The LUNA Corpus: an annotation scheme for a multilingual multidomain dialogue corpus

SIXTH FRAMEWORK PROGRAMME

Information Society Technologies

Kepa Joseba Rodriguez

Piedmont Consortium for Information Systems

csipiemonte

LUNA project

Potsdam, 6 June 2007

Outline

- **The LUNA project**
  - Consortium
  - Goals
  - Modules of the SLU toolkit

- **The LUNA corpus**
  - Function of the corpus in the project
  - Description
  - Historical background
  - Levels of annotation

The LUNA project

LUNA

Spoken Language UNderstanding in MultilinguAl Communication Systems.

www.ist-luna.eu

The LUNA project

The LUNA consortium

- Piedmont Consortium for Information Systems (CSI-Piemonte, IT)
- University of Trento (IT)
- Loquendo (IT)
- University of Avignon (FR)
- France Telecom (FR)
- RWTH University Aachen (DE)
- Polish-Japanese Institute of Information Technology (PL)
- Polish Academy of Sciences (PL)

The LUNA project

Goals of the project

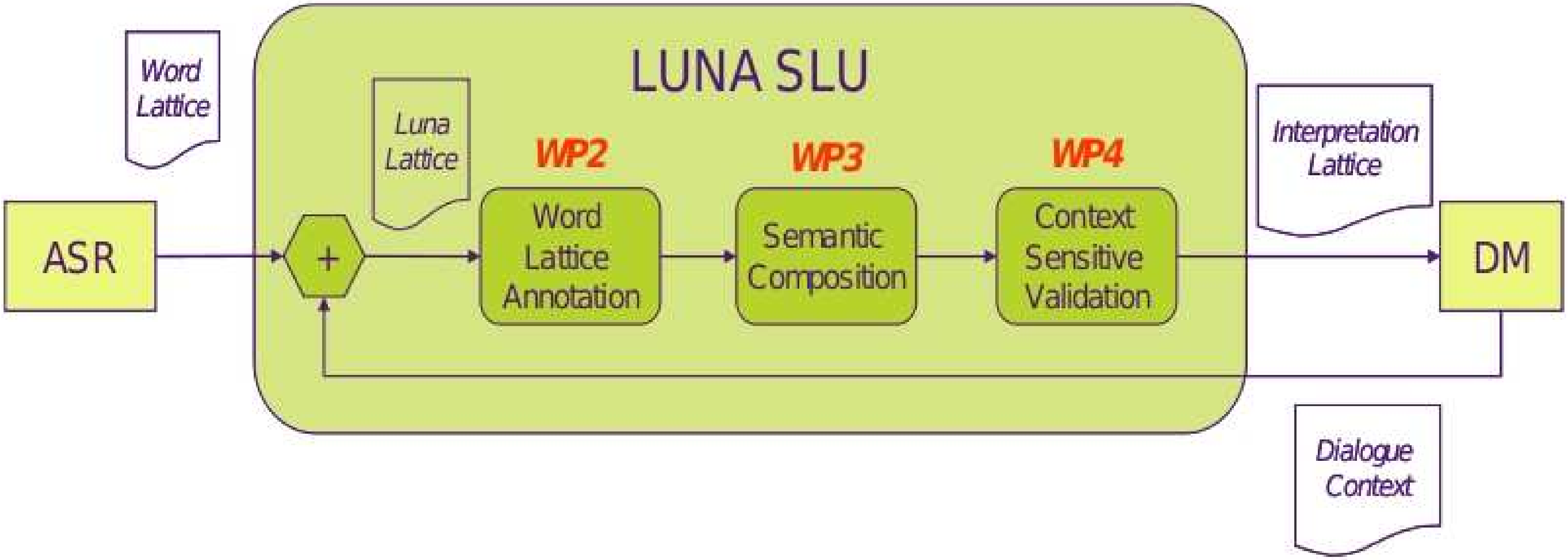
- **The focus of the LUNA project is the real time understanding of spontaneous speech in dialogue systems.**

- **Three steps are considered for the SLU interpretation process:**
  - Generation of semantic concept tags.
  - Composition into conceptual structures.
  - Context sensitive validation using information provided by the dialogue manager.

