dipper/Götze/Stede 2004). Moreover, XML provides means of validating annotations formally according to a document grammar that can, for example, be encoded as a *Document Type Definition (DTD)*. If an XML document conforms to such a grammar, the document is said to be valid with regard to the document grammar. A document grammar might, for example, require that nouns have to be included in noun phrases. Of course, validity constraints are only formal constructs and do not prevent the annotator from annotating incorrect structure due to wrong analyses. The linguistic adequacy can only be determined through human inspection. Inter-annotator agreement and methods of automatic consistency checking, however, may help to find potential problems.

Consistency in annotation is the most important factor in determining the quality of the annotated resource.

Consistency here means that the same linguistic phenomena are annotated in the same way, and similar or related phenomena must receive annotations that represent their similarity or relatedness if possible. Consistency is important for all major applications of annotated corpora, regardless of whether they are used as training data in NLP applications, as gold standard in the evaluation of NLP applications, or as data for qualitative or quantitative linguistic studies. If one phenomenon receives different annotations in the corpus, then a machine learning algorithm cannot learn the regularities concerning the phenomenon, and the evaluation of NLP applications based on inconsistent data is misleading. Even if applications treated this phenomenon consistently, the evaluation would punish the system in cases where the system is consistent, and the annotation is inconsistent. In the case of corpus-based linguistic studies, the linguist is misled by the annotation, either finding only a part of the occurrences of a phenomenon in a quantitative study, or being forced to assume two different phenomena where only a single one exists.

440 Depending on the techniques used for
the creation of the corpora, different
strategies for providing consistency can
be applied:

(i) Annotation guidelines

445 (ii) Semi-automatic annotation

(iii) Manual or automatic consistency
checking

(iv) Multiple annotation by different
annotators

450 These strategies can be applied
independently or in combination, most of
them are independent of the type of
annotation. However, some of them
necessitate the adaptation of the method
455 to the annotation scheme.

Annotation guidelines are crucial for
manual annotation. They describe the
general principles in the design of the
annotation scheme as well as given
460 examples of different phenomena, and
tests for difficult cases. They
constitute a set of evolving laws of
good annotation practice rather than a
comprehensive grammar (the importance of
465 annotation guidelines is also stressed
in article 17, section 3.1). These
guidelines provide a list of all symbols
used in the annotation such as terminal
and non-terminal symbols and their basic
470 definitions. The annotation guidelines
provide a resource for the user of the
corpus: phenomena can only be searched
for in corpora if users know how they

are annotated. For example, if linguists
475 search for relative pronouns in the Penn
Treebank, they need to know that
relative pronouns are annotated as WDT.
If they search for subordinating
conjunctions, they should be aware that
480 these received the same part-of-speech
tag, IN, as prepositions in the Penn
Treebank. Additionally, annotation
guidelines help in training new
annotators and as a reference for
485 annotators when they are unclear on how
to annotate certain phenomena. A
detailed set of annotation guidelines
can help preventing different annotators
from making different decisions
490 concerning the same phenomenon. Examples
of annotation guidelines are the Penn
Treebank part-of-speech tagging
guidelines (Santorini 1990), the Penn
Treebank II bracketing guidelines (Bies
495 et al. 1995), the PropBank annotation
guidelines (Babko-Malaya 2005), and the
TimeBank 1.2 documentation (Pustejovsky
et al. 2006). Dipper/Götze/Skopeteas
(2007) exemplify a joint effort of a
500 number of annotation projects to create
common guidelines for phonology,
morphology, syntax, semantics, and
information structure as realised in the
Potsdam Commentary Corpus.
505 Another possibility of ensuring
consistency is the use of software that
assists the annotator in the annotation

process. *Semi-automatic annotation* is a process in which a program (1) suggests
510 annotations, which then have to be approved or corrected by the annotator, and/or (2) visualises the annotation so that missing links in the annotation become evident. The first type of
515 program generally uses a machine learning component that is trained on a previously annotated data set. This ensures that suggestions are made for phenomena that are consistent with
520 previous annotations. Thus, to create new annotations, a conscious effort on the part of the annotator is required. Examples for such annotation programs are *annotate* (Plaehn 1998) and
525 *TreeBanker* (Carter 1997). The second type of program presents the annotation in such a way that it helps users see gaps in the annotation. Such programs can be useful in the annotation of
530 anaphoric or coreference chains (see article 28). If a link between two coreferent entities is missing, the intended single chain is interrupted, resulting in two different discourse
535 entities. Tools that help with this type of annotation are, for example, *MMA2* (Müller/Strube 2003), *PALinkA* (Orasan 2003), or *WordFreak* (cf.
<https://sourceforge.net/projects/wordfreak/>).
540 ak/).

Automatic consistency checking is a

very general label for annotation-specific strategies to discover inconsistencies. The strategies are
545dependent on the type of annotation as well as on the annotation scheme. They are based on the assumption that humans will always make mistakes, no matter how careful the annotators are. Thus, it is
550a good strategy to employ global search strategies, which can find questionable annotations (see also *transverse correction* in article 17). In treebank annotation, for example, one can check
555whether there are clauses in the treebank that have more than one subject. If such examples are found, they need to be checked manually because, in some cases, the double
560subject may result from coordination rather than from an annotation error. Since these searches are highly dependent on the type of annotation and the annotation scheme, it is difficult
565to envisage a general tool. Thus, the searches are either implemented as specialized programs or as queries in a tool that is capable of searching the annotated structures. Returning to the
570example with the two subjects in a clause, the tools TIGERSearch (König/Lezius 2000) or `tgrep` (also `tgrep2` (Rohde 2001) or `tregex` (Levy/Andrew 2006)) query tree
575structures, so that they can find

clauses with two subjects. A more general approach is to perform a statistical analysis to detect rare constructions, which then need to be
580checked by humans. Such an approach is based on the assumption that very rare constructions are likely to be errors (cf. e.g., Dickinson/Meurers 2005).

