



**SIXTH FRAMEWORK PROGRAMME**
**PRIORITY 2**
**Information Society Technology**

# LUNA

**spoken Language UNderstanding in multilinguAl communication systems**

**Project n. 33549**

# D1.3 – Specifications of the Annotation protocol for the Data

| | | | |
|---|---|---|---|
| **Due date of deliverable:** | March 3, 2007 | | |
| **Actual submission date:** | April 13, 2007 | | |
| **Start Date of Project** | September 4, 2006 | **Duration** | 36 months |
| **Organization name of lead contractor for this deliverable** | University of Trento | | |
| | | **Version** | 1.0 |
| **Dissemination level** | [PU] | | |

## Authors and reviewers tables

| Authors | Company Name |
|---|---|
| Kepa Joseba Rodriguez | CSI |
| Giuseppe Riccardi | UT |
| Massimo Poesio | UT |
| Frederic Bechet | UAPV |
| Géraldine Damnati | FT |
| Krzysztof Marasek | PJIIT |
| Ryszard Gubrynowicz | PJIIT |
| Agnieszka Mykowiecka | PAS |
| Joanna Wisniewska | PAS |
| Cosmin Popovici | LOQ |

| Reviewers | Company Name |
|---|---|
| Renato de Mori | UAPV |
| Fabrice Lefevre | UAPV |
| Malgorzata Marciniak | PAS |
| Franck Panaget | FT |
| Stefan Hahn | RWTH |

# Abstract

In the present document we describe a multi-level transcription and annotation protocol for the LUNA multilingual multidomain spoken language corpora. This protocol will be used by the partners of the LUNA project for the transcription and semantic annotation of the human-human and human-machine spoken dialogs collected for the different application domains and languages. The annotated data will be used in workpackages WP2, WP3 and WP4 to train statistical models for the LUNA spoken language systems.

# TABLE OF CONTENTS

# 1 Introduction

In the last decade there have been several research projects supporting Spoken Language Understanding and Dialog annotation with different level of syntactic, semantic and discourse information.
In the early nineties the DARPA ATIS project researchers used manual or semiautomatic annotation of spoken queries in the flight information reservation domain (Levin and Pieraccini 1995, Miller et al. 1994). In the DARPA Communicator Dialog Travel Planning program (Walker et al., 2001) a large corpus of spoken dialogs was collected and annotated for a flight information and travel booking task. The focus of the evaluation was the task completion rate and user satisfaction measures. In the VoiceTone SLU system (Gupta et al., 2005) a corpus of human-machine dialogs was annotated using a list of named entities and semantic labels. In order to minimize the human labeling effort the data that should be annotated was selected using an active learning framework. In the French evaluation program MEDIA (Bonneau-Maynard et al., 2005) a complex semantic model was used for annotation in order to build a hierarchical representation of the natural language query and coreference between dialog semantic segments.
The LUNA SLU systems will exploit the “fat pipe” information flow (see D1.2) from the speech recognition transcription up to the semantic features and the discourse level relations. We then need to have a coherent and multilayer annotation scheme for the spoken dialog corpora. Moreover we need to support multilingual and multidomain contexts of the data collections done across the consortium sites.
In the following chapters we describe the transcription and annotation protocol that the LUNA partners will use for the semantic annotation of the human-human and human-machine dialogs collected for the different application domains and languages. Semantic interpretation involves several aspects, like the meaning of tokens referred to a domain or the relation between different semantic objects in the utterance and discourse level. In order to capture these different aspects we decided to implement a multi-dimensional annotation scheme. The annotation of some levels is mandatory for a corpus depending on the task needs, domain and/or resource availability. The annotation of the other levels is recommended.
The second chapter concerns the **organization of the data**, explaining how annotations on different levels are stored in different files and how the data stored in these files are related. The next three chapters are related to the **preparation of the corpus for the semantic annotation**. Chapter three specifies how the speech signal is segmented before the transcription and annotation process begins. In chapter four we define a **common transcription standard** that will be used for all the languages present in the project. In chapter five we explain how the transcribed dialog turns will be annotated with **part of speech** tags and **morphosyntactic information**, and pre-segmented based on **shallow syntactic information**. Chapter six specifies the **assignment of attribute-value pairs to semantic segments** based on domain ontologies. The annotation of this level is mandatory, as the annotation of the other levels depends on it. The next two chapters concern the annotation of **relations between the previously annotated semantic segments**. The annotation of these levels is recommended. Chapter seven discusses two prominent approaches for the annotation of **predicate structure**, PropBank (Kingsbury and Palmer, 2003), which starts from a syntactic representation, and FrameNet (Fillmore, 1985), that uses frame semantics. After that we present an annotation scheme that is based in frames and domain knowledge. In chapter eight we present a scheme for the annotation of **coreference** and anaphoric relations between the semantic segments. The last chapter, the annotation of **dialog acts**, is more related to the dialog management as

to the semantic interpretation. The annotation in this level will be used to build prototype systems and investigate the context dependent interpretation. The annotation of dialog acts is recommended.

# 2 Organization of the data

The collected data will be annotated using specialized tools for the different levels and sometimes for the different languages (like in the case of annotation of Part of Speech). These tools use different formats to store the annotated data. In order to merge data annotated using different formats we will follow the idea of standoff annotation. The term standoff describes formats where the primary data (i.e. transcription) and different levels of annotation are stored in separated files (Thomson and McKelvie, 1997). We will separate not only the primary data from the annotations, but the individual annotation layers will be separated from each other as well.

The coordination between the different annotations occurs at two different levels.
The words of each dialog are stored in a file (`name_words.xml`) in the same order as they occur in the conversation and marked with an ID. The markables (segments that will be annotated with a tag) for the annotation in the other levels are defined as segments of words. These segments use the IDs of the words in order to define their boundaries.
On the other side there are links between markables of different levels of annotation.

The files corresponding to each of the audio files will be stored in a folder that has the name of the audio file without the extension (i.e. for `B61365.wav` the name of the folder will be `B61365`). The name of the files will be the name of the folder with addition of the name of the level after an underscore as represented in the following table.

| Audio file | Folder | File | Content |
|---|---|---|---|
| `name.wav` | `name` | `name.trs` | Transcription |
| | | `name_words.xml` | Words |
| | | `name_turns.xml` | Dialog turns |
| | | `name_chunks.xml` | Chunks |
| | | `name_attvalue.xml` | Domain attribute-value pairs |
| | | `name_frames.xml` | Predicate structure |
| | | `name_coref.xml` | Coreference |
| | | `name_dialacts.xml` | Dialog acts |

# 3 Segmentation of the speech signal

In order to speed-up the process of annotation we will first segment the dialog in dialog turns and proceed on to transcribe each turn. At a later stage semantic and dialog annotation could take intra-utterance context into account. This turn segmentation will be done manually using the tool Transcriber[1] (Barras et al, 1998), an open source tool that stores the data in standard XML.

During the segmentation Transcriber annotates the time stamps of the beginning and the end of the turn. This annotation will be completed with manually introduced information about the identity and gender of the speaker.

Example of segmentation:

```
<?xml version="1.0" encoding="<UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="project LUNA"
audio_filename="conference-1347" version="1"
version_date="070309">
<Speakers>
<Speaker id="spk1" name="operator" check="no"
type=″male″ dialect="native" accent=""
scope="local"/>
<Speaker id="spk2" name="costumer" check="no"
type=″male″ dialect="native" accent=""
scope="local"/>
</Speakers>
<Episode>
<Section type="report" startTime="0"
endTime="116.8">
……
<Turn speaker="spk1" startTime="33.234"
endTime="36.527">
<Sync time="33.234"/>
</Turn>

<Turn speaker="spk2" startTime="36.527"
endTime="41.647">
<Sync time="36.527"/>
</Turn>

<Turn speaker="spk1" startTime="41.647"
endTime="46.680">
<Sync time="41.647"/>
</Turn>

<Turn speaker="spk2" startTime="46.680"
endTime="50.153">
<Sync time="46.680"/>
</Turn>
```

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

# 4 Transcription

In general, the goal of the transcription is to capture what the speaker said, and not to what the speaker meant. Annotators should transcribe exactly what they hear without any modifications.

For the transcription the annotator will operate by listening to each manually segmented turn. In this phase the annotator transcribes the corpus based only on what she/he hears without being able to access contextual information.

The transcription of the dialogs in LUNA will be done using Transcriber (fig. 1).


Figure 1: Transcriber window

For the transcription of the dialogs in the project the following conventions will be used:

### Capitalization

The use of capitals should follow the standards of the language. The exception is the initial word of a turn. It will be capitalized only if it would be capitalized in the middle of the sentence.
As stated below, capital letters are used also in acronyms and spellings.

### Numbers

Numbers will be spelled out following the standards of each language. If in a particular domain or in a particular language the use of digits would be preferable in order to simplify the work of the annotators, these digits will be converted automatically into the spelled form after the transcription.

### Spelling

Words that are spelled will be transcribed using capital letters separated by spaces and marked with the symbol [spelled].

Example: “No Tarino, comune di Torino, T O R I N O” (*no Tarino, Municipality of Turin, T O R I N O*)

Transcription in Transcriber window:
no Tarino comune di Torino *[pron=SPELLED-]* T O R I N O *[-pron=SPELLED]*

XML internal representation is:
no Tarino comune di Torino

```
<Event desc="SPELLED" type="pronounce" extent="begin"/>
T O R I N O
<Event desc="SPELLED" type="pronounce" extent="end"/>
```

### Acronyms

Acronyms that are pronounced as a single word will be written in capitals without dots or spaces between the letters.

