Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus

Kepa Joseba Rodríguez[♣], Stefanie Dipper[♠], Michael Götze[♠], Massimo Poesio[△], Giuseppe Riccardi[◇], Christian Raymond[◇], Joanna Wisniewska[‡]

[♣]Piedmont Consortium for Information Systems (CSI-Piemonte)
KepaJoseba.Rodriguez@csi.it
[♠]Department of Linguistics. University of Potsdam.
{dipper|goetze}@ling.uni-potsdam.de
[△]Center for Mind/Brain Sciences. University of Trento.
massimo.poesio@unitn.it
[◇]Department of Information and Communication Technology. University of Trento.
{christian.raymond|riccardi}@dit.unitn.it
[‡]Institute of Computer Science. Polish Academy of Science.
jwisniewska@poczta.uw.edu.pl

Abstract

The LUNA corpus is a multi-lingual, multi-domain spoken dialogue corpus currently under development that will be used to develop a robust natural spoken language understanding toolkit for multilingual dialogue services. The LUNA corpus will be annotated at multiple levels to include annotations of syntactic, semantic, and discourse information; specialized annotation tools will be used for the annotation at each of these levels. In order to synchronize these multiple layers of annotation, the PAULA standoff exchange format will be used. In this paper, we present the corpus and its PAULA-based architecture.[1]

1 Introduction

XML standoff markup (Thompson and McKelvie, 1997; Dybkjær et al., 1998) is emerging as the cleanest way to organize multi-level annotations of corpora. In many of the current annotation efforts based on standoff a single multi-purpose tool such as the NITE XML Toolkit (Carletta et al., 2003) or WordFreak (Morton and LaCivita, 2003) is used to annotate as well as maintain all annotation levels (cf. the SAMMIE annotation effort (Kruijff-Korbayová et al., 2006b)).

However, it is often the case that specialized tools are developed to facilitate the annotation of particular levels: examples include tools for segmentation and transcription of the speech signal like PRAAT (Boersma and Weenink, 2005) and TRANSCRIBER (Barras et al., 1998), the SALSA tools for FrameNet-style annotation (Burchardt et al., 2006), and MMAX (Müller and Strube, 2003) for coreference annotation. Even in these cases, however, it may still be useful, or even necessary, to be able to visualize more than one level at once, or to 'knit' together[2] multiple levels to create a file that can be used to train a model for a particular type of annotation. The Linguistic Annotation Framework by (Ide et al., 2003) was proposed as a unifying markup format to be used to synchronize heterogeneous markup formats for such purposes.

In this paper, we discuss how the PAULA representation format, a standoff format inspired by the Linguistic Annotation Framework, is being used to synchronize multiple levels of annotation in the LUNA corpus, a corpus of spoken dialogues in multiple languages and multiple domains that is being created to support the development of robust spoken language understanding models for multilingual dialogue services. The corpus is richly annotated with linguistic information that is considered relevant for research on dialogue, including chunks, named entities, argument structure, coreference, and dialogue acts. We chose to adopt specialized tools for each level: e.g.,

[1]The members of the LUNA project consortium are: Piedmont Consortium for Information Systems (IT), University of Trento (IT), Loquendo SpA (IT), RWTH-Aachen (DE), University of Avignon (FR), France Telecom R&D Division S.A. (FR), Polish-Japanese Institute of Information Technology (PL) and the Institute for Computer Science of the Polish Academy of Sciences (PL), http://www.ist-luna.eu.
This research was performed in the LUNA project funded by the EC, DG Infso, Unit E1 and in the Collaborative Research Center 632 "Information Structure", funded by the German Science Foundation, http://www.sfb632.uni-potsdam.de.

[2]In the sense of the knit tool of the LT-XML suite.

Proceedings of the Linguistic Annotation Workshop, pages 148–155,
Prague, June 2007. ©2007 Association for Computational Linguistics

transcription using TRANSCRIBER, coreference using MMAX, attributes using SEMANTIZER, etc. To synchronize the annotation and allow cross-layer operations, the annotations are mapped to a common representation format, PAULA.