The most time-consuming strategy for
585detecting inconsistencies in the annotation is the multiple annotation of the corpus by different annotators or by the same annotator after a sufficient period of time. This means that every
590part of the corpus is annotated at least twice. A comparison of these two annotations reveals annotation errors or problematic cases in which the guidelines provide no guidance or are
595not specific enough to cover the present phenomena. Such multiple annotations allow for the evaluation of *inter-annotator agreement* (also referred to as *inter-coder reliability*) or - if it is a
600single annotator - as *intra-annotator agreement*), i.e., the degree to which the different annotators agree on a single annotation for a specific sentence or paragraph. If a high inter-
605annotator agreement is reached, one can conclude that the corpus has been annotated consistently. Brants (2000), for example, reports 92.43% agreement between two annotators in assigning

610 syntactic annotation to the German NEGRA
Corpus and, after a discussion and
correction phase, an improved agreement
of around 95%.

High inter-annotator agreement also
615 suggests conclusion that the annotation
scheme is well-balanced between
providing enough specialized information
and being too specific. If the
annotation scheme is too specific, it
620 becomes difficult for the annotators to
distinguish the relevant cases, and the
annotation becomes inconsistent. One
example of such a situation is the
annotation of a text with word senses.

625 Most of these annotations are based on
the inventory of WordNet (Fellbaum 1998)
or related wordnets for other languages.
If WordNet provides too few senses for a
word, then certain distinctions are
630 lost, and the annotator needs to decide
which of the existing categories fits
the word or whether to use a
superordinate, less specific category.

However, if WordNet provides a very
635 fine-grained set of senses, then it is
often difficult for the annotator to
decide which is the correct sense for
the word in question (see also Véronis
2001 and Palmer/Dang/Fellbaum

640 forthcoming). Thus, finding a good
granularity for an annotation is
important for ensuring a consistent
annotation of the corpus (see also

Bayerl et al. 2003b). Additionally,
645 recent studies show that the granularity
also influences the quality of NLP
applications based on these corpora (cf.
(Kilgarriff/Rosenzweig 2000) for word
sense disambiguation, as well as (Kübler
650 2005) and (Kübler/Hinrichs/Maier 2006)
for parsing). Finally, we would like to
point out that inter-annotator agreement
is not an absolute measure of quality;
there is always the possibility that two
655 annotators just agree by chance. A
widely used means for measuring inter-
annotator agreement is the kappa
statistic (Cohen 1960). It compares the
observed proportion of agreement with
660 the expected proportion of chance
agreement and indicates whether an
inter-annotator agreement is at a
satisfactory level. As a rule of thumb,
a kappa coefficient of less than or
665 equal to 0.67 means that the inter-
annotator agreement is too close to
chance agreement and that one can
therefore not draw any conclusions about
it. If it is between 0.67 and 0.80 it
670 allows tentative conclusions; only a
value of 0.80 and above allows for
definite conclusions about inter-
annotator agreement (Krippendorff 1980).
For discussions on the interpretation of
675 the kappa coefficient see for example
(Carletta 1996), (Di Eugenio/Glass 2004)
and (Krenn/Evert/Zinsmeister 2004);

article 28 in this volume discusses an alternative measure. New developments
680 concerning the evaluation of linguistically annotated data are presented among others at the biennial Linguistic Resources and Evaluation Conference (LREC) as well as in the
685 Journal on Language Resources and Evaluation.

4 Representation of Annotation

This section deals with the question
690 as to how different linguistic levels of annotation are to be technically realized and how they are related to a shared source of primary data. We distinguish conceptual *levels* from
695 technical *layers*: a conceptual level need not correspond to a single technical layer and vice versa (cf. Bayerl et al. 2003a). Different levels, such as the word level (which is a
700 fundamental but still not fully understood level, see the discussion in article 25), the part-of-speech level, and the phrasal level might, for example, be realized by means of one
705 technical layer, as is the case in the Penn Treebank bracketing format (see figure 1).

The type of representation of the annotated corpus is a crucial
710 prerequisite for ensuring its reusability. It is important to use a

data format for which there are tools to access and search the corpus. The issue is complicated by the fact that the
715 standards that are developed by international standardization committees are often not widely accepted, e.g., the base tagset for the transcription of speech of the Text Encoding Initiative
720 (TEI). In most cases, the data format of a specific corpus is chosen to fit the primary application for which it is created. Thus, part-of-speech-tagged corpora, which are mainly used for
725 training statistical part-of-speech taggers, are represented in pure text files, either in a column format, in which each word with its part-of-speech tag is placed on a separate line, or in
730 running text, in which the part-of-speech tag is separated from the word by a special character. Once the annotation becomes more complex, or when there are multiple annotation levels, the issue of
735 representation becomes more difficult. In general, linguistic annotations can belong to one of two conceptually different annotation models: either a sequential model (sometimes also
740 referred to as *graph-based model*) or a hierarchical model (cf. article 33 and article 36).

4.1 Hierarchies and Sequences

745 There are ongoing efforts for defining

a representational standard for corpora in which multiple types of annotation are present, for example, morphological, morpho-syntactic, syntactic, lexical-
750 semantic, information structure, or discourse annotation. Bird/Lieberman (2001) propose a graph-based representation, in which each type of annotation is treated as an independent
755 layer of graph annotation. The graph approach is very flexible for the representation of text-based corpora as well as speech corpora, each necessitating a different interpretation
760 of the fundamental nodes in the graph. If the underlying data type is text, the position in the sequence of characters serves as the reference point for distinct layers. In contrast, if the
765 underlying data type consists of speech data, the time stamp of each token in the utterance will serve as the basis for the nodes. Additionally, annotation graphs are flexible enough to allow for
770 crossing segment boundaries between layers as well as crossing edges inside a single layer. While this approach is very flexible for the representation of different types of corpora and
775 annotations, it is difficult to imagine a general tool that would allow the user to access the whole range of annotations without having an overly complex and cryptic user interface.