Example: “Buongiorno sono User dall’ARPA di Grugliasco.” (*Good morning my name is User of ARPA Grugliasco*)

Transcription in Transcriber window:
Buongiorno sono User dall’ ARPA di Grugliasco.

Acronyms that are pronounced as a sequence of individual letters will be transcribed in capitals without dots or spaces between the letters and using the symbol [spelled].

Example: “j'ai un un un c'est mis erreur technique alors le code c'est E.R.C.H.X.” (*I have a a a it's written technical error so the code it's E.R.C.H.X*)

Transcription in Transcriber window:
j'ai un un un c'est mis erreur technique alors le code c'est *[pron=SPELLED-]* ERCHX *[-pron=SPELLED]*

XML internal representation is:
j'ai un un un c'est mis erreur technique alors le code c'est

```
<Event desc="SPELLED" type="pronounce" extent="begin"/>
ERCHX
<Event desc="SPELLED" type="pronounce" extent="end"/>
```

In the case of acronyms that are spelled with foreign pronunciation see below.

### Foreign words

If the speaker uses foreign words (usually English words) with foreign pronunciation we transcribe the word following the orthographic standards of the original language and adding a tag that indicates that it is a foreign word and the name of the language.

Example: “un hôtel à Toulouse avec piscine si possible cet hôtel doit avoir wellness service“(*a hotel in Toulouse with swimming pool if possible the hotel must have wellness service*)

Transcription in Transcriber window:
un hôtel à Toulouse avec piscine si possible cet hôtel doit avoir [lang=English-]wellness service[-lang=English]

XML internal representation is:
un hôtel à Toulouse avec piscine si possible cet hôtel doit avoir
```
<Event desc="en" type="language" extent="begin"/>
wellness service
<Event desc="en" type="language" extent="end"/>
```

If an acronym is spelled with foreign pronunciation we combine the tags [Lang=lang] and [spelled].

Example: CD ROM

Transcription in Transcriber window:
[lang=English-][pron=SPELLED-]CD[-pron=SPELLED]ROM[-lang=English]

XML internal representation is:
```
<Event desc="en" type="language" extent="begin"/>
<Event desc="SPELLED" type="pronounce" extent="begin"/>
CD
<Event desc="SPELLED" type="pronounce" extent="end"/>
ROM
<Event desc="en" type="language" extent="end"/>
```

**Punctuation**

The transcription will not include punctuation marks.

**Word truncations**

Truncation is a phenomenon where a speaker starts uttering a word but interrupts the utterance of the word before the end. The annotator will not try to interpret the word, but will transcribe only the actually spoken part of the word. The end of the truncation will be marked with the symbol “**~**”.

Example: *“Legia stadion nie albo szwol szwoleżerów albo rozbrat lub coś tam takiego”*(Legia’s stadium no or szwoleżerów or rozbrat)

Transcription in Transcriber window:
Legia stadion nie albo szwol+*[lex=~]* szwoleżerów albo rozbrat lub coś tam takiego

XML internal representation is:
Legia stadion nie albo
```
szwol<Event desc="~" type="lexical" extent="previous"/>
szwoleżerów albo rozbrat lub coś tam takiego
```

Truncations at the beginning or at the end of an utterance that are due to a bad segmentation of the speech signal are not annotated like truncations that are due to a false start of the speaker himself. In this case annotators are asked to propose a completion of the truncated word in unambiguous cases.

Example: “je voudrais des inform(ations)“ (*I would like inform(ation)*)

Transcription in Transcriber window:
je voudrais des inform[lex=~-]ations[-lex=~]

XML internal representation is:
```
je voudrais des inform<Event desc="~" type="lexical"
extent="begin"/>ations<Event desc="~" type="lexical"
extent="end"/>
```

Example : *“(m)erci au revoir“* (*(th)ank you bye*)

Transcription in Transcriber window:
[lex=~-]m[-lex=~]corefrci au revoir

XML internal representation is:
```
<Event desc="~" type="lexical"
extent="begin"/>m<Event desc="~" type="lexical" extent="end"/>erci
au revoir
```

In ambiguous cases the annotator is not asked to propose the completion of the word.

Example : *“*je voudrais des in…*“* (*I would like in…*)

Transcription in Transcriber window:
je voudrais des in*+[lex=~]*

XML internal representation is:
```
je voudrais des in<Event desc="~" type="lexical"
extent="previous"/>
```

**Mispronounced**

In case of mispronunciation the correct form must be transcribed with an indication that it has been mispronounced.

Example: “je souhaiterais avoir des renseignements sur ma facture” (*I would like to have information on my bill*)

Transcription in Transcriber window:
je souhaiterais avoir des [pron=*-] renseignements [-pron=*] sur ma facture

XML internal representation is:
```
je souhaiterais avoir des
<Event desc="*" type="pronounce" extent="begin"/>
renseignements sur ma facture
<Event desc="*" type="pronounce" extent="end"/>
```

**Unintelligible words**

Words that cannot be recognized will be transcribed with the symbol **

Example: Legia stadion nie albo *unintelligible* albo rozbrat lub coś tam takiego (*Legia's stadium no or *unintelligible* or rozbrat*)

Transcription in Transcriber window:
Legia stadion nie albo [pron=**] albo rozbrat lub coś tam takiego

XML internal representation is:

```
Legia stadion nie albo
<Event desc="**" type="pronounce" extent="instantaneous"/>
 albo rozbrat lub coś tam takiego
```

**Overlapping speech**

The transcription is related only to the speaker that dominates the dialog. If it is not possible to understand the segment of speech, it will be annotated as "no transcribed".

**Pause fillers, hesitations and human noises**

Pause fillers that are not considered to be words (like ah, hmm...) will be not transcribed but they will be replaced by the symbol [fil]. The same symbol will be used for the representation of human articulatory noises like breath, laugh, cough, etc.

Example: "hm, nie" (*hmm no*)

Transcription in Transcriber window:
*[lex=FIL]* nie

XML internal representation is:

```
<Event desc="FIL" type="lexical" extent="instantaneous"/> nie
```

**Noise**

Non human noises will be annotated with the tag [noise].
We distinguish in the annotation when a noise doesn't overlap speech, when the noise overlaps the beginning or the end of a word and when the noise overlaps a segment.

Example where noise doesn't overlaps the speech signal: "je veux acheter des actions L'Oréal" (*I want to buy some L'Oréal shares*)

Transcription in Transcriber window:
je veux acheter [noise] des actions L'Oréal

XML internal representation is:

```
je veux acheter
<Event desc="noise" type="noise" extent="instantaneous"/> des
actions L'Oréal
```

Example with overlapping on the beginning of the word: *“Quando l’ aveva richiesto?”* (When did you request it?)

Transcription in Transcriber window:
quando l’aveva [noise]+richiesto

XML internal representation is:
```
quando l' aveva <Event desc="noise" type="noise" extent="next"/>
richiesto
```

Example with overlapping on the end of the word: “to Legia stadion, czy szwoleżerów, bo to jest ...” (*is this Legia's stadium or szwoleżerów this is ...*)

Transcription in Transcriber window:
to Legia stadion+[noise] czy szwoleżerów bo to jest

XML internal representation is:
```
to Legia stadion
<Event desc="noise" type="noise" extent="previous"/> czy
szwoleżerów bo to jest
```

Example with noise overlapping a segment: “Vous voulez acheter des actions au règlement comptant sur votre compte-titres ordinaire. Quelle quantité ?” (*You want to buy shares with immediate payment on standard portfolio. How many?*)

Transcription in Transcriber window:
[noise-] vous voulez acheter des actions au règlement comptant sur votre compte-titres ordinaire quelle quantité [-noise]

XML internal representation is:
```
<Event desc="noise" type="noise" extent="begin"/>
vous voulez acheter des actions au règlement comptant sur votre
compte-titres ordinaire quelle quantité ?
<Event desc="noise" type="noise" extent="end"/>
```

If desired it is possible for some domain applications to do a more fine grained distinction between different kinds of noise. In this case we recommend the classification of noises using three different categories: continuous noise ([sta]), intermittent noise ([int]) and babble or cocktail party noise ([babble]).

**Silence**

To our experience it is very hard to annotate objectively, agreement is very low on this. Because of this reason we will annotate pauses only longer than 1 sec.

Example where the silence happens inside of the turn: rozumiem, rozumiem, dobra, dobra (I see, I see, good, good)

Transcription in Transcriber window:
rozumiem rozumiem [sil] dobra dobra

XML internal representation is:

```
 rozumiem  rozumiem
<Event desc="sil" type="noise" extent="instantaneous"/>
 dobra dobra
```

# 5 POS/Shallow parsing

The transcribed material will be annotated on the word level with part of speech (POS-tags) and morphosyntactic information like agreement features using tools available for each language.

The POS taggers available for the different languages of the project use different tagsets for the annotation of categories. In order to have a more consistent annotation in this level for the different languages and to be compatible with international standards, we will define for each language a core set of tags based on the recommendations of EAGLES for the morphosyntactic annotation of text corpora (EAGLES, 1996a) and computational lexicons (EAGLES, 1996b).

Based on the POS annotation, the transcribed material will be segmented according to selected syntactic criteria. The goal of this shallow parsing activity is to group the words into basic constituents like nominal phrases, prepositional phrases, and verbal groups. For the selection of tags that will be used to annotate the categorie of each chunk we will use the recommendations for syntactic annotation of corpora of EAGLES (EAGLES, 1996c).