- **The SLU models will be applied to different conversational systems in Italian, French and Polish.**

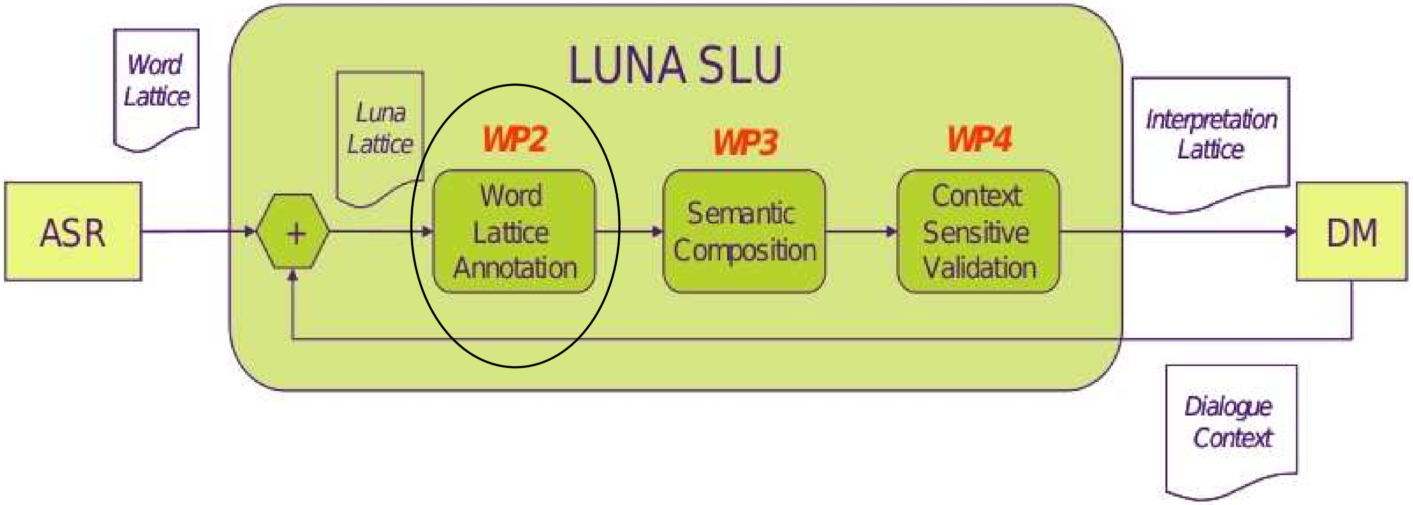
The LUNA project

Modules of the SLU toolkit



The LUNA project

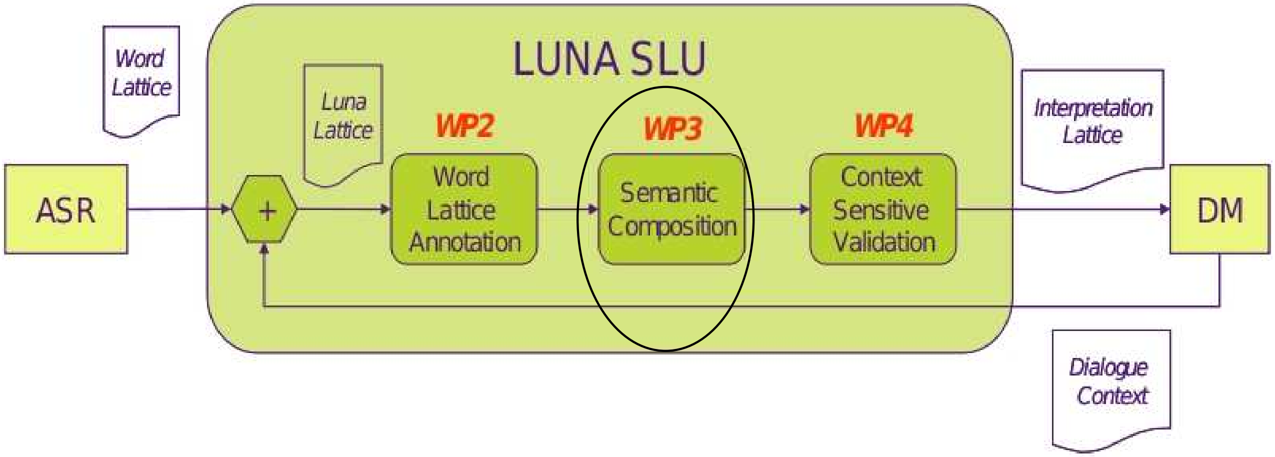
The Word Lattice Annotation Module



- **Input: LUNA lattice**
  - word lattice produced by the ASR enriched with context information coming from the dialogue manager.
- **Output: concept lattice**
  - LUNA lattice annotated with semantic concepts.
- **Semantic concepts: basic units of meaning in each application domain.**