The structure of the paper is as follows. In Section 2, we present the LUNA project and the LUNA corpus with its main annotation levels. In Section 3, we introduce the PAULA exchange format, focusing on the representation of time alignment and dialogue phenomena. Finally we show how PAULA is used in the LUNA corpus and discuss alternative formats.

2 The LUNA project

The aim of the LUNA project is to advance the state of the art in understanding conversational speech in Spoken Dialogue Systems (Gupta et al., 2005), (Bimbot et al., 2006).

Three aspects of Spoken Language Understanding (SLU) are of particular concern in LUNA: generation of semantic concept tags, semantic composition into conceptual structures and context sensitive validation using information provided by the dialogue manager. In order to train and evaluate SLU models, we will create an annotated corpus of spoken dialogues in multiple domains and multiple languages: French, Italian, and Polish.

2.1 The LUNA corpus

The LUNA corpus is currently being collected, with a target to collect 8100 human-machine dialogues and 1000 human-human dialogues in Polish, Italian and French. The dialogues are collected in the following application domains: stock exchange, hotel reservation and tourism inquiries, customer support service/help-desk and public transportation.

2.2 Multilevel annotation

Semantic interpretation involves a number of subtasks, ranging from identifying the meaning of individual words to understanding which objects are being referred to up to recovering the relation between different semantic objects in the utterance and discourse level to, finally, understanding the communicative force of an utterance.

In some annotation efforts–e.g., in the annotation of the French MEDIA Corpus (Bonneau-Maynard and Rosset, 2003)– information about the meaning of semantic chunks, contextual information about coreference, and information about dialogue acts are all kept in a single file. This approach however suffers from a number of problems, including the fact that errors introduced during the annotation at one level may make other levels of annotation unusable as well, and that it is not possible for two annotators to work on different types of annotation for the same file at the same time. Most current annotation efforts, therefore, tend to adopt the 'multilevel' approach pioneered during the development of the MAPTASK corpus and then developed as part of work on the EU-funded MATE project (McKelvie et al., 2001), in which each aspect of interpretation is annotated in a separate **level**, independently maintained. This approach is being followed, for instance, in the ONTONOTES project (Hovy et al., 2006) and the SAMMIE project (Kruijff-Korbayova et al., 2006a).

For the annotation of the LUNA corpus, we decided to follow the multilevel approach as well. That allows us to achieve more granularity in the annotation of each of the levels and to investigate more easily dependencies between features that belong to different levels. Furthermore, we can use different specialized off-the-shelf annotation tools, splitting up the annotation task and thus facilitating consistent annotation.

2.3 Annotation levels

The LUNA corpus will contain different types of information. The first levels are necessary to prepare the corpus for subsequent semantic annotation, and include segmentation of the corpus in dialogue turns, transcription of the speech signal, and syntactic preprocessing with POS-tagging and shallow parsing.

The next level consists of the annotation of domain information using attribute-value pairs. This annotation will be performed on all dialogues in the corpus.

The other levels of the annotation scheme are not mandatory, but at least a part of the dialogues will be annotated in order to investigate contextual aspects of the semantic interpretation. These levels include the predicate structure, the relations between referring expressions, and the annotation of dialogue acts.

2.3.1 Segmentation and transcription of the speech signal

Before transcription and annotation can begin, it is necessary to segment the speech signal into dialogue turns and annotate them with speaker identity and mark where speaker overlap occurs. The goal of this segmentation is to be able to perform a transcription and annotation of the dialogue turns with or without dialogue context. While dialogue context is preferable for semantic annotation, it slows down the annotation process.

The tool we will use for the segmentation and transcription of the speech signal is the open source tool TRANSCRIBER[3] (Barras et al., 1998).

The next step is the transcription of the speech signal, using conventions for the orthographic transcription and for the annotation of non-linguistic acoustic events.

2.3.2 Part Of Speech Tagging and Chunking

The transcribed material will be annotated with POS-tags, morphosyntactic information like agreement features, and segmented based on syntactic constituency.