**Example 1:**
Z placu Zamkowego do Wilanowa jedzie autobus sto szesnaście. (*From Zamkowy square to Wilanów goes the bus hundred sixteen [116].*)

**POS-annotation**

```
<words>

<w id='1' word='z' lemma='z' POS='Prep' morph='-' />

<w id='2' word='placu' lemma='plac' POS='Nc' AGR='m3.gen.sg' />

<w id='3' word='Zamkowego' lemma='Zamkowy' POS='ADJp'
morph='masc.gen.sg.pos'>

<w id='4' word='do' lemma='do' POS='Prep' morph='-' />

<w id='5' word='Wilanowa' lemma='Wilanów' POS='Np'
morph='m3.gen.sg' />

<w id='6' word='jedzie' lemma='jechać' POS='VV'
morph='3.sg.pres.imperf' />

<w id='7' word='autobus' lemma='autobus' POS='Nc'
morph='m3.nom.sg' />

<w id='8' word='sto' lemma='sto' POS='NUM' morph='non-masc.nom.pl'
/>

<w id='9' word='szesnaście' lemma='szesnaście' POS='NUM'
morph='non-masc.nom.pl' />
```

```
.....
</words>
```

**Chunking**

```
<chunks>

<chunk id='1' span=”word_2..word_3” cat='NP' />

<chunk id='2' span=”word_5” cat='NP' />

<chunk id='3' span=”word_6” cat='VP' />

<chunk id='4' span=”word_7..word_9” CAT='NP' />
.....

</chunks>
```

**Example 2:**
Non mi funziona il certificato digitale di webapp (*My webapp digital certificate does not work*)

**POS-annotation**

```
<words>
.....

<w id=“word_9” word=”non” lemma=”non” pos=”ADV” morph=”-” />

<w id=“word_10” word=”mi” lemma=”mi” pos=”PPERS”
morph=”1.comm.sing” />

<w id=“word_11” word=”funziona” lemma=”funzionare” pos=”VVFIN”
morph=”3.sing.pres” />

<w id=“word_12” word=”il” lemma=”il” pos=”ART”
morph=”def.masc.sing” />

<w id=“word_13” word=”certificato” lemma=”certificato” pos=”NC”
morph=”masc.sing” />

<w id=“word_14” word=”digitale” lemma=”digitale” pos=”ADJ”
morph=”masc.sg” />

<w id=“word_15” word=”di” lemma=”di” pos=”PREP” morph=”-” />

<w id=“word_16” word=”webapp” lemma=”webapp” pos=”NC”
morph=”femm.sing” />
.....

</words>
```

**Chunking**

```
<chunks>
.....

<chunk id=”chunk_4” span=”word_4..word_6” cat=”NP” />

<chunk id=”chunk_5” span=”word_7..word_8” cat=”PP” />
.....

</chunks>
```

# 6 Domain attribute level

At this level, domain entities will be annotated following an approach used for the French Media dialog corpus (Bonneau-Maynard, S. and Rosset, S., 2003).

A similar kind of semantic dictionary will be used for the annotation of attribute-value pairs but with a simpler method. First, we do not represent context at this level and second, we split the attribute information into other levels of the annotation.

Our approach begins with the specification of the domain knowledge in domain ontologies that will be used to define the semantic dictionaries.

The next step is the segmentation of the utterances in semantic segments following domain and language specific rules.

## 6.1 Building domain ontologies

The knowledge about the domains is described in domain ontologies (example in fig. 2).

As format for the domain ontologies we have decided to use OWL (DL dialect) (Smith et al. 2004). The main reasons to choose this standard are:

1. It has been defined as a standard and recommended by the W3C.
2. Using one of the most important international specification formats brings LUNA closer to other international projects.
3. We can take advantage of already developed work and free resources available for future development.

Some of the classes that will be used in the different ontologies of LUNA are domain independent, i.e. the temporal expressions. An additional advantage of the use of a unified format for all the domains of the project is that working in the same ontological framework will enable the partners to share common parts of their ontology.

Protégé[2] was chosen to help the development of the ontologies. Protégé is an open source ontology editor that allows constructing ontologies in various formats.

## 6.2 Definition of concept dictionaries

We use the domain ontologies for the definition of the concept dictionaries that will be used for the annotation of the semantic segments.

A concept dictionary contains:

- Concepts: that corresponds to the classes of the ontology.
- Values: that corresponds to the individuals of the ontology.
- Constraints on the admissible values for each concept.

[2] http://protege.stanford.edu

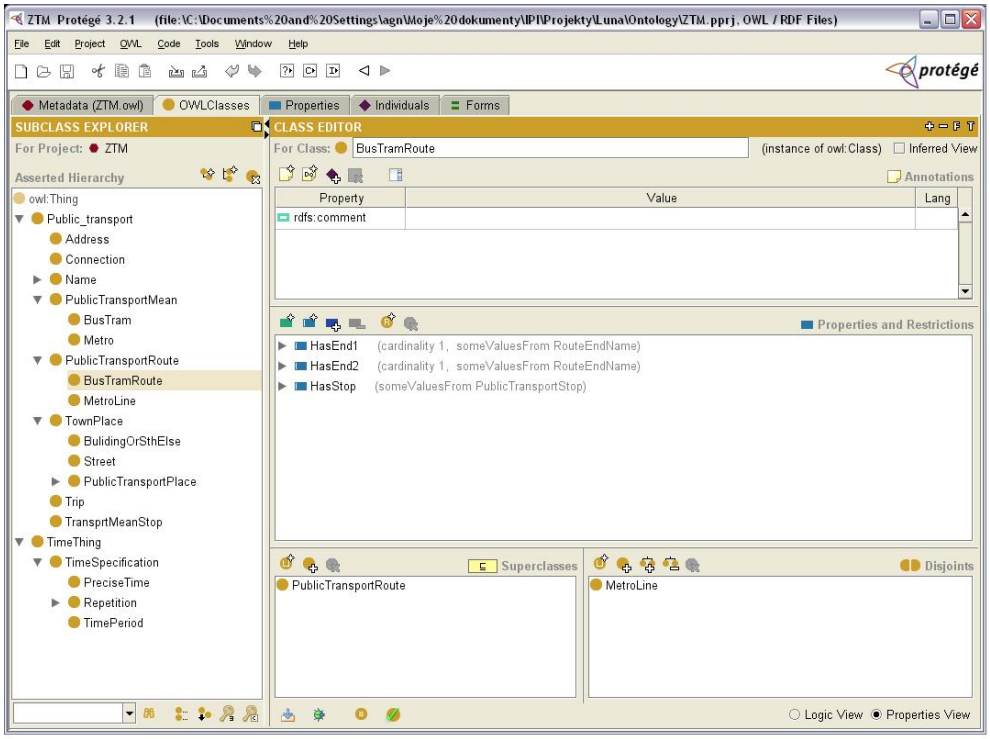


Figure 2: public transportation ontology in Protégé

## 6.3 Segmentation

Semantic annotation is performed on the transcription of the dialogs that have been previously chunked following syntactic criteria.

The first step is to merge these basic syntactic chunks into semantic units. The criteria to concatenate the chunks produced by the shallow parser involve linguistic and domain knowledge. Separate sets of segmentation rules will be defined for each language and domain.

**Example 1:**
“un hôtel à Toulouse dans euh dans le centre oui dans le centre ville” (*An hotel in Toulouse in /disfl/ in the center yes in the center of the town*)

Segmentation provided by the shallow parser:
`[un hôtel] [à Toulouse] [dans] [euh] [dans le centre] [oui] [dans le centre ville]`

Semantic segments:
`[un hôtel] [à Toulouse] [dans euh dans le centre oui dans le centre ville]`

In this example, the annotator uses linguistic knowledge. A PP was broken by disfluencies and the annotator concatenates the chunks that where uttered to be part of the PP.

**Example 2:**
"Da stamattina / non riesco / ad entrare / nell'applicativo sanità" (*Since this morning I cannot access the health application*)

Segmentation provided by the shallow parser:
`[Da stamattina] [non] [riesco] [ad entrare] [nell'applicativo] [sanità]`

Semantic segments:
`[Da stamattina] [non riesco] [ad entrare] [nell'applicativo sanità]`

In this example, the annotator uses domain knowledge. The chunks [non] (*no*) and [riesco] (*can*) are put together in the semantic segment [non riesco], which corresponds in the domain to a type of a problem. The chunks [nell'applicativo] (*in the application*) and [*sanità*] (*health, sanitation*) are put together to give the name of a concrete application.

## 6.4 Annotation

The annotation of a semantic segment is represented as an attribute value pair where:

- The attribute is the name of the concept corresponding to the segment. The possible names of the attributes correspond to the concepts of the semantic dictionary.
- The value of the attribute. The possible values for a attribute correspond to the possible values for a concept in the semantic dictionary.

**Annotation examples**

O: Mi collego / un momento / alla sua macchina
*O: I'm jogging on to your PC for a moment*

```
<concept id="concept_34" span="word_131..word_132"
attribute="action" value="access" />
<concept id="concept_35" span="word_135..word_137"
attribute="hardware" value="pc" />
```

U: ... dalla regione, / assessorato alla sanità/. Da stamattina / non riesco / ad entrare / nell'applicativo sanità.
*U: ... from the Region, Health Department. Since this morning I cannot access the health application.*

```
<concept id="concept_28" span="word_97..word_99"
attribute="customer-region" value="health-department" />
<concept id="concept_29" span="word_100..word_101"
attribute="tempEx-timeRel" value="today-morning" />
<concept id="concept_30" span="word_102..word_103"
attribute="problem-type" value="incident" />
<concept id="concept_31" span="word_104..word_105"
attribute="action" value="access" />
```

```
<concept id=”concept_32” span=”word_106..word_108”
attribute=”software-application” value=”health-application” />
```

U: Z / placu Zamkowego / do / Wilanowa/ jedzie /autobus 116/
*U: From Zamkowy square to Wilanów goes the bus 116*

```
<concept id=”concept_2” span=”word_15..word_16”
attribute=”streetName” value=”plac Zamkowy” />
<concept id=”concept_3” span=”word_18”
attribute=”buildingOrPlaceName” value=”Wilanów” />
<concept id=”concept_4” span=”word_20..word_21”
attribute=”busLinieName” value=”116” />
```

# 7 Predicate structure

The annotation of predicate structure in dialog is useful for the interpretation of the relation between entities and events occurring in the dialog. A further issue is that the annotation of predicate structure can help to implement intent-oriented semantics as described in (Gupta et al., 2005) and also may be used to resolve anaphoric expressions.