780 In contrast to flexible annotation
schemes designed specifically for
multiple layers of annotation, there are
annotation schemes that are developed to
optimally encode a specific single level
785of annotation. One example of such an
approach is the so-called *Pie-in-the-
Sky* initiative (Meyers 2005) which aims
at optimizing the representation of
semantic information and which should
790serve as a basis for a general standard
in corpus annotation. Hence, semantic
information is the main level of
organization and the other types of
information need to conform to this
795primary annotation level. Because of the
restrictions imposed by the underlying
organization of annotations, it would
be, e.g., impossible to cross a boundary
that is imposed by the semantic
800annotation. While this restriction may
be seen as a disadvantage for
representing information other than on
the semantic level, one needs to keep in
mind that such a representation allows
805for a very simple and direct access to
the semantic data. Thus, it is, for
example, much easier to search for a
specific type of predicate-argument
structure than it would be in a graph-
810based representation. A similar
representation is suggested by
(Hinrichs/Kübler/Naumann 2005). Their
representation, however, is based

primarily on syntactic information.

815 Apart from this level of annotation, the authors include morphological and morpho-syntactic as well as anaphoric and coreferential annotations. Again, the decision that the syntactic

820 constituents serve as the basis for the annotation of other levels restricts the annotation especially on the anaphoric and coreference level. However, it also ensures that the two levels are

825 consistent with regard to each other.

Thus, a markable, i.e., a potential referring expression, on the referential level will always correspond to a syntactic constituent. This is in

830 contrast to other annotations which have been performed more independently. In PropBank, for example, some of the semantic roles intentionally conflict with the syntactic information in the

835 underlying Penn Treebank

(Palmer/Gildea/Kingsbury 2005: 81ff).

Consequently, if the user needs to align the semantic information from PropBank with the syntactic information from the

840 Penn Treebank, these mismatches must be resolved somehow.

4.2 Embedded and Standoff XML Markup

Besides conceptual decisions,

845 especially the development or adoption of an appropriate tagset and the decision for a hierarchical or a graph-

based annotation approach, annotated corpora differ in the way they combine annotations and textual resources. Since the creation of large linguistically annotated corpora is an extremely time consuming endeavor, many of these corpora are based on technical decisions made in the 1980s. The resulting physical representation is therefore realized as a record-and-field or column-based format and frequently as a bracketing format, often influenced by the syntax of the programming language LISP. Nowadays linguistic (and other) annotations use the syntax of XML, at least as an interchange format. Often, existing linguistic corpora are converted into an XML-annotated resource as well (see, e.g., TiGer Treebank, Prague Dependency Treebank).

The main reason for the use of XML in linguistic annotation is obvious: XML is the lingua franca for text annotation in general (cf. article 33). Hence, XML is supported by most relevant software products, ranging from text editors, databases, web browsers to libraries for programming languages. Based on experience with the technical development in the last decades, XML was developed as a pure text format without any implicit formatting. This ensures that XML will be accessible in the future, even after it has ceased to play

an important role. An additional reason for the widespread use of XML is its flexibility. As a result of this
885flexibility, the XML annotation in corpora varies tremendously. The most striking difference concerns the connection of markup and annotations. It is possible to embed the markup used to
890annotate portions of text in the text itself or to refer to this text by means of links. The first technique is called *embedded* or *inline* annotation whereas marking by referencing is usually called
895*standoff* annotation (Thompson/McKelvie 1997). Both approaches have advantages and drawbacks. Standoff annotation is more flexible and allows for the distribution of different levels of
900annotation over several independent layers without duplicating the textual resource that is annotated by the different levels. The distribution of annotations of different linguistic
905levels (e.g., syntactic and discourse structure) over separate annotation layers not only leads to better conceptual modeling but also avoids problems which arise due to the fact
910that the XML standard forbids overlapping elements. But since XML was designed with the embedded annotation technique in mind, only a few XML software products allow for the
915processing of standoff annotated corpora

in a way acceptable for non-XML experts. This, and the fact that single annotation layers cannot be interpreted or exchanged separately, since they are
920 dependent on the layers they themselves point to, goes against the vision of sustainability of annotated text (see Hilbert/Schonefeld/Witt 2005). To reach more sustainable annotations, also
925 standoff annotated documents should be stored and exchanged as XML documents with embedded XML annotations. Ideally each annotation level is encoded in a single XML file. Several of these XML
930 documents can be merged into a single document if they share their textual primary data (see Witt et al. 2005). Alternatively, a parser which is described by Ide/Suderman (2006) can be
935 used to yield a single XML inline representation. This parser integrates annotations distributed in separate XML standoff layers. Both approaches to merging can deal with overlapping
940 hierarchies (cf. articles 33 and 36), which can be represented by means of milestones (DeRose 2004). Such tools enable users, for example, to merge relevant parts of the Penn Treebank with
945 the annotations of PropBank, NomBank, Timebank, and the Penn Discourse Treebank.

5 Documentation

950 It should be clear from the previous
sections that the creation of large
linguistically annotated corpora is
extremely laborious. Of course, this
holds also for smaller corpora, but the
955efforts and the procedures necessary for
the creation of corpora do not simply
scale linearly with the corpus size.