For the POS-tags and morphosyntactic features, we will follow the recommendations made in EAGLES (EAGLES, 1996), which allows us to have a unified representation format for the corpus, independently of the tools used for each language.

2.3.3 Domain Attribute Annotation

At this level, semantic segments will be annotated following an approach used for the annotation for the French MEDIA dialogue corpus (Bonneau-Maynard and Rosset, 2003).

We specify the domain knowledge in domain ontologies. These are used to build domain-specific dictionaries. Each dictionary contains:

- Concepts corresponding to classes of the ontology and attributes of the annotation.
- Values corresponding to the individuals of the domain.
- Constraints on the admissible values for each concept.

[3] http://trans.sourceforge.net

The concept dictionaries are used to annotate semantic segments with attribute-value pairs. The semantic segments are produced by concatenation of the chunks produced by the shallow parser. A semantic segment is a unit that corresponds unambiguously to a concept of the dictionary.

(1) buongiorno lei [può iscriversi]$_{concept1}$ [agli esami]$_{concept2}$ [oppure]$_{concept3}$ [ottenere delle informazioni]$_{concept4}$ come la posso aiutare[4]

```
<concept1 action:inscription>
<concept2 objectDB:examen>
<concept3 conjunctor:alternative>
<concept4 action:obtain_info>
```

2.3.4 Predicate structure

The annotation of predicate structure facilitates the interpretation of the relation between entities and events occurring in the dialogue.

There are different approaches to annotate predicate structure. Some of them are based upon syntactic structure, with PropBank (Kingsbury and Palmer, 2003) being one of the most relevant, building the annotation upon the syntactic representation of the TreeBank corpus (Marcus et al., 1993). An alternative to syntax-driven approaches is the annotation using semantic roles as in FrameNet (Baker et al., 1998).

For the annotation of predicate structure in the LUNA corpus, we decided to use a FrameNet-like approach, rather than a syntax-based approach:

1. Annotation of dialogue interaction has to deal with disfluencies, non-complete sentences, ungrammaticality, etc., which complicates the use of deep syntactic representations.

2. If we start from a syntactic representation, we have to follow a long way to achieve the semantic interpretation. Syntactic constituents must be mapped to θ-roles, and then to semantic roles. FrameNet offers the possibility of annotating using directly semantic criteria.

[4] Good morning, you can register for the exam or obtain information. How can I help you?

For each domain, we define a set of frames. These frames are defined based on the domain ontology, with the named entities providing the frame elements. For all the frames we introduce the negation as a default frame element.

For the annotation, first of all we annotate the entities with a frame and a frame element.

Then if the target is overtly realized we make a pointer from the frame elements to the target. The next step is putting the frame elements and the target (if overtly realized) in a set.

(2) buongiorno [lei]$_{fe1}$ [può iscriversi]$_{fe2}$ [agli esami]$_{fe3}$ oppure [ottenere delle informazioni]$_{fe4}$ come la posso aiutare

set1 = {id1, id2, id3}
frame: inscription
frame-elements:{student, examen, date}
set2 = {id4}
frame = info-request
frame-elements:{student, addressee, topic}

```
<fe1 frame="inscription"
FE="student" member="set1"
pointer="fe2">
<fe2 frame="inscription"
FE="target" member="set1">
<fe3 frame="inscription"
FE="examen" member="set1"
pointer="fe2">
<fe4 frame="information"
FE="target" member="set2">
```

2.3.5 Coreference / Anaphoric relations

To annotate anaphoric relations we will use an annotation scheme close to the one used in the ARRAU project (Artstein and Poesio, 2006). This scheme has been extensively tested with dialogue corpora and includes instructions for annotating a variety of anaphoric relations, including bridging relations. A further reason is the robustness of the scheme that doesn't require one single interpretation in the annotation.

The first step is the annotation of the information status of the markables with the tags `given` and `new`. If the markables are annotated with `given`, the annotator will select the most recent occurrence of the object and add a pointer to it. If the markable is annotated with `new`, we distinguish between markables that are related to a previously mentioned object (associative reference) or don't have such a relation.