Here we discuss two of the most prominent approaches for annotation of argument structure, the PropBank (Kingsbury and Palmer, 2003), that starts from a syntactic representation and FrameNet (Baker et all, 1998) that uses frame semantics (Filmore, 1985).

## 7.1 PropBank

The goal of the annotation of the PropBank (Proposition Bank) is the argument structure of verbs in order to provide training data for supervised automatic role annotation. The scheme was used in order to annotate the TreeBank corpus (Marcus et al 1993), a corpus of written news in American English with syntax annotation.

In the PropBank model, two independent levels are distinguished: the level of arguments and adjuncts and the level of semantic roles.
The elements regarded as arguments are numbered from Arg0 to Arg5 depending on the valence of the verb and on the semantic proximity to the verb. Adjuncts are tagged as ArgM.
Regarding to the semantic roles, PropBank uses two different kinds: general roles like agent, theme, etc. and roles specific to each concrete verb like in the example given below “entity leaving”, “place left”, “attribute”, etc. Frame specific roles are defined in each frame file.

One can map arguments of PropBank with traditional thematic roles like the ones used in VerbNet (Kipper-Schuler, 2005) and syntactic functions as indicated in the table:

| **Argument** | **θ-Rol** | **Gramm. function** |
|---|---|---|
| Arg0 | agent, experiencer | subject |
| Arg1 | patient, theme, attribute, extension | direct object, attribute, predicative, passive subject |
| Arg2 | attributive, beneficiary, instrument, extension, final state | attribute, predicative, indirect object, adverbial complement |
| Arg3 | beneficiary, instrument, attribute, cause | predicative, circumstantial complement |
| Arg4 | destination | adverbial complement |
| ArgM | location, extension, destination, cause, time manner, direction | adverbial complement |

Modal verbs and negations are tagged as ArgMs.
In case of lexical ambiguity (same verb with different senses) the number of arguments is a justification for contrasting senses. This sense distinction is not as fine grained as in other approaches like FrameNet or WordNet (Fellbaum, 1987).

Example of annotation with two different senses of "leave":

**leave.01**

sense:move away from
roles: Arg0: entity leaving
Arg1: place left
Arg2: attribute

The move left the companies as outside bidders.

arg0: The move
rel: left
arg1: the companies
arg2: as outside bidders

**leave.02**

sense:give
roles: Arg0: giver
Arg1: thing given
Arg2: receiver

An ambitious expansion has left Magna with excess capacity

arg0: Am ambitious expansion
rel: left
arg2: Magna
arg1: with excess capacity

The annotation of phenomena like cross-tree dependencies, split arguments, merged arguments and symmetrical arguments is taken into account.

## 7.2 FrameNet

FrameNet is a corpus-based lexicography project based on frame semantics (Fillmore, 1985). This kind of semantics characterizes the semantic and syntactic properties of predicating words by their relation to semantic frames. Frames are a representation of situations involving different elements with different roles, the frame elements (henceforth FE).

**Semantic relations between frames**
In FrameNet frames are organized in a network expressing the following semantic relations:

1 Inheritance: equivalent to is-a in many ontologies. Frames can be modeled by an inheritance lattice with parent-children relation where the children inherit the

semantic properties of the parents. This includes FEs, their relationships and semantic types and frame relations to other frames.

Other relations between frames are:

2 Perspective_on
3 SubFrames
4 Precedes
5 Causative_of and Inchoactive_of
6 Using
7 See_also

**FrameNet tagset**

- **Frame elements:** they are organized in several domains like Body, Chance, Cognition, Communication, etc. Each domain contains several frames characterizing different word classes.
- **Grammatical function:** external argument, complement, modifier, extraposed, object, predicate, head and genitive determiner.
- **Phrase types:** types of semantic constituents able to express FEs of the principal predicating words (nouns, adjectives and verbs in the case of English).
- **Null instantiations:** non over realized FEs. The defined types are:
    1. Definite Null Instantiation: Missing element must be something that is already understood in the linguistic or discourse context.
    2. Indefinite Null Instantiation: Transitive verbs that can be used “intransitively” (i.e. eat, drink, bake…).
    3. Constructional Null Instantiation: Omission licensed by a grammatical construction.

**FrameNet annotation**

A FrameNet annotation is a 3-tuple:
<Frame element, grammatical function, phrase type>

The annotation depends on a target word which evokes a given frame.
The FrameNet annotation is a partial annotation: only some parts of the sentence will be annotated. Whole constituents and not only words are tagged.

Example:
frame ( Commerce_sell )
frame-elements {BUYER, SELLER, PAYMENT, GOODS}
[Robin]$_{SELLER}$ sold [a car]$_{GOODS}$ [to Abby]$_{BUYER}$ [for $5,000]$_{PAYMENT}$.

## 7.3 Discussion

The annotation scheme proposed by PropBank starts from a deep syntactic analysis of the sentence. The empty categories or arguments that are not phonetically realized constituted the main problem for the application of the SRL approach.

The annotation of dialog interaction has to deal with structures like disfluencies, non complete sentences, ungrammaticality, etc. Thus, the use of full syntactic analysis and deep syntactic representations is problematic.

The procedure proposed by FrameNet seems to be more suitable to be adapted to the annotation of spoken dialog corpora. One reason is that it does not require a deep syntactic representation. It is sufficient to just select the frame evoking words that become target words of the annotation. Another important motivation is that we can rely on a richer semantic representation useful to deal with different dialog scenarios.

One problem is that one can find cases where the frame elements are present but not the frame evoking word, like in the following example of the MEDIA corpus:

> U: un hôtel à Tolouse avec piscine si possible cet hôtel doit avoir un billard
> *(a hotel in Toulouse with swimming pool if possible the hotel must have a billiard)*

In the first utterance (un hôtel à Toulouse avec piscine si possible, "*a hotel in Toulouse with swimming pool if possible* ") the frame evoking word (something like "I would like to reserve") is not present and the other dialog participant can infer them using domain knowledge.

This suggests the necessity of finding a way to implement domain knowledge in the annotation of spoken dialog using frames.

A further question is related to the goal of the annotation in LUNA. In the case of the annotation with PropBank we have to follow a long way to achieve the semantic representation. PropBank-arguments must be mapped to thematic roles, which constitute an interface between the syntactic representation and the semantic interpretation. Then these thematic roles must be mapped to the semantic roles. As showed in (Giuglea and Moschitti, 2006a) and (Giuglea and Moschitti, 2006b) it is feasible, but the application of this method to dialog corpora seems problematic.

An alternative is to start directly annotating the semantic roles of the entities occurring in the dialog.

## 7.4 Annotation of predicate structure on the LUNA corpus

For the annotation of predicate structures in the LUNA corpus we have decided to use a FrameNet-like approach where the frames are specifically defined for our target domain. The frame elements are the category of entities defined in the domain ontologies. For all the frames we introduce negation as a default FE.

We will start the annotation with a set of frames for each domain. The annotators will have the option to annotate a frame as "`other`" if none of the specified frames can be used to annotate it. This information will be used to complete the list of frames for each domain. If there is no frame evoking expression and the annotator is not able to recognize the frame using only intra-utterance context and domain knowledge, the frame will be annotated with "`unknown`".

The annotation unit is the dialog turn in which it is possible to find several frames. We don't allow the annotation of predicate structure across different turns.

For the annotation first of all we annotate the segments annotated in the domain attribute level with a frame and a frame element. Then if the target is realised we make a pointer from the frame element to the target. There is the possibility of modifying the boundaries of the initial segments or, if necessary, of introducing new markables by hand.

The next step is putting the frame elements and the target (if realized) together in a set.

Example of annotation:
In the following example the tag `link_concept` signalizes the corresponding mark in the domain attribute level.