Typically, the starting point for the
creation of a smaller corpus is a single
960linguistic project that is concerned
with a particular research question. To
provide empirical evidence, language
material is collected, and a corpus is
created. This often results in small,
965special-purpose corpora that are barely
reusable in projects concerned with
slightly different research tasks. Even
though this is a deplorable situation
because the numerous small corpora would
970constitute a valuable resource for
linguistics, it is almost impossible to
change the situation. Typically
individual researchers start collecting
language material; afterwards they
975create a corpus, and later analyze or
interpret their data. A single person or
even a small group is likely to try to
minimize the effort which is invested in
the creation of a corpus. However, for
980the creation of large corpora it is
imperative to devote considerable effort
to building the prerequisites of corpus
reuse and distribution. Ideally, such

issues are considered in the early
 985stages of corpus building, otherwise
 there is a risk of wasting serious
 amounts of time and money on the
 creation of annotated data that end up
 as *data graveyards* (Schmidt et al. 2006)
 990that are not accessible for the research
 community.

One of the most important tasks for
 facilitating reusability is a thorough
 documentation that goes beyond
 995annotation guidelines described in
 section 3, which are exclusively
 directed at the human user. A
 comprehensive documentation that
 addresses all corpus-related tasks and
 1000that can also be explored by machines
 comprises different types of
 information:

-Linguistic tagsets: Linguistic concepts
 are marked with tags such as NN as part-
 1005of-speech tag for normal nouns or PP-LOC
 for locative prepositional phrases in
 the Penn Treebank tagset or ARG0 as
 semantic tag in the PropBank tagset.

-Content Models: XML based-annotation
 1010schemes often use document grammars for
 formally defining constraints on the use
 of tags as content of other tags. To
 update and to process corpora annotated
 according to formal document grammars
 1015(e.g., DTDs), an extensive documentation
 of the grammars is extremely important.

-Metadata: As described thoroughly in

article 13, the use of metadata, i.e., information describing corpora or sub-
1020 corpora, is extremely important for the organization of corpora in general and, especially, for the retrieval of information contained in corpora. The most prominent metadata schemes defined
1025 for linguistic data are the ones defined by the Isle Metadata Initiative (IMDI) and by the Open Language Archive Community (OLAC).

-Linguistic Data Categories/ Linguistic
1030 Ontologies: To ease the interoperability of different linguistic resources some researchers promote the use of linguistic ontologies. Because annotated corpora contain arbitrarily defined tags
1035 to refer to linguistic concepts (e.g., number: dual, case: genitive), an ontology, i.e., a formal representation of the concepts, can be used to associate these tags with general
1040 linguistic concepts. The Data Category Registry (DCR, ISO 12620-1 (2003)) and the General Ontology for Linguistics (GOLD, Farrar/Langendoen 2003, see also article 13) were developed for this
1045 purpose.

Large corpora should be documented on all of these levels. This is not only a prerequisite for the reuse of the corpora. Since large corpora are created
1050 and evaluated by several people over a long period of time, extensive

documentation is necessary for the creation of a consistent linguistic resource of high quality (cf. also 1055section 3).

6 An Architecture for the Sustainable Representation of Corpus Data

Many linguistic projects collect and 1060annotate corpora, it has become more and more apparent that many of the laboriously acquired resources are not useable and sometimes not even accessible after the research project 1065for which they were created has come to an end.

At the time of writing this article, the issue of sustainability of linguistic data is the subject matter of 1070a joint research project undertaken by the Collaborative Research Centers in Tübingen (SFB 441), in Hamburg (SFB 538), and in Potsdam (SFB 632) (the joint project's homepage:

1075<http://www.sfb441.uni-tuebingen.de/c2/index-engl.html>). Each of these centers has independently developed their own annotated corpora. This joint project has the practical 1080goal of transforming these heterogeneous corpus collections into a uniform data representation. At the same time, the project aims at developing methodologies and rules of best practice for new 1085corpus-oriented projects in general (see

also Dipper et al. 2006). Within this project, an architecture for the sustainable representation of corpus data was developed and published in
1090 Wörner et al. (2006). A generalized, i.e., less project-specific, version of this architecture is given in figure 3.

{figure 3 here}

1095

The architecture can be subdivided into an input-oriented and an output-oriented part. When dealing with existing corpora, the input-oriented
1100 component is necessary for unifying heterogeneous corpus formats. This merging process may result in a document that contains all the information of the source document but its representation
1105 belongs to another model, for example, data originally represented in a graph-based model is now represented according to a hierarchical model. In this case, the transformation is an information-
1110 preserving (*lossless*) procedure. In other cases, however, the generalized corpus model does not allow for the inclusion of all the information of the source. In such cases, the merging can
1115 be regarded as a generalization to a least common denominator. Whether a generalization or unification is used depends heavily on the diversity of the input formats. If they belong to

1120 different paradigms, especially if the
merging process needs to combine a
graph-based format and a hierarchy-
oriented format (see section 4), a non-
lossless transformation is more likely
1125 to be defined and implemented. As a
consequence, the result of such a
combination is not re-convertible to the
input format. To circumvent this
drawback, the generalized format needs
1130 to allow for the inclusion of
information whose sole purpose is to
enable back-transformations into the
original format.

A corpus represented in a format that
1135 conforms to the generic corpus model
must contain all the metadata of the
original corpus, but potentially
represented in another metadata scheme
(see article 13).

1140 The output-oriented part of the
proposed architecture for the
sustainable representation of linguistic
data follows the idea that the best way
to improve accessibility is to provide
1145 the same data in as many different
representations as possible. Therefore
the data can be partially or completely
converted into several linguistic and
non-linguistic formats (e.g., TEI, LAF
1150 (Ide/Romary/de la Clergerie 2003), or
XHTML).

