# WP2 - Semantic Parsing and Generation of Documents and Documents Components

*Claire GARDENT, Bikash GYAWALI,*
*Anastasia SHIMORINA*
***CNRS / LORIA***
*Samuel CRUZ-LARA*
***University of Lorraine / LORIA***

ITEA 3

# WP2

Goal: Provide tools and methods for:

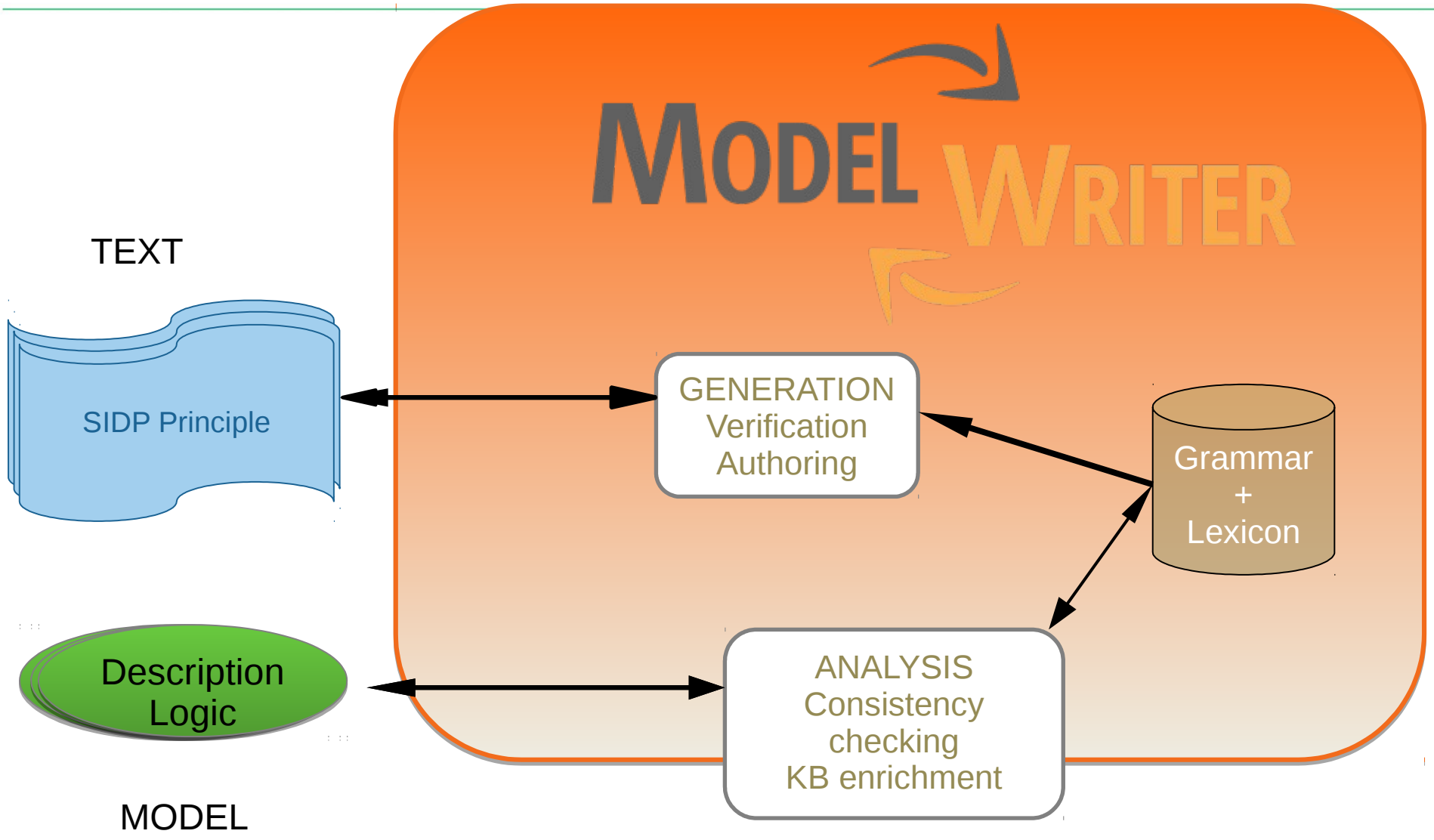
- Converting texts to models and models to text
- Annotating text fragments with model elements

Tasks:

- T2.1 Data Collection
- T2.2 Semantic Parsing
- T2.3 Natural Language Generation
- T2.4 Definition of a common target semantic language
- T2.5 Development of a Semantic Parser and of a Natural Language Generator