(Here the frame evoking expression is not overt realized. In the hotel-reservation domain the purpose of the user is to reserve a hotel. The annotator uses this domain knowledge to recognize the frame. )

U: un hôtel à Toulouse avec piscine si possible
*U: a hotel in Toulouse with swimming pool if possible*
frame: **reservation**
frame-elements: {**customer**, **lodging**, **location**, **facility**}

un hôtel / à / Toulouse / avec / piscine / si possible
fe_1 :lodging fe_2:location fe_3 :facitily
set1={fe_1, fe_2, fe_3}

```
<fe id=fe_1" span="word_11..word_12" link_concept="concept_1"
frame="reservation" frame-element="lodging" member="set_1" />
<fe id=fe_2" span="word_14" link_concept="concept_2"
frame="reservation" frame-element="location" member="set_1"
/>
<fe id=fe_3" span="word_16" link_concept="concept_3"
frame="reservation" frame-element="facility" member="set_1"
/>
```

(Now the frame evoking expression is realized)

U: cet hôtel doit avoir un billard
*U: the hotel must have a billiard*
frame: **reservation**
frame-elements: **{lodging, facility}**

cet hôtel / doit avoir / un billard
fe_4:lodging fe_5:target fe_6:facility
set2={fe_4, fe_5, fe_6}

```
<fe id=”fe_4” span=“word_19..word_20”
link_concept=”concept_4” frame=”reservation” frame-
element=”lodging” member=”set_2” />
<fe id=fe_5” span=“word_21..word_22” frame=”reservation”
frame-element=”target” member=”set_2” />
<fe id=fe_6” span=“word_23..word_24” link_concept=”concept_5”
frame=”reservation” frame-element=”facility” member=”set_2”
/>
```

O: je vous propose l'hôtel lafayette
*O: I propose you the hotel Lafayette*
frame: **lodge-offering**
frame-elements: {**operator**, **reservation**, **lodging**, **location**}

je / vous / propose / l'hôtel lafayette
fe_7:operator fe_8:customer fe_9:target fe_10:lodging
set3={fe_7, fe_8, fe_9, fe_10}

```
<fe id=fe_7” span=“word_25” frame=”lodge-offering” frame-
element=”operator” member=”set_3” pointer=”fe_9” />
<fe id=fe_8” span=“word_26” frame=”lodge-offering” frame-
element=”customer” member=”set_3” pointer=”fe_9” />
<fe id=fe_9” span=“word_27” link_concept=”concept_6”
frame=”lodge-offering” frame-element=”target” member=”set_3”
/>
<fe id=fe_10” span=“word_28..word_30”
link_concept=”concept_7” frame=”lodge-offering” frame-
element=”lodging” member=”set_3” pointer=”fe_9” />
```

U: ok, je le prends
*U: ok, I'll take it*
frame: **reservation**
frame-elements: **{customer, lodging, location}**

ok, / je / le / prends
fe_11:customer fe_12: lodging fe_13:target
set4={fe_11, fe_12, fe_13}

```
<fe id=fe_11” span=“word_32” frame=”reservation”
frame-element=”customer” member=”set_4” pointer=”fe_13” />
<fe id=fe_12” span=“word_33” link_concept=”concept_8”
frame=”reservation” frame-element=”lodging” member=”set_4”
pointer=”fa_13” />
<fe id=fe_13” span=“word_34” link_concept=”concept_9”
frame=”reservation” frame-element=”target” member=”set_4” />
```

# 8 Coreference / anaphoric relations

The term "*coreference*" is used in an informal way in corpus annotation to indicate the annotation of anaphoric information and the annotation of information about referring expressions. Anaphoric relations are relations between entities in a discourse model (Webber, 1978; Kamp and Reyle, 1993). A particularly well known case of anaphoric relation is identity (IDENT), which has as a special case –when it relates entities which corefer—the relation of COREFERENCE, defined as follows (van Deemter and Kibble, 2000).

expressions $\alpha_1$ and $\alpha_2$ *corefer* if and only if Referent($\alpha_1$) = Referent($\alpha_2$)

But the notion of anaphoric relation is often generalized to relations other than identity, as in the case of bridging references. Clark (1975) introduced the term bridging references to indicate cases of references that require the use of some reasoning in the identification of their textual antecedent. This definition of bridging was used in the annotation work by Vieira and Poesio (2000), who treated as bridging definite descriptions that either (i) have an antecedent denoting the same discourse entity, but using a different head noun or (ii) are related by a relation other then identity to an entity already introduced in the discourse—so called ASSOCIATIVE references (Hawkins, 1978). Müller and Strube (2001) and other authors used the term bridging expressions only to indicate this second class, expressions which stand in some conceptual relation to the antecedent but do not refer to the same object.

One of the main problems in the annotation of coreference in large scale corpora is that most of the words can be potentially anaphoric. This fact makes it necessary to constrain the selection of markables to be annotated. Most schemes relay on syntactic restrictions to determine if a sequence of words will be considered as a markable, i.e. schemes that restrict the mark-up to NPs, selecting which NPs will be considered referring expression and which not. Another way is restricting annotation to entities that are pre-defined for the domain like the annotation of coreference on the MapTask corpus (Thompson et al. 1993).

## 8.1 Overview of schemes

Here are some of the schemes used for the anaphoric annotation of corpora, with a particular interest for the schemes used in the annotation of dialog.

**MUC-7**

The MUC-7 coreference annotation scheme (Hirschman and Chinchor 1997) was designed to encode information deemed useful for information extraction.
Only coreference relations between nominal constituents were marked. It was possible to select as a markable not only a full NP, also parts of NPs.

**MapTask**

The MapTask corpus is a corpus of dialogs where each participant has a map with several landmarks in front of her/him, which may be different from that of the other participant. (Most of the landmarks are common to both maps, but not all.) A route is drawn on one of

the maps. The participant who has the route (**giver**) has to instruct the other participant (**follower**) to draw one on its map.

No syntactic constraints are specified for the selection of markables for the annotation of coreference, but the annotation is limited to references to MapTask landmarks; no anaphoric or coreference relations are marked.

**EML annotation scheme**

The EML annotation scheme (Müller and Strube 2001) distinguishes between anaphora and bridging references. Anaphora includes both pronominal anaphora and identity links expressed by predicates standing in an **is-a** (or hyponym-hyperonym) relation, as in *the car … the vehicle*. The bridging relations annotate include **cause-effect**, **part-whole** and **entity-attribute**.

This scheme has been used to annotate written text and dialog, but not to our knowledge to create a large corpus.

**ARRAU**

The ARRAU annotation scheme (Artstein and Poesio 2006) has been used to annotate nominal anaphora (including discourse deixis) and deictic references in a medium-scale corpus including both transcripts of spoken dialogs, e.g., from the TRAINS corpus (Gross et al 1993), and text, including texts from the Wall Street Journal portion of the Penn Treebank. The reliability of the scheme has been extensively tested in a series of studies (e.g., Poesio and Artstein, 2005; Artstein and Poesio, 2006), concerned especially with the problem of how to deal with ambiguous anaphoric expressions, which are very common in spoken dialog. The corpus is still under construction.

All NPs recognized by a (full) parser are treated as markables, including expletives and temporal expressions. Annotators then have to decide which of these NPs are **term-denoting**. Next, term-denoting NPs are classified as “old” or “new” depending on whether they refer to a previously mentioned object or not. (Bridging references are classified as new.) If a markable is annotated with “old”, its antecedent is then identified, and the part of text that evokes this antecedent marked. This **textual anchor** could either be a nominal phrase or a sequence of utterances evoking an abstract object.

If the markable is annotated as “new”, the next step is to annotate whether it is **related** to a previous object or not (i.e., if it is a bridging reference). In case it is related, the annotator should indicate the most recent mention of the related object and annotate the type of bridging relation.

An important feature of ARRAU is that in the case of an ambiguous anaphoric expression, the annotators need not choose among its possible antecedents; they can mark more than one. Annotators can also mark ambiguities between old and new readings.

## 8.2 Annotation scheme for the LUNA corpus

The approach to the annotation of coreference that we have chosen is very close to the one used in ARRAU. One of the reasons to select this scheme is that it offers a good and easily applicable scheme for the annotation of anaphoric relations, including bridging, in

dialog corpora. A further reason is the robustness of the scheme that doesn't require having always one single interpretation in the annotation

The one difference is that, in contrast with ARRAU, the annotation of coreference in LUNA will be restricted to the segments annotated with attribute-value pairs in the domain attribute level of our annotation scheme. There is the possibility of modifying the boundaries of these initial segments or, if necessary, of introducing new markables by hand.

**Annotation of agreement features**

A preliminary step will be to annotate markables with agreement features (gender, number) in case this information has not been automatically extracted during some of the automatic preprocessing steps.

**Annotation of the information status**

The next step in the annotation is the decision whether a markable refers to an entity that was already introduced in the dialog (given) or to an entity that was not previously mentioned in the dialog (new). The annotator may also mark an expression as being ambiguous between a given and a new interpretation as explained below.

**Annotation of markables marked as "given"**

In this case, the annotator will select the most recent mention of the referred object and add a pointer to it. If a markable refers to more than one previous object or if there are several possible candidates to be the antecedent it is possible to point to more than just one object as explained below in the sections on plural markables and ambiguity.

**Annotation on markables marked as "new"**

In this case two possibilities will be distinguished: whether the markable is related to an object that was previously mentioned in the dialog (associative reference) or not.

In case of relatedness, the annotator will select the previously introduced referred markable and make a pointer to it. The annotator will then indicate the type of relation: one of the relations and properties defined in the domain ontology, or one of the set relations (element, subset).

If a markable refers to more than one object or if there are several possible candidates to be the antecedent it is possible to point to more than just one object as explained below in plural markables and ambiguity sections.

**Plural markables**

A plural markable is one which refers to a set of objects already mentioned in the dialog.
In this case the markable will be annotated with "multiple referents" and the annotator will add a pointer to each of the antecedents.

**Ambiguity**

An ambiguous markable is a markable which has two or more alternative interpretations.

In this case the markable will be annotated as “ambiguous” and the annotator will add a pointer to each of the antecedents.

**Example of annotation:**
The tag `link_concept` signalizes the corresponding markable in the domain attribute level.

(In the following example, all entities mentioned are new to the discourse.)

O: à [ Paris ]$_{coref1}$ je vous propose [ l’hôtel Ibis Montparnasse ]$_{coref2}$ et [ l’hotel Lafayette ]$_{coref3}$.
*In Paris I propose the Hotel Ibis Montparnasse and the Hotel Lafayette*

```
<coref id=”coref_1” span=”word_2” link_concept=”concept_1”
inf_status=”new” related=”no” />
<coref id=”coref_2” span=”word_6..word_9” link_concept=”concept_2”
inf_status=”new” related=”no” />
<coref id=”coref_3” span=”word_11..word_13”
link_concept=”concept_3” inf_status=”new” related=”no” />
```

(Now the two hotels mentioned in the previous utterance get referred to using a plural reference.)