For a general discussion of
sustainability we refer the reader to

the seminal paper by Bird/Simons (2003)
1155in which they address the problem of
portability and sustainability of
digitized language data in general with
a special emphasis on recorded spoken
language. They suggest rules of best
1160practice for the creation, storage and
distribution of linguistic resources,
which they specify along the following
seven dimensions: content, format,
discovery, access, citation,
1165preservation, and rights. For example,
they recommend the use of Unicode for
character encoding and XML for
annotation. They strongly recommend
using open, non-proprietary standards
1170for storing language data and
descriptions. They suggest that creators
of corpora document the process for
access as part of the metadata,
including any licenses and charges.
1175Finally, Bird and Simons recommend
making an additional paper print-out of
the data, for this is still the most
sustainable way of preserving
information. An updated version of the
1180rules of best practice is available on
the OLAC pages ([http://www.language-
archives.org](http://www.language-archives.org)). For an exhaustive
discussion on aspects of language
documentation interested readers may
1185also consult (Gippert/Himmelmann/Mosel
2006).

7. Conclusion

In this article, we discussed
1190 linguistically annotated corpora and
described an approach for the
sustainable representation of such data.
The availability of large collections of
electronic texts and the need for
1195 corpora augmented with linguistic
information especially for natural
language engineering purposes has led to
the creation of larger and larger
linguistically annotated corpora. In
1200 addition to these large corpora, an
immense amount of rather small special-
purpose corpora have been annotated as
well. Naturally, the creation of all
these corpora is a laborious process but
1205 the effort involved in the creation of
large corpora is not only greater than
for the creation of smaller resources,
it also requires different strategies:
different annotators are involved,
1210 heterogeneous software might be used,
potentially more levels of information
are annotated, the creation of large
corpora is more time-consuming (in some
cases taking even decades), and so on.
1215 For these reasons, every effort should
be made that such resources are
accessible and reusable after their
creation. Therefore, the linguistic
community should not any longer tolerate
1220 corpus-oriented projects ignoring
aspects of sustainability and agreement

on rules of best practice in corpus creation.

12258 Acknowledgements

We would like to thank Stefanie Dipper, Piklu Gupta, and Georg Rehm for their useful comments on earlier versions of this article.

1230

9 Literature

Babko-Malaya, O. (2005) PropBank Annotation Guidelines.

<http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf>

Baker, C., Fillmore, C. and Lowe, J. (1998) The Berkely FrameNet Project. In *Proceedings of COLING-ACL*. Montréal, Canada.

1240 Bayerl, P., Lungen, H., Goecke, D., Witt, A. and Naber, D. (2003a): Methods for the semantic analysis of document markup. In Roisin, C., Munson, E. and Vanoirbeek, C. (eds) *Proceedings of the ACM Symposium on Document Engineering*.

1250 Bayerl, P., Lungen, H., Gut, U. and Paul, K.I. (2003b) Methodology for reliable schema development and evaluation of manual annotations. In *Workshop Notes for the Workshop on Knowledge Markup and Semantic Annotation*, Second International Conference on Knowledge Capture (K-CAP 2003). Sanibel, Florida.

1255 Bies, A., Ferguson, M., Katz, K. and

- MacIntyre, R. (1995) Bracketing Guidelines for Treebank II Style Penn Treebank Project, University of Pennsylvania,
 1260ftp.cis.upenn.edu/pub/treebank/doc/manual/
 1/.
- Bird, S. and Liberman, M. (2001) A Formal Framework for Linguistic Annotation. *Speech Communication*,
 126533:1,2, 23-60.
- Bird, S. and Simons, G. (2003) Seven Dimensions of Portability for Language Documentation and Description. *Language* 79, 557-582.
- 1270 Böhmová, A., Hajič, J., Hajičová, E., Hladká, B. (2003) The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Abeillé, A. (ed) *Treebanks: Building and Using Parsed Corpora*. Kluwer, Amsterdam, 103-127.
- Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G. and Uszkoreit, H. (2004) TIGER: Linguistic Interpretation
 1280of a German Corpus. In Hinrichs, E.W. and Simov, K. (eds) *Research on Language and Computation 2:4*, Special Issue, 597-620.
- Brants, T. (2000) Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece.

- 1290 Burchardt, A., Erk, K., Frank, A.,
Kowalski, A., Pado, S. and Pinkal, M.
(2006) The SALSA Corpus: a German Corpus
Resource for Lexical Semantics. In
Proceedings of the LREC Workshop on
1295 *Merging and Layering Linguistic
Information*. Genoa, Italy.
- Carletta, J. (1996) Assessing
agreement on classification tasks: The
Kappa statistic. *Computational
1300 Linguistics* 22:2, 249-254.
- Carlson, L., Marcu, D. and Okurowski,
M.E. (2003) Building a Discourse-Tagged
Corpus in the Framework of Rhetorical
Structure Theory. In van Kuppevelt, J.
1305 and Smith, R. (eds) *Current Directions
in Discourse and Dialogue*. Kluwer
Academic Publishers, 85-112.
- Carter, D. (1997) The TreeBanker: a
Tool for Supervised Training of Parsed
1310 Corpora. In *Proceedings of the ACL
ENVGRAM Workshop*. Madrid, Spain.
- Cohen, J. (1960) A Coefficient of
Agreement for Nominal Scales.
*Educational and Psychological
1315 Measurement* 20:1, 37-46.
- DeRose, S. (2004) Markup Overlap: A
Review and a Horse. In *Proceedings of
Extreme Markup Languages 2004*. Montréal,
Canada.
- 1320 Dickinson, M. and Meurers, D. (2005)
Detecting Annotation Errors in Spoken
Language Corpora. In *Proceedings of the
Special Session on Treebanks for Spoken*