The LUNA project

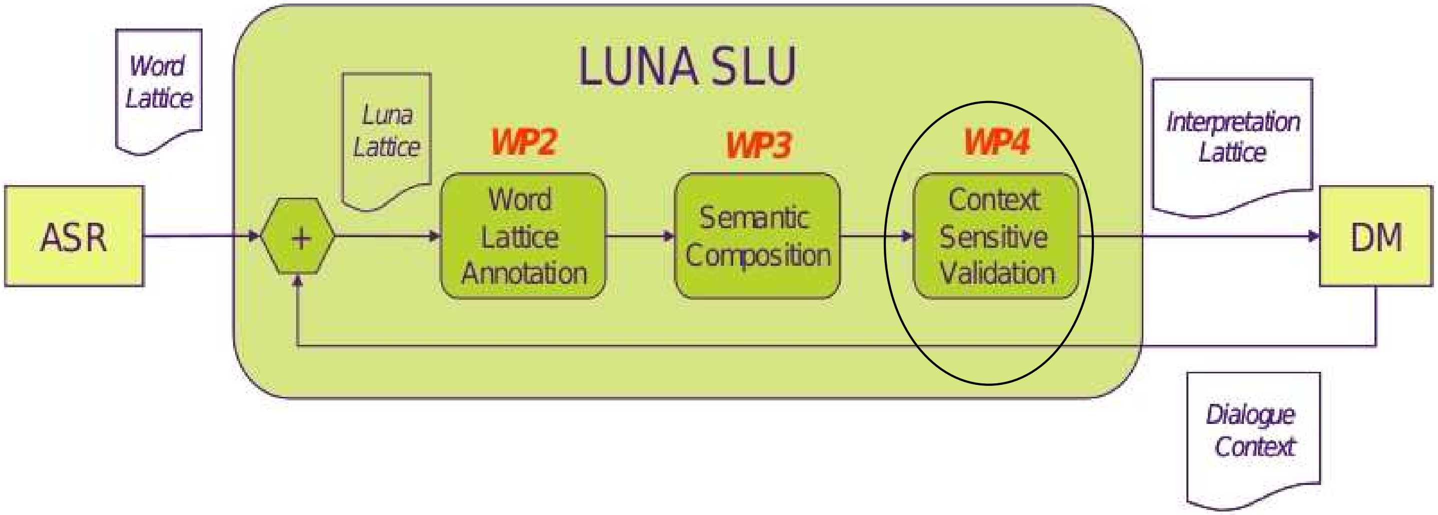
The Semantic Composition Module



- **Input: concept lattice (with dialogue context) output by the previous module.**
- **Output: lattice of semantic structures representing all the possible interpretations of the utterance.**
- **Kind of models involved in the module:**
  - Semantic composition model: builds a set of all the possible interpretations of the utterance.
  - Semantic confidence score model: adds a confidence score to each hypothesis of the lattice.
  - Decision module: re-evaluates the interpretations following an interpretation strategy

The LUNA project

The Context Sensitive Validation Module



- **Contextual information can introduce several modifications:**
  - Specification of concepts: dialogue context can desambiguate concepts detected in the WLA module.
  - Specification of interpretations: e.g. resolution of referring expressions.
  - Rescoring of the interpretation lattice.

- **The input/output format of the module is identical to the output format of the previous module.**

Outline

- Function of the corpus in the LUNA project
- Description
- Background: the MEDIA evaluation project
- The multi-level annotation scheme

The LUNA corpus

Function of the corpus

- **Training of the modules of the SLU toolkit.**
  - Statistical models of understanding.
  - Statistical models of dialogue.
- **Evaluation of the modules of the SLU toolkit.**
  - Different evaluation metrics.
- **Ressource for other tasks like retraining of ASR and NLP tools.**

The LUNA corpus

Description of the corpus

- **Target: collection and annotation of**
  - 3000 Human-Human and
  - 8100 Human-Machine dialogues
  - in French, Italian and Polish.
- **French subcorpus:**
  - Application domains: travel information and reservation, IT help desk, telecom costumer care and financial information transaction
  - Human-Machine dialogues: 7100
- **Italian subcorpus:**
  - Application domain: IT helpdesk
  - Human-Human dialogues: 2500
  - WOZ dialogues: 500
- **Polish subcorpus:**
  - Application domain: public transportation information
  - Human-Human dialogues: 500
  - WOZ dialogues: 500

The LUNA corpus

Historical background: the MEDIA evaluation project.