If there are alternative interpretations, which of a list of candidates can be the antecedent, the annotator can annotate the markable as `ambiguous` and add a pointer to each of the possible antecedents.

(3) **Wizard:** buongiorno [lei]$_{cr1}$ [può iscriversi]$_{cr2}$ [agli esami]$_{cr3}$ oppure ottenere [delle informazioni]$_{cr4}$ come la posso aiutare

```
<cr1 inf_status="new" related="no">
<cr2 inf_status="new" related="no">
<cr3 inf_status="new" related="no">
<cr4 inf_status="new" related="no">
```

Caller: [iscrizione]$_{cr5}$ [esami]$_{cr6}$[5]

```
<cr5 inf_status="given"
single_phrase_antecedent="cr2"
ambiguity="unambiguous">
<cr6 inf_status="given"
single_phrase_antecedent="cr3"
ambiguity="unambiguous">
```

2.3.6 Dialogue acts

In order to associate the intentions of the speaker with the propositional content of the utterances, the segmentation of the dialogue turns in utterances is based on the annotation of predicate structure. Each set of frame elements will correspond to an utterance.

Each utterance will be annotated using a multidimensional annotation scheme partially based on the DAMSL scheme (Allen and Core, 1997) and on the proposals of ICSI-MRDA (Dhillon et al., 2004).

We have selected nine dialogue acts from the DAMSL scheme as initial tagset, that can be extended for the different application domains. Each utterance will be annotated with as many tags as applicable.

(4) **Wizard:** [buongiorno]$_{utt1}$ [lei può iscriversi agli esami]$_{utt2}$ oppure [ottenere delle

[5] Register for the exam.

informzaioni]$_{utt3}$ [come la posso aiutare]$_{utt4}$

```
<utt1 d-act="opening/closing">
<utt2 d-act="statement"
link-frame="set1">
<utt3 d-act="statement"
link-frame="set2">
<utt4 d-act="info-request">
```

Caller: [iscrizione esami]$_{utt5}$

```
<utt5 d-act="answer;statement"
link-frame="set3">
```

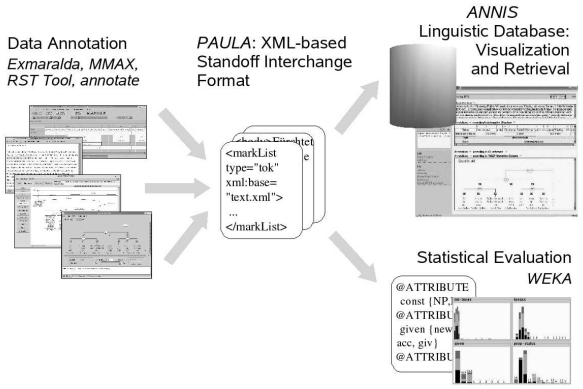

Figure 1: PAULA annotation scenario

3 PAULA - a Linguistic Standoff Exchange Format

PAULA stands for *Potsdamer Austauschformat für linguistische Annotation* ("Potsdam Interchange Format for Linguistic Annotation") and has been developed for the representation of data annotated at multiple layers. The application scenario is sketched in Fig 1: researchers use multiple, specialized off-the-shelf annotation tools, such as EXMARALDA or MMAX, to enrich data with linguistic information. The tools store the data in tool-specific formats and, hence, it is not straightforward to combine information from different sources and, e.g., to search for correlations across multiple annotation layers.

This is where PAULA comes in: PAULA maps the tool-specific formats to a common format and serves as an interchange format between these tools.[6] Moreover, the annotations from the different sources are merged into one single representation. PAULA makes this data available for further applications, such as searching the data by means of the tool ANNIS[7], or to feed statistical applications like WEKA[8].