- Language and Discourse at the 15th Nordic Conference of Computational Linguistics (NODALIDA-05)*. Joensuu, Finland.
- Di Eugenio, B., Glass, M. (2004) The kappa statistic: a second look.
- 1330 *Computational Linguistics* 30:1, 95-101.
- Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A. and Webber, B. (2005) Attribution and the (Non-) Alignment of Syntactic and Discourse Arguments of Connectives. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, MI.
- Dipper, S., Götze, M. and Skopeteas, S. (eds) (2007) Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure. In *Interdisciplinary Studies on Information Structure (ISIS)*, Working Papers of the SFB 632. Universitätsverlag Potsdam, Germany.
- Dipper, S., Götze, M. and Stede, M. (2004) Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*. Lisbon, Portugal.
- Dipper, S., Hinrichs, E.W., Schmidt, T., Wagner, A. and Witt, A. (2006) Sustainability of Linguistic Resources. In *Proceedings of the LREC Workshop on*

Merging and Layering Linguistic Information. Genoa, Italy.

1360 Ellegård, A. (1978) The syntactic structure of English texts: A computer-based study of four kinds of text in the Brown University Corpus. Göteborg: Gothenburg Studies in English 43.

1365 Farrar, S. and Langendoen, D.T. (2003) A linguistic ontology for the Semantic Web. *GLOT International* 7:3, 97-100.

Fellbaum, C. (ed) (1998) *WordNet: An Electronic Lexical Database*. MIT Press.

1370 Forst, M., Bertomeu, N., Crysmann, B., Fouvry, F., Hansen-Schirra, S., Kordoni V. (2004) Towards a dependency-based gold standard for German parsers - The TiGer Dependency Bank. In *Proceedings*

1375 *of the COLING Workshop on Linguistically Interpreted Corpora (LINC '04)*, Geneva, Switzerland.

Francis, W.N. and Kučera, H. (1979) *Manual of information to accompany a*

1380 *standard corpus of present-day edited American English, for use with digital computers*. Technical report, Department of Linguistics, Brown University.

Francis, W.N. and Kučera, H. (1982)

1385 *Frequency Analysis of English*. Houghton Mifflin, Boston.

Garside, R., Leech, G. and Sampson, G. (eds) (1987) *The computational analysis of English: A corpus-based approach*.

1390 Longman, London.

Garside, R., Leech, G. and Váradi, T.

- (compilers) (1992) Lancaster Parsed Corpus. A machine-readable syntactically analyzed corpus of 144,000 words,
1395 available for distribution through ICAME. Bergen: The Norwegian Computing Centre for the Humanities.
- Gippert, J., Himmelmann, N.P. and Mosel, U. (2006) Essentials of Language
1400 Documentation. DeGruyter, Berlin.
- Hilbert, M., Schonefeld, O. and Witt, A. (2005) Making CONCUR work. Paper given at Extreme Markup Languages, sponsored by IDEAlliance. Montreal,
1405 Canada.
- Hinrichs, E.W., Kübler, S., Naumann, K., Telljohann, H. and Trushkina, J. (2004) Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank. In
1410 *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. Tübingen, Germany, 51-62.
- Hinrichs, E.W., Kübler, S. and Naumann, K. (2005) A Unified
1415 Representation for Morphological, Syntactic, Semantic, and Referential Annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann
1420 Arbor, MI.
- Hockenmaier, J. and Steedman, M. (2005) CCGbank Manual. Technical Report MS-CIS-05-09, Department of Computer and Information Science, University of
1425 Pennsylvania.

- Ide, N., Romary, L. and de la Clergerie, E. (2003) International Standard for a Linguistic Annotation Framework. In *Proceedings of HLT-1430NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*, Edmonton, Canada.
- Ide, N. and Suderman, K. (2006) An Open Linguistic Infrastructure for 1435 American English. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.
- ISO 12620-1 (2003). Terminology and 1440 other language resources. Data categories. Part 1: Specification of data categories and management of a data category registry for language resources.
- 1445 Kilgarriff, A. and Rosenzweig, J. (2000) Framework and Results for English SENSEVAL. *Computers and the Humanities* 34:1/2, 15-48.
- King, T.H., Crouch, R., Riezler, S., 1450 Dalrymple, M. and Kaplan, R. (2003) The PARC 700 Dependency Bank. *Proceedings of the EAACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*. Budapest, Hungary.
- 1455 König, E. and Lezius, W. (2000) A description language for syntactically annotated corpora. In *Proceedings of the COLING Conference*. Saarbrücken, Germany.

- 1460 Krenn, B., Evert, S. and Zinsmeister, H. (2004) Determining Intercoder Agreement for a Collocation Identification Task. In *Proceedings of KONVENS 2004*. Vienna, Austria.
- 1465 Krippendorff, K. (1980) *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills.
- Kübler, S. (2005) How Do Treebank
- 1470 Annotation Schemes Influence Parsing Results? Or How Not to Compare Apples And Oranges. In *Proceedings of the International Conference on Recent Advances in Natural Language*
- 1475 *Processing, RANLP 2005*. Borovets, Bulgaria.
- Kübler, S., Hinrichs, E.W. and Maier, W. (2006) Is it Really that Difficult to Parse German? In *Proceedings of the*
- 1480 *2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006*. Sydney, Australia.
- Leech, G. (1997) Introducing Corpus Annotation. In Garside, R., Leech, G.
- 1485 and McEnery, T. (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London.
- Levy, R. and Andrew, G. (2006) Tregex and Tsurgeon: tools for querying and
- 1490 manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