- **Annotation of semantic segment as tuplet:**
  - Mode: **`positive, negative, interrogative, if-possible`**.
  - Attribute: name of the concept.
  - Value
  - Link: pointer to related segments.
  - Comment on the segment.
- **MEDIA proposes a taxonomy of dialogue acts.**

The LUNA corpus

Example of MEDIA annotation

U: un hôtel / à toulouse / avec piscine si possible

(a hotel in toulouse with swimming pool if possible)

```
1: +/objectBD : hotel
2: +/localisation-ville-hotel : toulouse
3: ~/hotel-services : piscine
```

U: cet / hôtel / doit avoir un billard

(the hotel must have a billiard hall)

```
4: +/lienRef-coRef : singulier      reference = {(1,2,3)}
5: +/objectBD : hotel
6: +/hotel-services : billard
```

S: je vous propose l'hôtel lafayette

(I propose you the hotel Lafayette)

```
7: +/nom-hotel : lafayette
```

The LUNA corpus

- **Critique:**
  - Compact format: information at different levels is put together
  - The definition of the attribute mode.
    - **`Affirmative/Negative`** belongs to the semantic of the sentence
    - **`Interrogative`** belongs to the communicative level / dialogue acts
    - **`If-possible`** signalizes only the importance of a parameter.

- **Proposal:**
  - A modular approach splitting the annotated information in different levels.
  - The annotation levels should correspond to the modules of the toolkit.

- **Advantages:**
  - Easier for the annotators.
  - Helpful to investigate the interaction between different levels of representation.

Levels of annotation

- **Segmentation of the speech signal**
- **Word transcription / orthographic annotation**
- **Morphosyntactic annotation: POS and chunking**
- **Domain attribute level**
- **Predicate structure**
- **Coreference and anaphoric relations**
- **Dialogue acts**

Levels of annotation

Segmentation of the speech signal

- Segmentation of the speech signal in dialogue turns.
- The turns will be annotated with speaker identity, gender and time stamps.
- Goal: give the possibility of transcribing segments without a dialogue context.
- Interesting issue to be investigated: how the availability of context can influence the transcription/annotation.

Levels of annotation

Word transcription / orthographic annotation

The principal features of the transcription scheme are:

- **Spelling: using orthographical standards for each language.**
- **Capitalization: following the standards of the languages.**
  - Initial words of sentences will be capitalized only if they would be capitalized in the middle of the sentence.
- **Numbers: spelled out following the standards of each language.**
- **Punctuation: the transcription doesn’t include punctuation marks**

Levels of annotation

Acoustic events:

- **Lexical events.**
  - Word truncations.
  - Pause fillers, hesitations, human noises.
- **Foreign words.**
- **Pronunciation:**
  - Spelled words.
  - Mispronounciation.
  - Unintelligible words.
- **Noises:**
  - Non human noises
  - Silence: only intra-turn silences longer than 1 sec.

Levels of annotation

Example transcription

[Operator:] allora m'ha detto che non riusciva ad accedere al computer *`[silence]`* e le manca la procedura *`[pron=unintelligible]`*
so, you have told me that you cannot access to the computer, and that you need the procedure

[Caller:] esatto
exactly

[Operator:] allora avrei bisogno dell' *`[lex=filler]`* *`[pron=spelled-]`* RWS *`[-pron=spelled]`* del *`[pron=spelled-]`* PC *`[-pron=spelled]`*
so I need the RWS of the computer