PAULA is an XML-based standoff format for linguistic annotations, inspired by the "dump format" of the Linguistic Annotation Framework (Ide et al., 2003).[9] With PAULA, not only is the primary data separated from its annotations, but individual annotation layers (such as parts of speech and dialogue acts) are separated from each other as well. The standoff approach allows us to mark overlapping segments in a straightforward way: by distributing annotations over different files (XML as such does not easily account for overlapping segments, since its object model is a hierarchical, tree-like structure). Moreover, new annotation layers can be added easily.

PAULA assumes that a representation of the primary data is stored in a file that optionally specifies a header with meta information, followed by a tag `<body>`, which contains a representation of the primary data. In Fig. 2, the first box displays the transcription, with all contributions from the first speaker coming first, and the contributions from the other speaker(s) following (put in italics in the Figure).

The basic type of "annotation" are *markables*, encoded by the XML element `<mark>`. Markables specify "anchors", i.e., locations or ranges that can be annotated by linguistic information. The locations and ranges are positions or spans in the source text or timeline, which are referenced by means of XLinks and XPointer expressions. For instance, the "Token" markables in Fig. 2 define spans that cor-

[6] Currently, we provide PAULA import filters for the following tools and formats: Exmaralda, MMAX, RST Tool/URML, annotate/TIGER XML. Export from PAULA to the tool formats is at present supported for the original source format only. We plan to support the export of selected annotations to other tools. This is, however, not a trivial task since it may involve loss of information.

[7] ANNIS: `http://www.sfb632.uni-potsdam.de/annis`

[8] WEKA: `http://www.cs.waikato.ac.nz/ml/weka`

[9] The term 'standoff' describes the situation where primary data (e.g., the transcription) and annotations of this data are stored in separate files (Thompson and McKelvie, 1997).

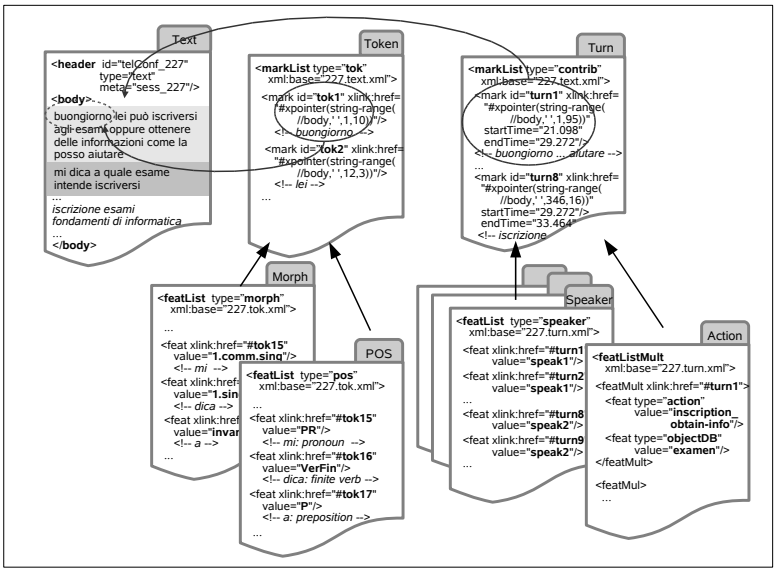

Figure 2: PAULA sample annotation

respond to words. The first markable, with the ID `tok1`, specifies the span that starts at character position 1 and is of length 10: *buongiorno*. Similarly, the speakers' individual turns are specified by the "Turn" markables. Here, the first markable (ID `turn1`) specifies the entire dialogue turn of the first speaker (which corresponds to the part marked in light grey within the text file). Additionally, the markable encodes the time range that is occupied by that turn: it starts at time point 21.098, and ends at time point 29.272.

Markables represent a special kind of annotation: they mark linguistic units. The actual annotation, though, specifies properties of these units, such as part of speech or dialogue acts. For the encoding of these properties, PAULA provides `<feat>` elements, which point to `<mark>` elements by referencing their IDs. Token markables are annotated by "Morph" and "POS" features. The name of the annotated feature is specified by the attribute `type` of the `<featList>` element; the value of the feature is given by the attribute `value` of the `<feat>` elements. For instance, the token with ID `tok15` is annotated with `morph="1.comm.sing"` and `pos="PR"`. Similarly, the Turn markables are specified for the speakers uttering the turns ("Speaker" features), and details of the dialogue acts ("Action") are given. The file with the dialogue act annotations specify multiple features within one tag `<feat>`, rather than distributing the features over several files, as we do in the case of morphology and POS annotations. This way, we explicitely encode the fact that the individual annotations (`action="inscription_obtain-info"` and `objectDB="examen"`) jointly form one complex annotation.

