1 Motivation

When creating a dialogue corpus with multi-level annotations, the corpus developers must take a number of important decisions, e.g.,

- **Annotation Tools**: Develop new, adapt, or use existing ones?
- **Representation Format**: Develop a new one, adapt, or use an existing one?

Scarce resources favor the use of existing tools and formats, but the required functionality is often not available. Toolkits such as NITE XML Toolkit allow for adapting both tools and annotation formats, but still require programming skills and effort.

We sketch an approach to corpus development, in which specialized off-the-shelf tools are used for creating annotations, which are subsequently merged into one standoff data representation, the **PAULA Interchange Format for Linguistic Annotation**.

Until recently, this approach was implemented for text corpora only. With the **LUNA dialogue corpus**, the first speech corpus is modeled with PAULA.

4 Annotation Levels and Workflow

- **Word transcription / orthographic annotation**
- **Morphosyntactic annotation**: POS and syntactic chunks.

- **Domain attribute level**
  - Attribute-value pairs representation
  - Target of attribute-value specified using domain ontologies

- **Predicate structure**
  - The corpus is annotated using a FrameNet-like approach
  - Based on domain knowledge, we define a set of frames for each domain

- **Coreference**
  - Different kinds of anaphoric relations like:
    - Identity
    - Bridging: exploiting the relations and properties of the domain ontologies
    - Set-element
  - The annotation scheme allows us to have more than one interpretation of the coreference.

- **Dialogue acts**
  - Initial tagset: 9 selected dialogue acts from the DAMSL scheme. Extensible for the different application domains.
  - The utterances are defined based on the predicate structure and annotated in several dimensions.
  - The annotation on this level will be used to build prototypes in the different application domains.

5 PAULA (Potsdam Interchange Format for Linguistic Annotation)

**PAULA annotation scenario:**
- Use of specialized off-the-shelf annotation tools
- Merging of different annotations into one representation
- Further processing with ANNIOS (for visualization and querying) and other tools (e.g. for statistical analysis)

PAULA realizes a **standoff-architecture**, i.e. separate XML files for:
- Text/Tokens; Meta-Information
- Markables (segments); structures (trees); features (annotations)
- Linking by XLink and XPointer expressions

Other features:
- Import (EMMA/DA, TIGER-XML, MAXAM, RTTool) and export functionality (Weka4 etc.)
- Support of various data structures (graphs, trees, pointers) and conflicting hierarchies; reference to externally defined tagsets
- PAULA/inline: an integrated version of the standoff-annotations for efficient querying and further processing

6 Representing Speech and Dialogue

Extending PAULA for dialogue and speech, annotations can additionally refer to points and spans of the timeline. In dialogue annotations, this often requires representation of partial information:

(a) si allora [noisenoise]+tredi di zero ottantasei

(b) allora avrei bisogno dell' [lexfiller] RWS del PC

PAULA: <feat xlink:href="#tok_34" value="noise"/>

PAULA: <mark xlink:href="#xpointer( string-range/body,'',59,0)="/">

7 Related Work

**MEDIA (France)**
- Goal: Evaluation of understanding capabilities of dialogue systems.
- Annotation of words, acoustic events, semantic segments with attribute-value pairs, coreference.
- Main Features: Annotation with only one tool: Semantizer: Representation: all levels together in one file.

**NITE XML Toolkit (Edinburgh)**
- AMI Meeting Corpus, SAMMIE Corpus, etc.
- Main Features: open-source libraries for richly annotated corpora; support for multimodal and dialogue corpora; query support; Java API and media support; Rich Corpus Meta Specification

**ELAN/ACM (MPI Nijmegen)**
- Corpora in the Dobes ("Documentation of Endangered Languages").
- Main Features: audio and video support, import and export facilities, other off-the-shelf tools for data management, metadata management, visualization and querying.

**Links**
- ELAN: http://www.mpi.nl/ELAN/ACM
- ANNIOS: http://www.bci.tuwien.ac.at/anni/software.html
- PAULA: http://www.sfb632.uni-potsdam.de/projects/d1/paula/doc/