[Caller:] si allora tredici zero ottantasei
yes, 13 0 86

Levels of annotation

POS and chunking

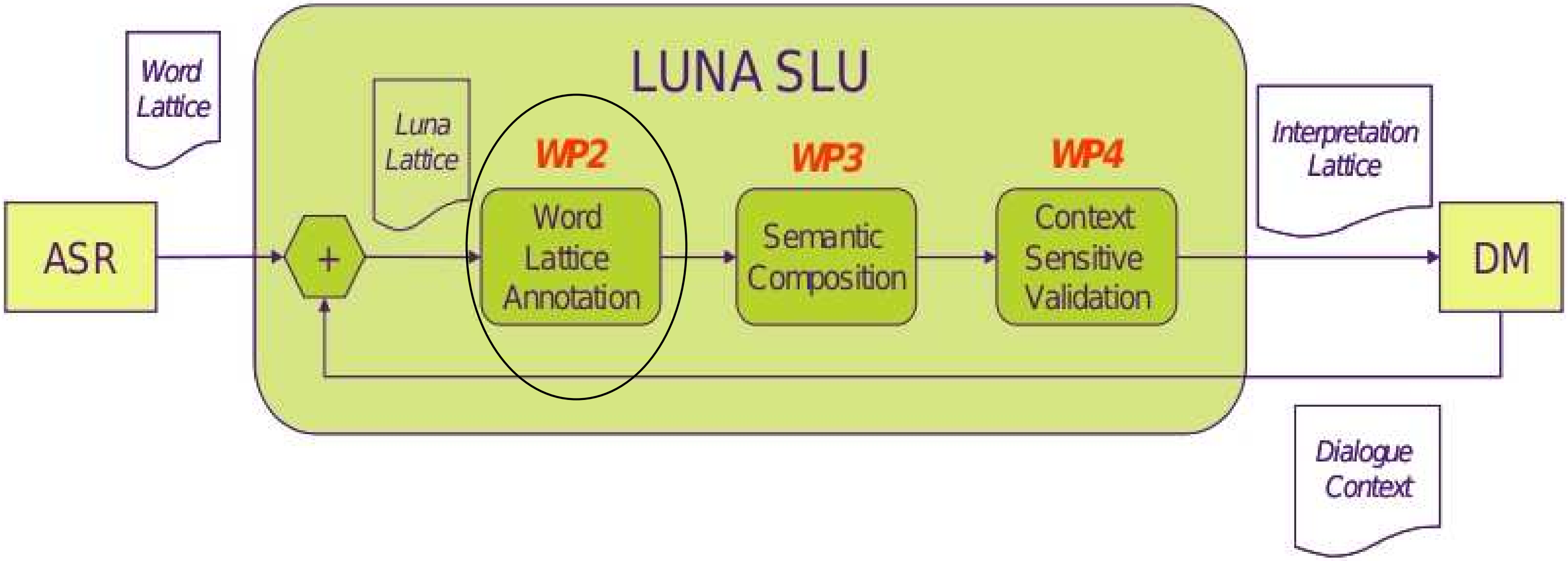
- The annotation on this level is done automatically using available tools for each language.

- Produced chunks can be the basis of the annotation in the next levels.

- For the POS-tags and morphosyntactic features, we follow the recommendations made in EAGLES. That allows us to have a unified representation for the corpus independent from the tools used for each language.

Levels of annotation

Domain attribute level



Levels of annotation

- **Starting from the output of the chunker we produce semantic segments.**

- **These segments will be annotated with attribute-value pairs. Similar approach as in MEDIA.**

- **Domain knowledge is specified in domain ontologies.**

- **With the ontologies we build domain specific concept dictionaries. Each dictionary contains:**
  - **Concepts:** corresponding to classes of the ontology and attributes of the annotation.
  - **Values:** corresponding to individuals of the domain.
  - **Constraints** on the admissible values for each concept.

Levels of annotation

Example domain attribute annotation

[Operator:] allora m'ha detto che [non riusciva]c1 ad [accedere]c2 [al computer]c3 e [le manca]c4 [la procedura]c5

so, you have told me that you cannot access the computer, and that you need the procedure

```
c1 trouble : unable_to
c2 action : access
c3 computer-hardware : pc
c4 trouble : lack_of
c5 computer-software : procedure
```

[Caller:] esatto

exactly

[Operator:] allora avrei bisogno [dell' RWS]c6 [del PC]c7

so I need the RWS of the computer

```
c6 concept : code-identificationCode
c7 computer-hardware : pc
```