U : [ ils ]$_{coref4}$ ont [ une piscine ]$_{coref5}$ ?
*Do they have a swimming-pool?*

```
<coref id=”coref_4” span=”word_14” link_concept=”concept_4”
inf_status=”given” multiple_phrase_antecedent=”coref_2;coref_3”
ambiguity=”unambiguous” />
<coref id=”coref_5” span=”word_16..word_17”
link_concept=”concept_5” inf_status=”new” related=”no” />
```

O: [l’hôtel Lafayette]$_{coref6}$ possède [ une piscine ]$_{coref7}$
*The Hotel Lafayette has a swimming pool*

```
<coref id=”coref_6” span=”word_18..word_20”
link_concept=”concept_6” inf_status=”given”
single_phrase_antecedent=”coref_3” ambiguity=”unambiguous” />
<coref id=”coref_7” span=”word_22..word_23”
link_concept=”concept_7” inf_status=”given”
single_phrase_antecedent=”coref_5” ambiguity=”unambiguous” />
```

U: bien, et [ cet hôtel ]$_{coref8}$ accepte [ les chiens ]$_{coref9}$?
*OK, and does this hotel accept dogs?*

```
<coref id=”coref_8” span=”word_26..word_27”
link_concept=”concept_8” inf_status=”given”
single_phrase_antecedent=”coref_6” ambiguity=”unambiguous” />
<coref id=”coref_9” span=”word_29..word_30”
link_concept=”concept_9” inf_status=”new” related=”no”/>
```

O : [ l’hôtel Lafayette ]$_{coref10}$ n’accepte pas [ les animaux ]$_{coref11}$

*Hotel Lafayette does not accept animals.*

```
<coref id="coref_10" span="word_31..word_33"
link_concept="concept_10" inf_status="given"
single_phrase_antecedent="coref_7" ambiguity="unambiguous" />
<coref id="coref_11" span="word_37..word_38"
link_concept="concept_11" inf_status="new" related="yes"
related_phrase="coref_9" relation="superClasseOf"
ambiguity="unambiguous />
```

U: je vais prender [ l'autre hôtel ]$_{coref12}$ alors
*Then I will reserve the other hotel.*

```
<coref id="coref_12" span="word_42..word_44"
link_concept="concept_12" inf_status="given"
single_phrase_antecedent="coref_2" ambiguity="unambiguous" />
```

# 9 Dialog acts

Dialog acts mark the intention of an utterance in a specific dialog and help to obtain information about relationships between utterances.

The basic units of dialog acts are the utterances, sequences of words that are sub components of the turns.

## 9.1 Background: coding schemes for large scale annotation

Follows a list of schemes used for large scale annotation: Verbmobil dialog coding scheme (Alexandersson et all, 1997), HCRC dialog structure coding manual (Carletta et al. 1997), DAMSL (Allen and Core, 1997), SWBD-DAMSL (Jurafsky et al, 1997), ICSI-MRDA (Dhillon et al. 2004)

### Verbmobil coding scheme

Verbmobil was a project aimed to perform translation of spontaneous speech within a dialog situation (Wahlster, 1993). The data consisted of the dialogs in travel planning and appointment scheduling domain.

Verbmobil coding scheme provides a detailed taxonomy of dialog acts but its applicability is limited only to the chosen domain. This taxonomy classifies the dialog acts in three types: control dialog, manage task and promote task.

In order to capture the fact that single utterances may have more than only one function the coding scheme allows the assignation of multiple acts to single utterances.

### HCRC Dialog Structure Coding Manual

The HCRC Dialog Structure Coding Manual assumes to be a general coding scheme for dialog. It was used to annotate the HCRC Map Task Corpus, a corpus of task oriented dialogs where a participant has to mark a route on one of the participants map.

The scheme describes three different levels of dialog structure:

1 Dialog moves: it’s the lowest level and corresponds to utterances
2 Dialog games: they reflect the goal structure of the dialog. Games can be made up of moves and games. The coding scheme doesn't provide a separate coding scheme for games.
3 Transaction level: transactions are made up of conversational games. This level gives the subdialog structure of the dialogs.

### DAMSL

The Dialog Act Markup in Several Layers (DAMSL) was initially developed for the annotation of task oriented collaborative dialogs with two participants, assuming that the scheme should work properly in dialogs where more than two persons take part.

The unit of the annotation is the utterance, which is annotated in four different dimensions:

1 Communicative status: a kind of meta-information about the utterance. It records completeness and interpretability of the utterance.
2 Information level: this dimension provides a characterization of the semantic content of the utterance.
3 Forward looking function: how the current utterance affects the subsequent dialog and beliefs of the participants
4 Backward looking function: how the current utterance relates the previous discourse.

An utterance does not have to be annotated with information from all the four levels.

Although the scheme was initially developed for the annotation of task oriented dialogs its generality makes it easily adaptable to other kinds of task and dialogs, like different kinds of collaborative dialogs or information seeking dialogs.

**SWBD-DAMSL**

The tagset of SWBD-DAMSL implements an extended version of the DAMSL. The annotation scheme doesn't allow combinations of tags for an utterance.

SWBD-DAMSL was applied in the annotation of a large scale corpus of spontaneous two-party telephone dialogs.

**MRDA labeling guide**

The ICSI-MRDA tagset extends and modifies the SWBD-DAMSL tagset for application to multi-party dialogs. It was used to annotate the ICSI Meeting Recorder corpus (Janin et al. 2003; Shiberg et al. 2004).

At the same time the mutual exclusiveness constraint of the SWBD-DAMSL annotation was removed and each utterance can be marked with so many tags as applicable (Dhillon et al. 2004).
A further feature of the ICSI-MRDA tagset is the availability of tags to annotate turn taking mechanisms, interruptions and abandoned utterances.

## 9.2 Dialog acts tagset for LUNA

For the annotation of dialog acts in LUNA we decided to start from a subset of DAMSL because of two reasons: the good coverage showed in the annotation of different kinds of dialog, especially of corpora of telephone dialogs, and the usability of the tagset in the annotation of large scale data.

The initial tagset consists in eight acts, four with forward looking function and four with backward looking function. This tagset will be completed with acts that can be necessary for the individual applications.

The initial tagset:

Forward looking function

- Statement
- Action-directive/open option
- Committing-speaker-future-action
- Info-request

Backward looking function
- Answer
- Accept
- Reject
- Signal-understanding
- Signal-non-understanding

**Forward looking function**

**Statement**

The speaker makes explicit claims about the world as in utterances like

> U: **Sono XXX dalla regione, assessorato alla sanità. Da stamattina non riesco ad entrare nell' applicativo sanità.**
> *U: **I am XXX from the Region, Health Department. Since this morning I cannot access the health application.***

or in answers to questions.

> O: Quando l'aveva richiesto?
> *O: When did you request it?*
> U: **Mi sembra a giugno**.
> *U: **believe in June**.*

**Action-directive/open option**

The purpose of the utterance is to influence the hearer's future non-communicative actions, as in requests or suggestions.

> O: allora, **deve andare su menù start, programmi, IuaLauncher e poi clikkare su IuaLauncher Start; poi chiude tutto, e riapre l'applicativo della sanità.** Quello che le ho lanciato io da remoto è un componente di APRIRE che non parte in automatico. Basta lanciarlo a mano. Adesso vede che gli atti funzionano?
> O: Then, **you should click on the Start menu, Programmes, IuaLauncher and then click on IuaLauncher Start; then close all the windows and open again the Health application.** What I launched from remote is an APRIRE component that does not start automatically. It is sufficient to launch it manually. Now, do you see it works now?

**Committing-speaker-future-action**

The utterance can potentially commit the speaker to some future action.

> U: **Sì. Allora aspetti che mi memorizzo come fare cosi non la distrurbo più**. Allora, devo andare su start, poi su programmi, poi …. Giusto?

*U: Yes. **Please wait a while that I memorise how to do it so as not to disturb you again.** So, I have to click on Start, then Programmes, and then…. Right?*

**Info-request**

The speaker's utterance introduces the obligation to provide information.

U: dzień dobry **chciałam się dowiedzieć jak euh mogę się dostać na przystanek** euh
*U: Good morning, **I would like to know how euh I could get at the stop Legia's stadium***

O: OK grazie. **Mi dica il problema?**
*O: Ok thank you. **Can you tell me your problem?***

**Backward looking function**

**Answer**

The utterance of the speaker is complying with a previous info-request action.

O: Votre problème concerne: un dysfonctionnement de votre ligne fixe ou d'Internet ou plutôt une demande d'information ? Je vous écoute.
*O: Your problem is about: your fixed line or your Internet access or rather a request for information? I'm listening to you.*
U: **en fait, il n'y a plus de tonalité.**
*U: **in fact there is no dialling tone.***

**Accept**

The speaker is agreeing to part of the proposal, request or claim done by the other dialog participant.

O: Funziona. Aspetti che già che ci sono controllo l'antivirus. Sì, è da aggiornare. Signora, provi a riavviare il computer. Quano poi si riavvia, mi dice se vede ancora il ! sull'icona di Norton?
*O: It works. Please wait for me to check the antivirus as well. Yes, it must be updated. Madam, please restart the PC. When it is on again, can you tell me if you still see the ! on the Norton icon?*
U: **Sí va bene.**
*U: **Yes, all right.***

**Reject**

The speaker is disagreeing to part of the proposal, request or claim done by the other dialog participant.