**Text <==> Model**

# Synchronising Text and Model

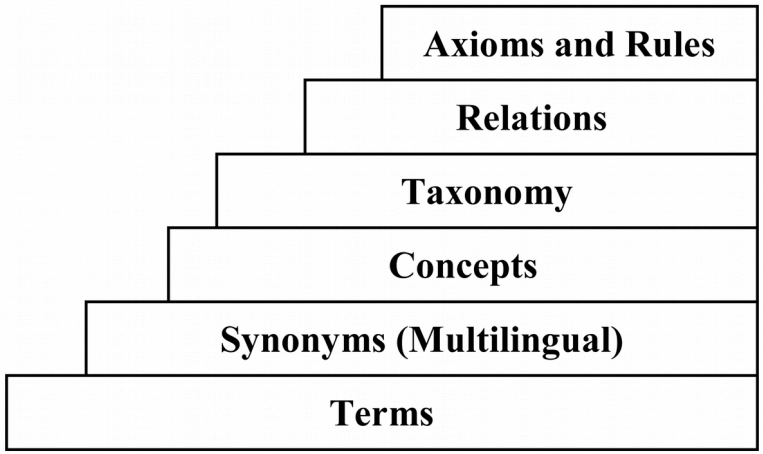


# Related Work on Ontology Learning

**Ontology Learning from Text**

[Mädche and Staab 2000, Volker et al. 2007, Tablan et al. 2006, Zouaq and Nkambou 2008]

==> *Restricted expressivity. Not applied to sentences (complex axioms).*



# Related Work on Semantic Parsing

## Deep parsing

[Currant et al. 2007, McCartney and Manning 2007 ]

*First Order Logic Representations close to initial text. Trained on newspaper text (Penn Tree Bank).*

==> *Not easily adaptable to Description Logic and SIDP text.*

## Domain Specific Semantic Parsing

[Ge and Mooney 2009,Wong and Mooney 2007]

==> *Require  parallel text-data training corpus.*

## Open Domain Semantic parsing

[Kwiatkowski et al 2010, Bordes et al. 2012, Kwiatkowski et al 2013, Berant et al., 2013,  Bordes et al. 2014, Wang et al. 2015]

==> *Restricted to questions. Require parallel question-answer training corpus.*

# Related Work on Text Generation

## Symbolic Approaches

[Dimitrios et al. 2007, Androtsopoulos et al. 2013, Power et al. 2010, Bontcheva et al. 2004 ]

==> *Heavily dependent on hand-written modules.*

## Machine Learning Approaches

[Wong et al. 2007, Belz 2008, Angeli et al. 2010, Chen et al. 2008, Konstas and Lapata 2012a and 2012b]

==> *Require parallel data-text training corpus.*

## Pattern-Based

[Duma et al. 2010, Blake et al. 2013, Schilder et al. 2013]

==> *Require large quantity of parallel or comparable text-data training corpus.*
==> *Limited Semantic Expressivity (set of RDF triples).*

# Contributions

## Reversible Processing: Text <--> Model

Analysis: SIDP Rule → Description Logic
Generation: Description Logic → SIDP Rule

Verification by generation

## Semantic Parsing of Complex Axioms

*Pipe shall be identified by labels*

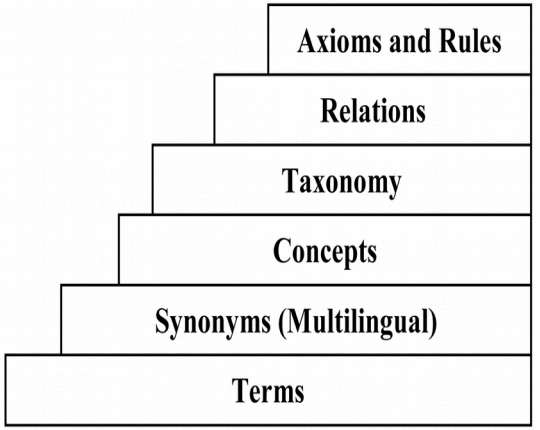
$$Pipe \sqsubseteq \exists identificationArg2^{-}.(Identification \sqcap \exists identificationArg3.Label)$$