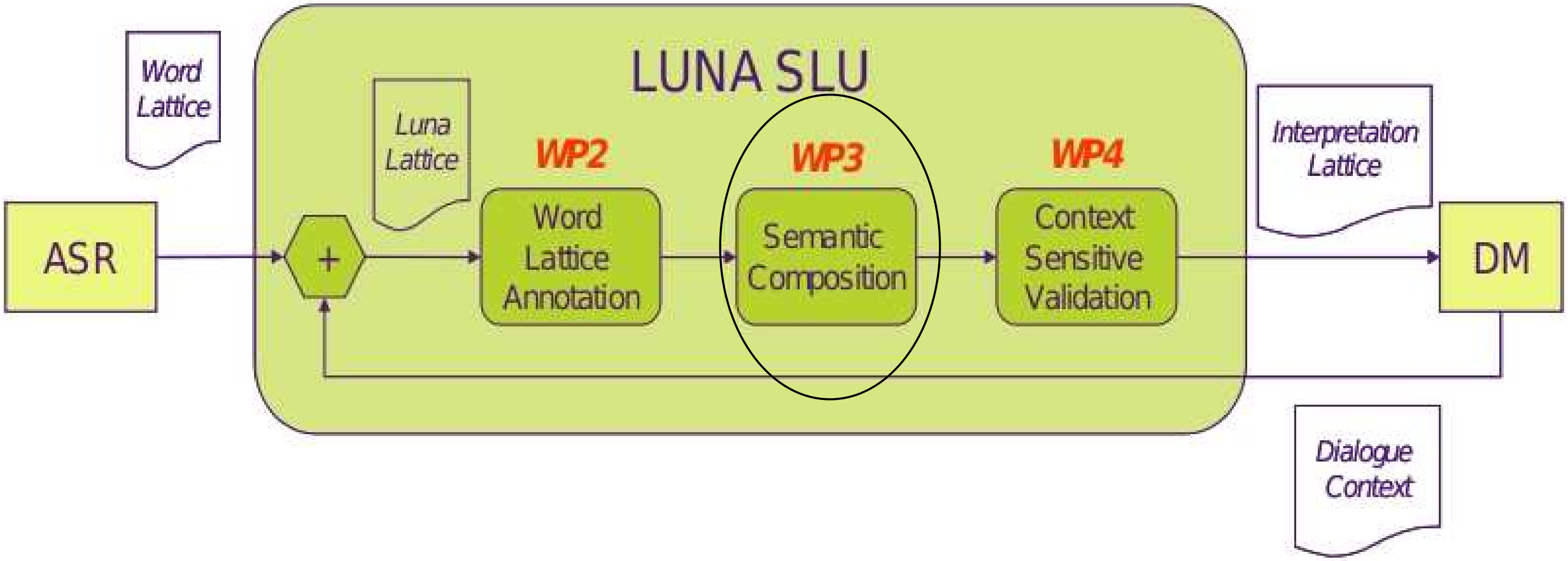
[Caller:] si allora [tredici zero ottantasei]c8

yes, 13 0 86

```
c8 code-identificationCode-rws : 13086
```

Levels of annotation

Predicate structure level



Levels of annotation

- **The corpus will be annotated using a domain adapted version of FrameNet.**

- **A short overview:**
  - Semantic frames: script-like conceptual structure that describes a type of situation, object or event along with its participants. They encode a part of the real-world knowledge in a schematic form.

  - Example of FrameNet:

```
Frame (CommercialTransaction)
• frame-elements: {buyer, seller, payment, goods}
• scenes (buyer gets goods, seller gets payment)
```

- **Definition of the frames for each domain starting from the domain ontologies.**

Levels of annotation

Example:

I am XXX from the Region, Health Department. **From this morning I cannot access the health application.**

frame: access
frame-elements: {user, application, temp}

From this morning I cannot <u>access</u> the health application.

We add the negation as default frame-element for all the frames.

From this morning I cannot <u>access</u> the health application.