- Mann, W.C. and Thompson, S.A. (1988).
 1495Rhetorical Structure Theory: Toward a
 functional theory of text organization.
Text 8 (3): 243-281.
- Marcus, M., Santorini, B. and
 Marcinkiewicz, M.A. (1993) Building a
 1500Large Annotated Corpus of English: The
 Penn Treebank. *Computational
 Linguistics* 19:2, 313-330.
- Marcus, M., Kim, G., Marcinkiewicz,
 M.A., MacIntyre, R., Bies, A., Ferguson,
 1505M., Katz, K. and Schasberger, B. (1994)
 The Penn Treebank: Annotating predicate
 argument structure. In *ARPA Human
 Language Technology Workshop*. Plainsbor,
 NJ.
- 1510 Meyers, A. (2005) Pie in the Sky
 Description.
[http://nlp.cs.nyu.edu/meyers/pie-in-the-
 sky/pie-in-the-sky-descript.html](http://nlp.cs.nyu.edu/meyers/pie-in-the-sky/pie-in-the-sky-descript.html).
- Meyers, A., Reeves, R., Macleod, C.,
 1515Szekely, R., Zielinska, V., Young, B.
 and Grishman, R. (2004) Annotating Noun
 Argument Structure for NomBank. In
*Proceedings of the 4th International
 Conference on Language Resources and
 1520Evaluation*. Lisbon, Portugal.
- Miltsakaki, E., Prasad, R., Joshi, A.
 and Webber, B. (2004) The Penn Discourse
 Treebank. In *Proceedings of the 4th
 International Conference on Language
 1525Resources and Evaluation*. Lisbon,
 Portugal.
- MUC 6 (1995) *Proceedings of the 6th*

Message Understanding Conference.

Columbia, MD, 1995.

- 1530 Müller, C. and Strube, M. (2003)
Multi-Level Annotation in MMAX. In
*Proceedings of the 4th SIGdial Workshop
on Discourse and Dialogue.* Sapporo,
Japan.
- 1535 Nelson, T.H. (1997) Embedded markup
considered harmful. *WWW Journal* 2:4,
129-134.
- Nilsson, J., Hall, J. and Nivre, J.
(2006) Mamba meets TIGER: Reconstructing
1540a treebank from antiquity. In
Henrichsen, P.J. and Skadhauge, P.R.
(eds) *Treebanking for Discourse and
Speech.* Copenhagen, Denmark, 119-132.
- Orasan, C. (2003) PALinkA: a highly
1545customizable tool for discourse
annotation. In *Proceedings of the 4th
SIGdial Workshop on Discourse and
Dialog.* Sapporo, Japan.
- Palmer, M., Dang, H. and Fellbaum, C.
1550(forthcoming) Making fine-grained and
coarse-grained sense distinctions.
Natural Language Engineering.
- Palmer, M., Gildea, D., and Kingsbury,
P. (2005) The Proposition Bank: A Corpus
1555Annotated with Semantic Roles.
Computational Linguistics 31:1.
- Plaehn, O. (1998) Annotate
Bedienungsanleitung. Universität des
Saarlandes, Sonderforschungsbereich 378,
1560Projekt C3.
- Pustejovsky, J., Hanks, P., Saurí, R.,

- See, A., Gaizauskas, R., Setzer, A.,
Radev, D., Sundheim, B., Day, D., Ferro,
L. and Lazo, M. (2003) The TIMEBANK
1565Corpus. In *Proceedings of Corpus
Linguistics 2003*, Lancaster, UK.
Pustejovsky, J., Ingria, B., Saurí,
R., Castano, J., Littman, J.,
Gaizauskas, R., Setzer, A., Katz, G. and
1570Mani, I. (2004) The Specification
Language TimeML. In Mani, I.,
Pustejovsky, J. and Gaizauskas, R. (eds)
The Language of Time: A Reader. Oxford
University Press, Oxford.
- 1575 Pustejovsky, J., Littman, J., Saurí,
R., and Verhagen, M. (2006) Timebank 1.2
Documentation. Brandeis University,
[http://www.timeml.org/site/timebank/docu-
mentation-1.2.html](http://www.timeml.org/site/timebank/documentation-1.2.html).
- 1580 Rohde, D. (2001) Tgrep2. Technical
report, Carnegie Mellon University.
<http://tedlab.mit.edu/~dr/Tgrep2>.
Sampson, G. (1995) *English for the
Computer: the SUSANNE corpus and
1585analytic scheme*. Clarendon Press,
Oxford.
Santorini, B. (1990) Part-Of-Speech
Tagging Guidelines for the Penn Treebank
Project, Department of Computer and
1590Information Science, University of
Pennsylvania, 3rd Revision, 2nd
Printing,
[ftp.cis.upenn.edu/pub/treebank/doc/taggu-
ide.ps.gz](ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz).
- 1595 Saxena, A. and Borin, L. (2002)

- Locating and Reusing Sundry NLP Flotsam
in an e-Learning Application. In
*Proceedings of the LREC Workshop on
Customizing knowledge in NLP
1600applications: strategies, issues, and
evaluation*. Las Palmas, Spain, 45-51.
Schmidt, T., Chiarcos, C., Lehmborg,
T., Rehm, G., Witt, A., and Hinrichs,
E.W. (2006) Avoiding Data Graveyards:
1605From Heterogeneous Data Collected in
Multiple Research Projects to
Sustainable Linguistic Resources. In
Proceedings of the E-MELD Workshop 2006.
East Lansing, MI.
- 1610 Sgall, P., Hajičová, E., Panevová, J.
(1986) *The Meaning of the Sentence in
its Semantic and Pragmatic Aspects*.
Reidel, Dordrecht and Academia, Prague.
Stede, M. (2004) The Potsdam
1615Commentary Corpus. In *Proceedings of
the ACL-04 Workshop on Discourse
Annotation*. Barcelona, Spain.
Svartvik, J. (ed) (1990) *The London
Corpus of Spoken English: Description
1620and Research*. Lund Studies in English
82. Lund University Press.
Teleman, U. (1974) Manual för
grammatisk beskrivning av talad och
skrivna svenska. *Lundastudier i nordisk
1625språkvetenskap* Serie C nr 6. Lund,
Sweden.
Thompson, H.S. and McKelvie, D. (1997)
Hyperlink Semantics for Standoff Markup
of Read-only Documents. In *Proceedings*

1630of *SGML Europe'97*. Barcelona, Spain.