## Executable Semantic Parsing on DL KBs

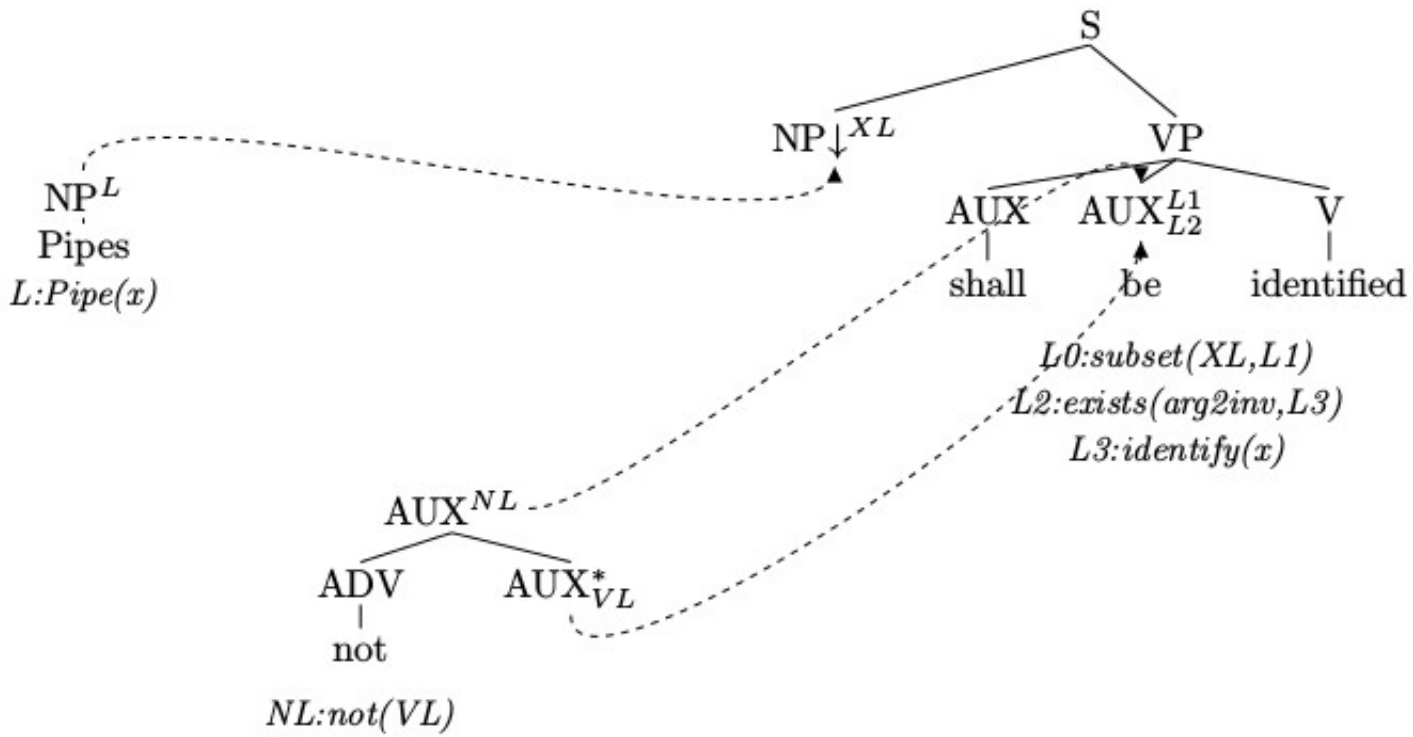
The output of semantic parsing is used to update a Description logic Knowledge Base and check the consistency of SIDPs (system installation design principle)

## Genericity

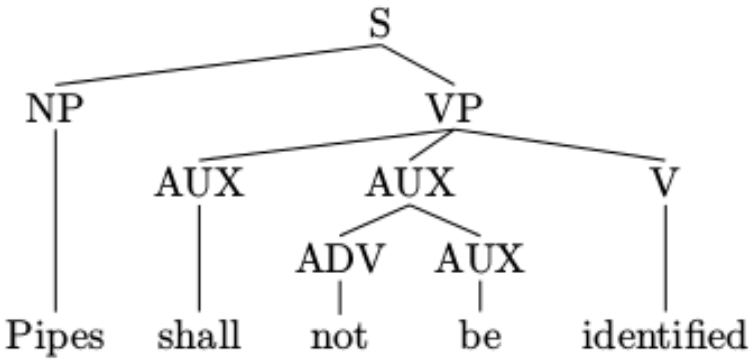
No training corpus required.
Adaptation to a new domain through grammar adaptation, extension or induction



# Grammar-Based Parsing and Generation



# Grammar-Based Parsing and Generation




*PL:Pipe(x) L0:subset(PL,NL) NL:not(VPL)*
*VPL:exists(identifyA2inv,VL) VL:Identify(x)*