O: hmhm , no jest jeszcze , ale to troszeczkę dalej z ząbkowskiej (street) 138

*O: hmhm, well there is an other, but a little further from ząbkowska 138*
U: hm , **nie**
*U: hm,* ***no***

A negative response to a question, statement or proposal is not necessarily a “reject”. If the previous question or statement is phrased in the negative a “no” can be an “accept” like in the following example.

O: Lei non aveva chiamato prima, vero? Perché non ho i suoi dati.
O: You have never called before, have you? Because I don't have your data.
U: **no.**
U: **no.**

**Signal-understanding**

Utterances that signal that the speaker understand what the other dialog participant said. They can have different forms, like discourse particles (I see, OK, etc.), repetitions or paraphrases.

O: Vous voulez acheter des actions au règlement comptant sur votre compte-titres ordinaire. Quelle quantité ?
O: *You want to buy shares with immediate payment on standard portofio. How many?*
U: trois
U: *three*
O: **Vous voulez acheter trois actions.** Au prix du marché ou à quel cours limite ?
O: ***You want to buy three shares.*** *Current rate or limited rate?*

**Signal-non-understanding**

The utterance indicates explicitly a problem in understanding a previous utterance of the other dialog participant.

O: **Je ne vous ai pas compris.** Pouvez-vous répéter? Vous voulez une information boursière sur quelle action ?
O: ***I didn't understand.*** *Could you repeat? Information about which share?*

## 9.3 Annotation of the corpus

The function of the annotation of dialog acts on the LUNA corpus is to associate the intentions of the speakers with the propositional content of the utterances. That has motivated us to use as segmentation criterion the annotation of the predicate structure on the previous level.

The criterion is that each set of constituents as defined in the previous annotation will be considered as an independent utterance.

In the example presented on page 25 *"un hôtel à Toulouse avec piscine si possible cet hôtel doit avoir un billard"* (a hotel in Toulouse with swimming pool if possible the hotel must have a billiard), we have two predicates or sets of constituents that are part of the same argument structure:

[un hôtel]coref$_1$ [à Toulouse]coref$_2$ [avec piscine]coref$_3$ si possible [cet hôtel]coref$_4$ [doit avoir]coref$_5$ [un billard]coref$_6$
set1={id1, id2, id3}
set2={id4, id5, id6}

The first partition we have is:
[un hôtel à Toulouse avec piscine] si possible [cet hôtel doit avoir un billard]

This segmentation can be corrected by hand in order to incorporate constituents that are outside of the markables to one of them. That can be done using syntactic criteria or prosody (criterion used in the next example).

[un hôtel à Toulouse avec piscine si possible] [cet hôtel doit avoir un billard]

In case of ambiguities words or syntactic constituents can be left out of the markables.

The next step is the assignment of tags. With the goal of capturing the intentions of the speaker, each dialog turn will be annotated with as many tags as possible, something that other annotation schemes like the Verbmobil dialog coding scheme or the MRDA labeling guides allow.

[un hôtel à Toulouse avec piscine si possible]$_{utt-2}$

```
<utt id="utt_2" span="word_11..word_18" link_fe_set="set_1"
da-tag-1="statement" da-tag-2="answer">
```

(The tag `link_fe_set` signalizes the corresponding set of frame elements in the predicate structure level.)

[cet hôtel doit avoir un billard]$_{utt-3}$

```
<utt id="utt_3" span="word_19..word24" link_fe_set="set_2"
da-tag-1="statement" da-tag-2="answer">
```

Sometimes full utterances that are outside of the annotated predicates are relevant for the dialog management, like in the next example:

je ne vous ai pas compris pouvez-vous répéter [vous voulez une information boursière sur quelle action] (*I didn't understand. Could you repeat? Information about which share?*)

For these cases we are using two criteria for this further segmentation:

1. List for each languages of expressions which corresponds to dialog acts that are considered important for a concrete domain application, like opening/closings, politeness formulae, etc.
2. Segmentation based on syntactic constrains: a verb and its arguments constitute an utterance.

[Je ne vous ai pas compris] [pouvez-vous répéter] [vous voulez une information boursière sur quelle action?]

[je ne vous ai pas compris]$_{utt-5}$
```
<utt id="utt_4" span="word_30..word_35" d-tag-1="statement"
da-tag-2="signal-non-understanding">
```

[pouvez-vous répéter]$_{utt-6}$
```
<utt id="utt_5" span="word_36..word_38"
da-tag-1="info-request" da-tag-2="signal-non-understanding">
```

[Vous voulez une information boursière sur quelle action?]$_{utt-7}$
```
<utt id="utt_6" span="word_39..word_46" link_fe_set="set_3"
da-tag-1="info-request" da-tag2="signal-non-understanding">
```

# 10 Conclusion

In this document we have described the annotation layers of the spoken dialog corpora that will be collected along the duration of the LUNA project. The annotation layers and features support each corpus data collection in a specific domain (e.g. call routing vs travel information), language (e.g. Italian vs. French) and modality (human-human vs. human-machine). The document will be the core of the annotation manual to be used in workpackage 5 during the data annotation process.

# Appendix 1: Annotated example

**Bienvenue sur Vocalia Bourse. Souhaitez-vous consulter votre portefeuille ou votre carnet d'ordres, obtenir une information boursière ou effectuer une transaction ?**
*[Welcome on Vocalia Bourse. Do you wish to consult your portfolio or your order account, obtain stocks information or purchase a transaction?]*

**U: une information boursière**
*U* : *stock exchange information*

Domain attribute level
```
<concept ID=”concept_1” span=”word_1..word_3” attribute=”object”
value=”timed_quantity” />
```

Predicate structure level
```
<fe id=”fe_1” span=”word_1..word_3” link_concept=”concept_1”
frame=”value-request”
frame-element=”target” member=”set_1” />
```

Coreference level
```
<coref id=”coref_1” span=”word_1..word_3” link_concept=”concept_1”
inf_status=”new” related=”no” />
```

Dialog acts
```
<utt id=”utt_1” span=”word_1..word_3” link_fe_set=”set_1”
da-tag-1=”statement” da-tag-2=”answer”>
```

**O: Vous voulez une information boursière sur une action, un marché ou un indice ?**
*O: an information about a share, a market or an index?*

Domain attribute level
```
<concept id=”concept_2” span=”word_6..word_8” attribute=”object”
value=”timed_quantity” />
<concept id=”concept_3” span=”word_10..word_11” attribute=”object”
value=”action” />
<concept id=”concept_4” span=”word_12..word_13” attribute=”object”
value=”market” />
<concept id=”concept_5” span=”word_14” attribute=”conjunct”
value=”alternative” />
<concept id=”concept_6” span=”word_15..word_16” attribute=”object”
value=”index” />
```

Predicate structure level
```
<fe id=”fe_2” span=”word_6..word_8” link_concept=”concept_2”
frame=”value-request” frame-element=”target” member=”set_2” />
<fe id=”fe_3” span=”word_10..word_11” link_concept=”concept_3”
frame=”value-request” frame-element=”hasValue” member=”set_2”
pointer=”fe_2” />
```

```
<fe id="fe_4" span="word_12..word_13" link_concept="concept_4"
frame="value-request" frame-element="hasValue" member="set_2"
pointer="fe_2" />
<fe id="fe_5" span="word_15..word_16" link_concept="concept_6"
frame="value-request" frame-element="hasValue" member="set_2"
pointer="fe_2" />
```

Coreference level

```
<coref id="coref_2" span="word_6..word_8" link_concept="concept_2"
inf_status="given" single_phrase_anteceden="coref_1"
ambiguity="umambiguous" />
<coref id="coref_3" span="word_10..word_11"
link_concept="concept_3" inf_status="new" related="no" />
<coref id="coref_4" span="word_12..word_13"
link_concept="concept_4" inf_status="new" related="no" />
<coref id="coref_5" span="word_15..word_16"
link_concept="concept_6" inf_status="new" related="no" />
```

Dialog acts

```
<utt id="utt_1" span="word_4..word_16" link_fe_set="set_2"
da-tag-1="info-request" da-tag-2="signal-understanding" />
```

**U: l'action <OOV>**
*U*: *share <out of vocabulary>*

Domain attribute level

```
<concept id="concept_7" span="word_17..word_19" attribute="object"
value="action" />
```

Predicate structure level

```
<fe id="fe_6" span="word_17..word_19" link_concept="concept_7"
frame="unknown" frame-element="unknown" member="set_3" />
```

Coreference level

```
<coref id="coref_6" span="word_15..word_16"
link_concept="concept_7" inf_status="new" related="yes"
related_phrase="coref_3" relation="instanceOf"
ambiguity="unambiguous" />
```

Dialog acts

```
<utt id="utt_2" span="word_17..word_19" link_fe_set="set_3"
da-tag-1="statement" da-tag-2="answer">
```

**O: Vous voulez une information boursière sur quelle action ?**
O: *an information about which share?*

Domain attribute level

```
<concept id="concept_8" span="word_22..word_24" attribute="object"
value="timed_quantity" />
```

```
<concept id="concept_9" span="word_26..word_27" attribute="object"
value="action" />
```

Predicate structure level

```
<fe id="fe_7" span="word_22..word_24" link_concept="concept_8"
frame="value-request" frame-element="target" member="set_4" />
<fe id="fe_8" span="word_26..word_27" link_concept="concept_9"
frame="value-request" frame-element="hasValue" member="set_4"
pointer="fe_7" />
```

Coreference level

```
<coref id="coref_7" span="word_15..word_16"
link_concept="concept_8" inf_status="given"
single_phrase_antecedent="coref_2" ambiguity="umambiguous" />
<coref id="coref_8" span="word_15..word_16"
link_concept="concept_9" inf_status="given"
single_phrase_antecedent="coref_3;coref_6" ambiguity="ambiguous"
/>
```