Véronis, J. (2001). Sense tagging:
does it make sense? Paper presented at
the Corpus Linguistics 2001 Conference,
Lancaster, U.K.

1635<http://www.up.univ->

[mrs.fr/veronis/pdf/2001-lancaster-
sense.pdf](http://www.up.univ-mrs.fr/veronis/pdf/2001-lancaster-sense.pdf).

Witt, A., Goecke, D., Sasaki, F. and
Lüngen, H. (2005) Unification of XML

1640Documents with Concurrent Markup.

Literary and Linguistic Computing 20:1,
103-116.

Wörner, K., Witt, A., Rehm, G. and
Dipper, S. (2006) Modelling Linguistic

1645Data Structures. In *Proceedings of*

Extreme Markup Languages 2006, Montréal,
Canada.

Authors:

1650 Heike Zinsmeister, Tübingen (Germany)

Erhard Hinrichs, Tübingen (Germany)

Sandra Kübler, Bloomington/IN (U.S.A.)

Andreas Witt, Tübingen (Germany)

With the exception of the first author,

1655the order is alphabetical.

```

( (S
  (PP-LOC (IN In)
    (NP (PDT such) (DT an) (NN environment) ))
  ( , , )
1660 (NP-SBJ (DT a) (NN market) (NN maker) )
      (VP (MD can)
          (VP (VB absorb)
              (NP (JJ huge) (NNS losses) )))
      ( . . ) ))

```

1665

Figure 1: Penn Treebank II: *In such an environment, a market maker can absorb huge losses.*

1670

```

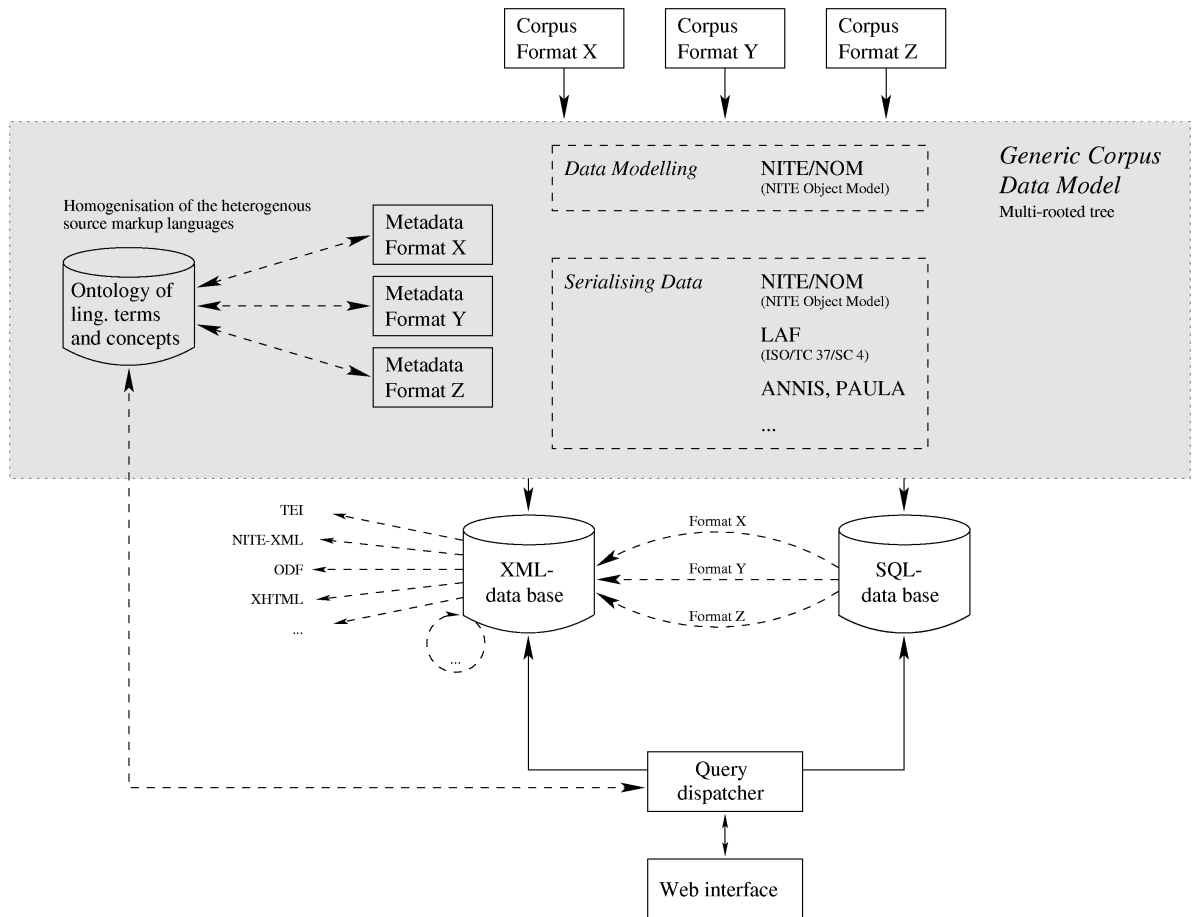
[ARGM-LOC In such an environment] , [ARG0 a
market maker] [ARGM-MOD can]
[rel absorb] [ARG1 huge losses] .

```

1675

Figure 2: PropBank; prototypical semantic roles of verbal arguments

1680



1685 Figure 3: An architecture for sustainable data representation