PAULA markables can also refer to points or areas within pictures or videos (by referring to coordinates) or point to other markables (Fig. 2 does not illustrate these options). Moreover, for the encoding of hierarchical structures like graphs, PAULA provides `<struct>` (structure) elements (see Fig. 3 below for an example).

The PAULA standoff format is a generic format that does not necessarily prescribe in detail how to represent annotations. Often there is more than one way to represent the data in PAULA standoff format. In the next section, we present the way we intend to represent dialogue data, which involve possibly overlapping contributions by several speakers, and often include time-alignment information.

4 Representing LUNA Dialogue Annotations in PAULA

In this section, we illustrate the use of PAULA for the LUNA corpus with a more elaborated example, fo-

cusing on the representation of frame annotation. In Fig. 3, the top elements represent the dialogue turns and the semantic units underlying the frame annotations, which are defined on the base of the dialogue turns. "FrameUnit" markables define the scope or extension of the frames, and roughly correspond to a sentence or turn. "FrameP" markables specify the frame participants, i.e., all elements that receive a semantic role within some frame.

The annotations at the bottom contain information about individual frames. The frames are encoded as `<struct>` elements, constituting complex objects that group semantic units to form frames instances. In Fig. 3, the frame with ID `frame_1` consists of the frame unit, the lexical unit and the frame participants. The "FrameAnno" box encodes the name of the frame: "inscription". The frames can be defined by external "Framesets", such as FrameNet (Baker et al., 1998), which in our example is stored in an external XML-resource called `frameSet.xml`.

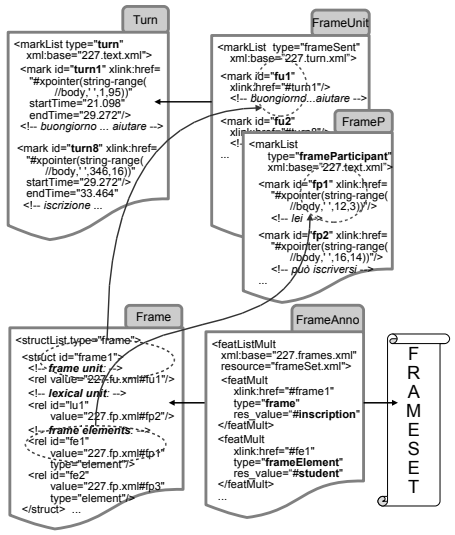

Figure 3: Frame annotation in PAULA

5 Alternative Formats

For richly annotated dialogue corpora, alternative representation formats have been proposed. Two of the most prominent ones are the NITE-XML[10] and the ELAN[11] format. Similar to PAULA, NITE-XML focuses on richly annotated corpus data. It comes with a rich data model and employs a rich meta specification, which determines—based upon the individual corpus characteristics— the concrete linearization of the respective XML representation. Furthermore, it is accompanied by a JAVA API and a query tool, forming a valuable toolkit for corpus engineers who can adapt available resources to their specific needs. The ELAN format is used by a family of tools developed primarily for language documentation, of which the most advanced one is ELAN, a robust, ready-to-use tool for multi-level annotation of video. Its underlying data model is the *Abstract Corpus Model (ACM)* (Brugman and Russel, 2004).

PAULA aims at an application scenario different from both of these formats. First, it builds upon the usage of specialized off-the-shelf annotation tools for the variety of annotation tasks. Both the NITE-XML and ELAN approaches require additional effort and skills from the user, to add the required functionality, which PAULA aims to avoid. Second, PAULA takes care of *merging* the annotations from different sources, which is not in focus of ELAN or NITE.