$$Pipe \sqsubseteq \exists \neg identifyA2^{-}.(Identify)$$

# Semantic Variations

| | |
|---|---|
| **Logical Operators** | |
| Only S shall be used by O | $\neg S \sqsubseteq \neg \exists useA2^{-}.(use \sqcap \exists by.O)$ |
| S should be used by all O | $O \sqsubseteq \exists by^{-}.(Use \sqcap \exists useA2.S)$ |
| S shall not be used by O | $S \sqsubseteq \neg \exists useA2^{-}.(Use \sqcap \exists by.O)$ |
| **Word Order** | |
| S shall be used by O only | $S \sqsubseteq \neg \exists useA2^{-}.(Use \sqcap \exists by.\neg O)$ |
| All S shall be used by O | $S \sqsubseteq \exists useA2^{-}.(Use \sqcap \exists by.O)$ |
| **Arity** | |
| S shall be used | $S \sqsubseteq \exists useA2^{-}.(use)$ |
| S shall be used by O | $S \sqsubseteq \exists useA2^{-}.(use \sqcap \exists by.O)$ |
| S shall be used by O on PO | $S \sqsubseteq \exists useA2^{-}.(use \sqcap \exists by.O \sqcap \exists on.PO)$ |
| **Sentence Structure** | |
| S shall be used by O before entering connections | $(Use \sqcap \exists useA2.S \sqcap \exists by.O) \sqsubseteq \exists before.(Enter \sqcap \exists enterA2.Connections)$ |
| **Modifiers** | |
| S shall be used directly by O | $S \sqsubseteq \exists useA2^{-}.(Use \sqcap \exists directly.(\exists by.O))$ |
| S shall be used by O between C and D | $S \sqsubseteq \exists useA2^{-}.(Use \sqcap \exists useA3.(O \sqcap \exists betweenA1^{-}.(Between \sqcap \exists betweenA2.C \sqcap \exists betweenA3.D)))$ |

# Parsing, Generating and Checking Consistency

PARSING: CKY algorithm + Robustness mechanism to skip unknown words

GENERATION: Tabular algorithm + Polarity filtering

CONSISTENCY CHECKING:

– The Semantic Representations output by the parser are converted to Description Logic

Pipes should be identified by labels

$$l_1 : Pipe(x) l_0 : subset(l_1, l_2)\ l_2 : exists(identifyA2inv, l_3)\ \ l_3 : and(l_4, l_5)$$
$$l_4 : Identify(z)\ \ l_5 : exists(by, l_6) l_6 : Label(y)$$

```
SubClassOf(Pipe ObjectSomeValuesFrom(identifyA2inv
ObjectIntersectionOf(Identify ObjectSomeValuesFrom(by Label))))
```

– The resulting formulae are added to the AIRBUS KB
– Hermit is used to check for consistency

# Experimental Setup and Results for Parsing

Grammar: 52 trees

Lexicon: 10781 lexical entries

Parsing algorithm: CKY + Robustness mechanism to skip unknown words

Input: 991 System Installation Design Principles

| | Complete Parse | Partial Parse | Failure |
|---|---|---|---|
| Simple SIDP | 132 | 329 | 24 |
| Complex SIDP | 0 | 496 | 10 |
| All SIDP | **132 (13%)** | **825 (83%)** | 34 (3%) |

# Updating the Model using Parsing Results (All Parses)

| | | |
|---|---|---|
| CONCEPTS | Nb. of new Concepts | 667 |
| | Nb. of Existing Concepts | 79 |
| PROPERTIES | Nb. of New Properties | 98 |
| | Nb. of SIDP Axioms (from Parsing) | 957 |
| SIDP AXIOMS | Nb of Invalid Axioms | 61 |
| | Nb. of Redundant Axioms | 125 |
| | Nb. of Inconsistent Axioms | 20 |
| ALL | Nb. of added SIDP Axioms | 749 |
| | Total Nb. Of Added Elements | **1514** |
| | Nb. of Axioms in Initial KB | 12469 |
| | Nb. of Axioms in Enriched KB | 14650 |

# Generation Results

Grammar: 52 trees

Lexicon: 10781 lexical entries

Generation algorithm: Tabular + Polarity Filtering

Input: 957 Description Logic Axioms derived from the AIRBUS System Installation Design Principles

| | Success | Failure |
|---|---|---|
| Simple SIDP | 448 | 13 |
| Complex SIDP | 470 | 26 |
| All SIDP | **918 (96%)** | 39 (4%) |

# Verifying Parsing Results Using Generation

| BLEU | < 0.33 | > 0.32 and < 0.67 | > 0.66 |
|---|---|---|---|
| Complete Parses (S) | 1 | 0 | 131 (14%) |
| Complete Parses (C) | | | |
| Partial (Simple) | 143 | 117 | 69 |
| Partial (Complex) | 396 | 91 | 9 |
| All Parses | 540 (56%) | *208* ***(22%)*** | *209* ***(22%)*** |

Regenerating from the DL formula derived through parsing from an SIDP:

- Produces a sentence identical to the input SIDP for complete parses
- Produces a sentence highly similar to the input SIDP in 44% of of the cases for partial parses

# Perspectives and Future Work

- Improve **lexicon** construction using chunking
- Improve **coverage** on complex sentences including conditions
- **Querying** the KB (support for AIRBUS engineers)
- Improve **robustness** and **genericity** (experiment with deep learning approaches using data expansion techniques and sequence to sequence models)