Dialog acts

```
<utt id="utt_3" span="word_20..word_27" link_fe_set="set_4"
da-tag-1="info-request" />
```

**U: oh non**
*U: oh no*

Dialog acts

```
<utt id="utt_4" span="word_28..word_29" da-tag-1="reject">
```

**O: Je ne vous ai pas compris. Pouvez-vous répéter ? Vous voulez une information boursière sur quelle action ?**
O: *I didn't understand. Could you repeat? Information about which share?*

Domain attribute level

```
<concept id="concept_10" span="word_41..word_43"
attribute="object" value="timed_quantity" />
<concept id="concept_11" span="word_45..word_46"
attribute="object" value="action" />
```

Predicate structure level

```
<fe id="fe_9" span="word_41..word_43" link_concept="concept_10"
frame="value-request" frame-element="target" member="set_5" />
<fe id="fe_10" span="word_45..word_46" link_concept="concept_11"
frame="value-request" frame-element="hasValue" member="set_5"
pointer="fe_5" />
```

Coreference level

```
<coref id="coref_9" span="word_41..word_43"
link_concept="concept_10" inf_status="given"
single_phrase_anteceden="coref_7" ambiguity="umambiguous" />
```

```
<coref id="coref_10" span="word_45..word_46"
link_concept="concept_11" inf_status="given"
single_phrase_anteceden="coref_8" ambiguity="umambiguous" />
```

Dialog acts

```
<utt id="utt_5" span="word_30..word_35"
da-tag-1="statement" da-tag-2="signal-non-understanding">
<utt id="utt_6" span="word_36..word_38" link_fe_set="set_3"
da-tag-1="info-request" da-tag2="signal-non-understanding">
<utt id="utt_7" span="word_39..word_46" link_fe_set="set_5"
da-tag-1="info-request" />
```

**U: Renault**
*U: Renault*

Domain attribute level

```
<concept id="concept_12" span="word_47" attribute="share"
value="renault" />
```

Predicate structure level

```
<fe id="fe_11" span="word_47" link_concept="concept_12"
frame="unknown" frame-element="unknown" member="set_6" />
```

Coreference level

```
<coref id="coref_11" span="word_47" link_concept="concept_12"
inf_status="new" related="yes" related_phrase="coref_10"
relation="instanceOf" ambiguity="unambiguous" />
```

Dialog acts

```
<utt id="utt_8" span="word_47" link_fe_set="set_6"
da-tag-1="statement" da-tag-2="answer">
```

# Appendix 2: Glossary

| | |
|---|---|
| **DAMSL** | Dialogue Act Markup in Several Layers |
| **DARPA** | Defense Advanced Research Projects Agency |
| **DARPA ATIS** | DARPA Airline Travel Information Systems |
| **EAGLES** | Expert Advisory Group on Language Engineering Standards |
| **EML** | European Media Laboratory Research GmbH |
| **FE** | Frame Element |
| **HCRC** | Human Communication Research Centre at the Universities of Edinburgh and Glasgow |
| **ICSI** | International Computer Science Institute at the University of California, Berkeley |
| **ICSI-MRDA** | ICSI Meeting Recorder Dialog Act Corpus |
| **ID** | Identifier |
| **MUC-7** | 7th Message Understanding Conference |
| **NP** | Noun Phrase |
| **OWL** | Web Ontology Language |
| **OWL-DL** | OWL-Description Logics (dialect of the OWL language) |
| **POS** | Part Of Speech |
| **SLU** | Spoken Language Understanding |
| **SWBD-DAMSL** | Switchboard DAMSL annotation scheme |
| **XML** | Extensible Markup Language |
| **W3C** | World Wide Web Consortium |

# Appendix 3: References

Alexandersson, J; Buschenbeck-Worlf, B.; Fujinami, T.; Maier, E.; Reithinger, N.; Schmitz, B.; siegel, M. (1997). Dialogue Acts in Verbmobil-2. Technical Report 204, BMBF.

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

Artstein, R. and Poesio, M. (2006). ARRAU Annotation Manual (TRAINS dialogues). Technical report. University of Essex.

Baker, C. F., Fillmore, C. J. and Lowe, J. B. (1998). The Berkeley FrameNet project. In Proceedings of COLING-ACL.

Barras, C.; Geoffrois, E.; Wu, Z. and Liberman M. (1998). Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. First International Conference on Language Resources and Evaluation (LREC)

Bonneau-Maynard, H. and Rosset, S. (2003). A semantic representation for spoken dialogues. In Eurospeech, Geneva.

Carletta J.; Isard, A.; Isard, S.; Kowto, J.; Doherty-Sneddon, G. and Anderson, A. (1997). The Reliability of a Dialogue Structure Coding Scheme. Computational Linguistics, vol 23, no1.

Clark, H. H. (1975). Bridging. In Proceedings of TINLAP-1.

Dhillon R.; Bhagat S.; Carvey H.; Shriberg E. (2004). Meeting Recorder Project: Dialog Act Labeling Guide, ICSI Technical Report TR-04-002.

EAGLES (1996). EAGLES recommendations for the morphosyntactic annotation of corpora, Technical Report EAG-TCWG-MAC/R, ILC-CNR, Pisa. Italy

EAGLES (1996b). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages, Technical Report EAG-LWG-Morphsyn, ILC-CNR, Pisa. Italy

EAGLES (1996c). Recommendations for the Syntactic Annotation of Corpora, Technical Report EAG---TCWG---SASG/1.8, ILC-CNR, Pisa. Italy.

Fellbaum, C. (1987). WordNet: An electronic lexical database. MA: MIT Press.

Fillmore, C. J. (1985). Frames and the semantics of understanding. Quaderni di Semantica IV(2).

Giuglea, A. M. and Moschitti, A. (2006a). Semantic Role Labeling via FrameNet, VerbNet and PropBank. In Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)

Giuglea, A. M. and Moschitti, A. (2006b). Shallow Semantic Parsing Based on FrameNet, VerbNet and PropBank. In Proceedings of the 17th European Conference on Artificial Intelligence.

Gross, D; Allen J. and Traum, D. (1993). The TRAINS 91 Dialogues, TRAINS Technical Note 92-1

Gupta N., Tur G., Hakkani-Tur D., Bangalore S, Riccardi G. and Rahim M. (2005). The AT&T Spoken Language Understanding System. IEEE Trans. on Speech and Language Processing, volume PP, Issue 99, pp. 1-10, 2005.

Hawkins, J. A. (1978). Definiteness and Indefiniteness. Croom Helm, London.

He, Y. and Young, S. (2006). Spoken Language Understanding using the Hidden Vector State Model. Speech Communication, special issue on Spoken Language Understanding, vol. 48. March 2006.

Hirschman, L. and Chinchor, N. (1997). MUC-7 coreference task definition, version 3.0. In Proceedings of the 7th Message Understanding Conference. Science Applications International Corporation.

Janin, A.; Baron, D.; Edwards, J. A.; Ellis, D.; Gelbart, D.; Morgan, N.; Peskin, B.; Pfau, T.; Shiberg, E.; Stolcke, A.; Wooters, Ch. (2003). The ICSI Meeting corpus. In Proceedings of ICASSP 2003 (IEEE International conference on Acoustics, Speech and Signal Processing), Hong Kong.

Jurafsky, D; Srhibert, E., Biasca, D. (1997). Switchboard SWBD-DAMSL shallow discourse function annotation: Coders manual. Technical Report 97-02, draft 13, University of Colorado, Institute of Cognitive Sciences.

Kamp, H. and Reyle, U. (1993). From Discourse to Logic. Dordrecht: Kluwer

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

Kipper-Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. PhD. Thesis. Computer and Information Science Dept. University of Pennsylvania.

Levin, E. and Pieraccini, R. (1995) "Concept-based spontaneous speech understanding system". In Proc. European Conference on Speech Communication and Technology, Vol.2. Madrid, Spain, pp. 555-558.

Marcus, M.; Santorini, B. and Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: the Penn Treebank. In Computational Linguistics 19.

Miller, S.; Schwartz, R.; Brobow, R. and Ingria, R. (1994). "Statistical Language Processing Using Hidden Understanding Models", Proc. of the Spoken Language Technology Workshop, 48-52, Plainsboro, New Jersey, Morgan Kaufmann publ, Los Altos, CA.

Müller, C. and Strube, M. (2001). Annotating Anaphoric and Bridging Relations with MMAX. Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue.

Smith, M.; Welty, Ch. And McGuiness, D.L. (2004). OWL Web Ontology Language Guide

Shriberg, E.; Dhillon, R; Bhagat, S.; Ang, J.; Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In Procedings of SIGdial 2004.

Thompsom, H. and McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents. In Proceedings of SGML Europe'97.

van Deemter, K. and Kibble, R. (2000). On Coreferring: Coreference in MUC and Related Annotation Schemes. Computational Linguistics, 26(2).

Vieira R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. Computational Linguistics, 26(4).

Wahlster, W. (1993). Verbmobil - Translation of face-to-face dialogues. In Proceedings of the Fourth Machine Translation Summit, Kobe, Japan.

Walker, M.; Aberdeen, J.; Boland, J.; Bratt, E.; Garofolo, J.; Hirschman, L.; Le, A.; Lee, S.; Narayanan, S.; Papineni, K.; Pellom, B; Polifroni, J.; Potamianos, A.; Brabhu, P.; Rudnicky, A.; Sanders, G.; Seneff, S.; Stallard, D. and Whittaker, S. (2001). DARPA Communicator dialog travel planning systems: The June 2000 data collection. In Proc. EUROSPEECH.

Webber, B. (1978). A Formal Approach to Discourse Anaphora. Ph.D. thesis, Harvard University.