Levels of annotation

Example: annotation of predicate structure

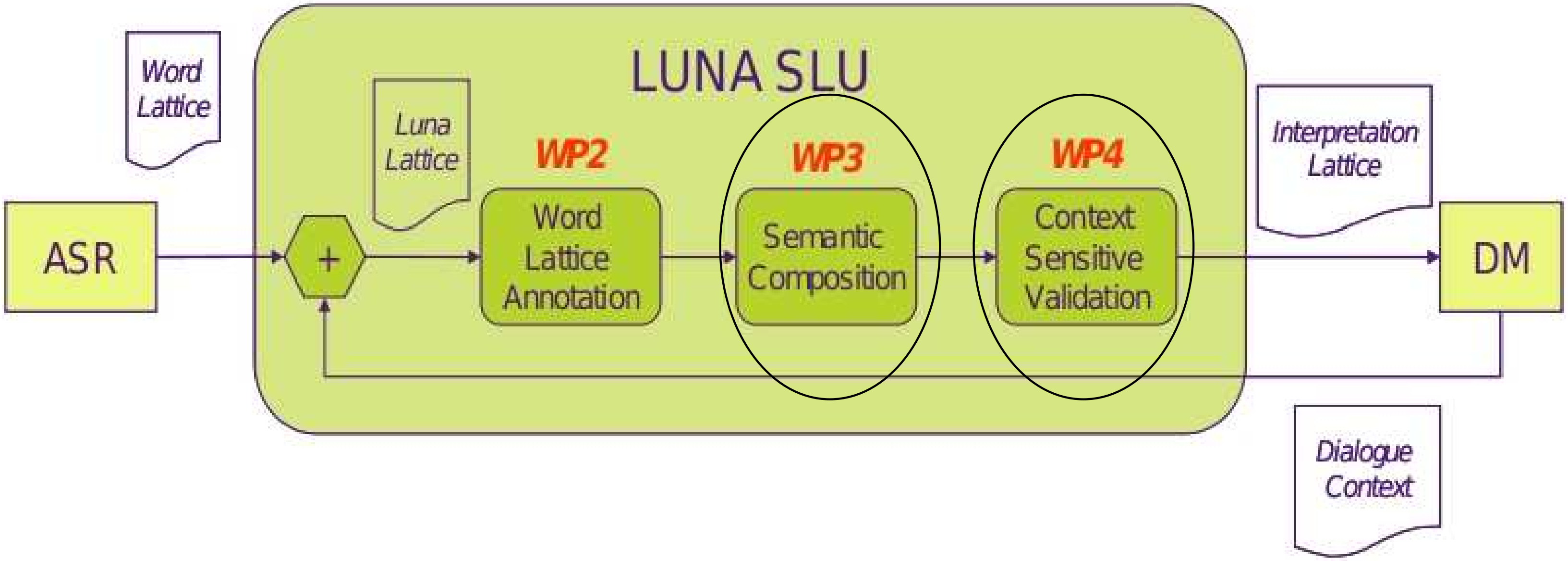
[Operator:] allora m'ha detto che [non riusciva]fe1 ad [accedere]fe2 [al computer]fe3 e le [manca]fe4 [la procedura]fe5

so, you have told me that you cannot access the computer, and that you need the procedure

```
frame : access
frame-elements : {user, hardware}
frame-set :{fe1, fe2, fe3}
   fe id:fe1 f-element: negation
   fe id:fe2 f-element: target
   fe id:fe3 f-element: hardware
frame : need
frame-elements : {user, requirement}
frame-set :{fe4, fe5}
   fe id:fe4 f-element: target
   fe id:fe5 f-element: requirement
```

Levels of annotation

Coreference level



Levels of annotation

- **We annotate different kinds of anaphoric relations like:**
  - Identity
  - Bridging : exploiting the relations and properties of the domain ontologies.
  - Set-element

- **Annotation scheme close to the one used in the ARRAU (AnaphoRa Resolution And Underspecification) project.**
  `http://cswww.essex.ac.uk/Research/nle/arrau`
  - Includes instructions for the annotation of associative descriptions.
  - A single interpretation is not required.

- **First step is the annotation of information status of the markables. They will be classified in new and `old/given`.**

- **If the markable is annotated with `given`, we add a pointer to the antecedent.**

Levels of annotation

- **If the markable is `new`, we annotate whether it is related to previous markables or not.**

- **In case of relatedness:**
  - the annotator points to the previous introduced markable
  - indicates the type of relation:
    - Set-relation
    - One of the relations and properties defined in the domain ontology.

- **Plural markables:**
  - refer to a set of objects already mentioned.
  - will be annotated with **`multiple_referents`.**
  - the annotator will add a pointer to each of the referents.

- **Ambiguity: a markable has two or more interpretations**
  - it will be marked as **`ambiguous`** and
  - the annotator will add a pointer to each of the possible antecedents.

Levels of annotation

Example: annotation of coreference

[Operator:] allora m'ha detto che non riusciva ad accedere [al computer]c1 e le manca [la procedura]c2

so, you have told me that you cannot access the computer, and that you need the procedure

```
<coref id=“c1” info-status=“given” … />
<coref id=“c2” info-status=“given” … />
```

[Caller:] esatto

exactly

[Operator:] allora avrei bisogno [dell' RWS]c3 [del PC]c4

so I need the RWS of the computer

```
<coref id=“c3”, info-status=“new”, related=“yes”,
single-related-phrase=“c1”, relation=“rwsOf” />
<coref id=“c4”, inf_status=“given”,
single-phrase-atecedent=“c1” />
```