6 Discussion and Future Directions

We presented the LUNA dialogue corpus and its representation format, the standoff exchange format PAULA.

In contrast to other formats, PAULA focuses on an application scenario in which different annotations come in their own specific format and are to be merged into one corpus representation. This includes, for instance, the use of specialized off-the-shelf annotation tools for specific annotation tasks, as well as distributed and incremental annotation. The creation of the LUNA dialogue corpus is a prototypical example for this scenario.

However, the usefulness of a format also depends on its interoperability and the available tools. With its import filters, PAULA already serves the needs of linguists of different linguistic communities, while more export functionality is still to be integrated. With the export to WEKA, a first step in this direction is done. Furthermore, ANNIS –a web-based tool for visualizing and searching complex multi-level

[10] NITE: `http://http://www.ltg.ed.ac.uk/NITE`

[11] ELAN: `http://www.lat-mpi.eu/tools/elan`

annotations– is available and will be developed further.

In our next steps, we will focus on a deliberate extension of the PAULA format for further and more complex dialogue annotations, which will enable the use of PAULA as an exchange format also in this domain.

References

J. Allen and M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers.

R. Artstein and M. Poesio, 2006. *ARRAU Annotation Manual (TRAINS dialogues)*. Univerity of Essex, U.K.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*. Association for Computational Linguistics.

C. Barras, W. Geoffrois, Z. Wu, and M. Libermann. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proceedings of the First International Conference on Language Ressources and Evaluation (LREC)*.

F. Bimbot, M. Faundez-Zanuy, and R. deMori, editors. 2006. *Special Issue on Spoken Language Understanding*, volume 48 of *Speech Communication*. Elsevier.

P. Boersma and D. Weenink. 2005. Praat: doing phonetics by computer (Version 4.3.14). `http://www.praat.org`.

H. Bonneau-Maynard and S. Rosset. 2003. A semantic representation for spoken dialogues. In *Proceedings of Eurospeech*, Geneva.

H. Brugman and A. Russel. 2004. Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2065–2068, Paris: ELRA.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, and S. Pado. 2006. SALTO – A Versatile Multi-Level Annotation Tool. In *Proceedings of LREC 2006*.

J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003. The NITE XML Toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers – special issue on Measuring Behavior,*, 35(3).

R. Dhillon, S. Bhagat, H. Carvez, and E. Shriberg. 2004. Meeting Recorder Project: Dialog Act Labeling Guide. Technical report, TR-04-002 ICSI.

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

EAGLES. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R.

N. Gupta, G. Tur adn D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Rahim. 2005. The AT&T Spoken Language Understanding System. *IEEE Transactions on Speech and Audio*, PP(99).

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Proc. HLT-NAACL*.

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 Engineeri ng and Architecture of Language Technology*.

P. Kingsbury and M. Palmer. 2003. PropBank: the Next Level of TreeBank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*.

I. Kruijff-Korbayova, C. Gerstenberger, V. Rieser, and J. Schehl. 2006a. The SAMMIE multimodal dialogue corpus meets the NITE XML toolkit. In *Proc. LREC*, Genoa.

I. Kruijff-Korbayová, V. Rieser, J. Schehl, and T. Becker. 2006b. The Sammie Multimodal Dialogue Corpus Meets the Nite XML Toolkit. In *Proceedings of the Fifth Workshop on multi-dimensional Markup in Natural Language Processing, EACL2006*. EACL.

M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English. *Coputational Linguistics*, (19).

D. McKelvie, A. Isard, A. Mengel, M. B. Moeller, M. Grosse, and M. Klein. 2001. The MATE workbench - an annotation tool for XML corpora. *Speech Communication*, 33(1-2):97–112.

T. Morton and J. LaCivita. 2003. WordFreak: an open tool for linguistic annotation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations*.

Ch. Müller and M. Strube. 2003. Multi-Level Annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*.

H. Thompson and D. McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe'97*. `http://www.ltg.ed.ac.uk/~ht/sgmleu97.html`.