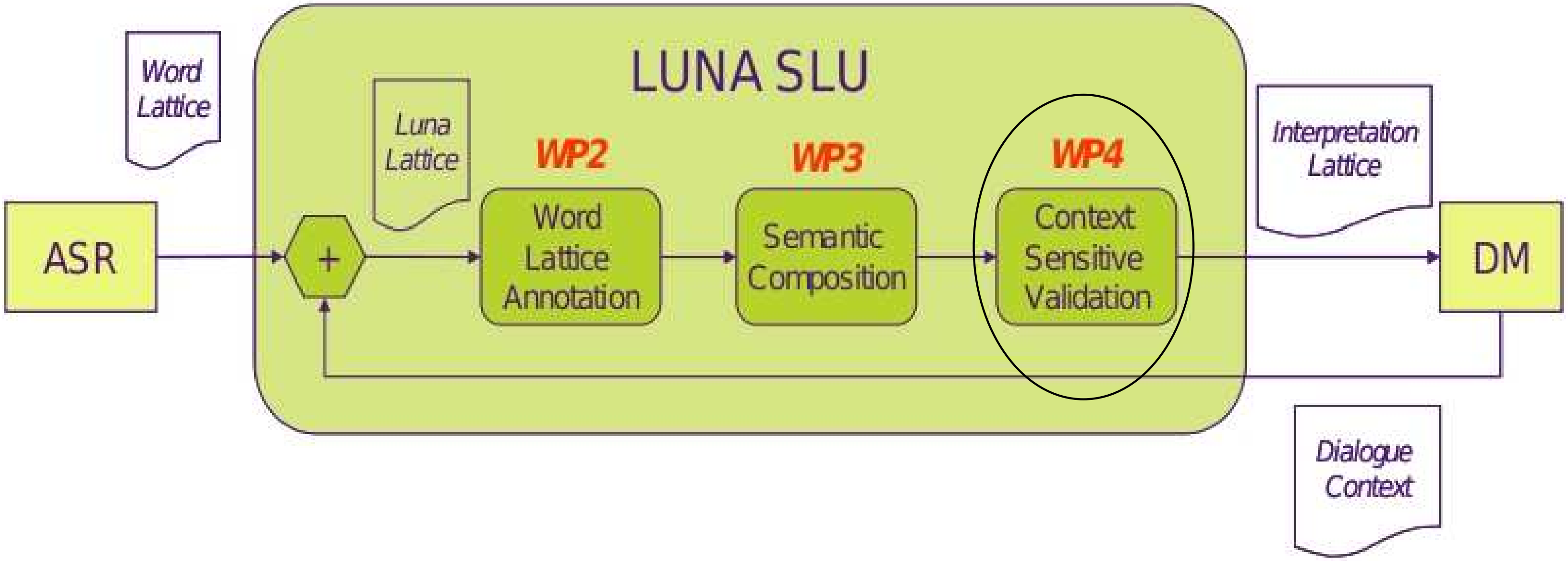
[Caller:] si allora [tredici zero ottantasei]c5

yes, 13 0 86

```
<coref id=“c5”, info-status=“new”, related=“yes”,
single-related-phrase=“c3” relation=“instanceOf”
/>
```

Levels of annotation

Dialogue acts



Levels of annotation

- **The segmentation of the dialogue turns in utterances is based on the annotation of the predicate structure.**
- **Each set of frame elements will correspond to an utterance.**
- **Additional to these utterances we can define other utterances without semantic content to encode opening/closings, accepts, etc.**
- **Annotation scheme partially based on the DAMSL**
- **Tageset of 9 dialogue acts that can be extended for individual application domains.**
- **Each utterance will be annotated with as many tags as applicable.**

Levels of annotation

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

Levels of annotation

Example: annotation of dialogue acts

[Operator:] allora m'ha detto che [non riusciva ad accedere al computer]u1 e [le manca la procedura]u2

so, you have told me that you cannot access to the computer, and that you need the procedure

```
u1: statement, info-request
u2: statement, info-request
```

[Caller:] [esatto]u3

exactly

```
u3: statement, answer
```

[Operator:] [allora avrei bisogno dell' RWS del PC]u4

so I need the RWS of the computer

```
u4: statement, info-request
```

[Caller:] [si]u5 [allora tredici zero ottantasei]u6

yes, 13 0 86

```
u5: accept
u6: statement, answer
```

Summary and next steps

We have talked about:

- **The steps of the semantic understanding process in LUNA**
- **The modules of the SLU toolkit**
- **The use of a corpus to build and evaluate the modules**
- **The annotation levels of the LUNA corpus**

Summary and next steps

And the next steps are:

- **Here and now.... discuss about the corpus and the scheme.**
- **Start the annotation of the data.**
- **Elaborate protocols for the control of quality based in:**
  - Statistical mesures of agreement.
  - Machine learning techniques to detect errors.
- **Elaborate protocols for the evaluation of the system prototypes.**