Formal Languages and Natural Languages

Guy Perrier

1 - Introduction

1. Generalities about natural languages

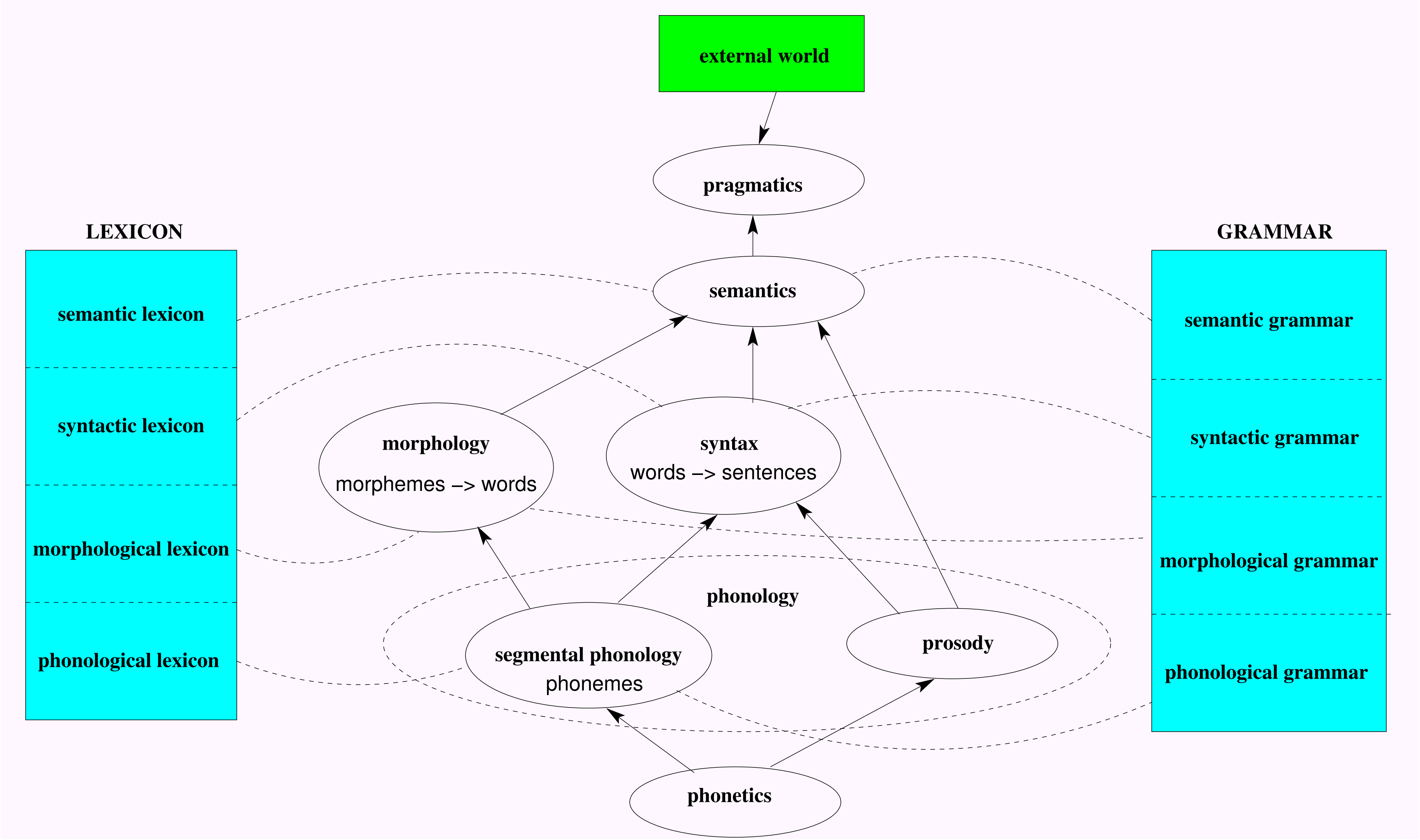
2. Formal languages and natural languages

3. Symbolic and statistical approaches of computational linguistics

1.1 - Generalities about natural languages

- To compute with natural languages requires to **model** them with **mathematics**.
- Natural languages are complex systems of signs aiming at expressing **meaning** from **sounds**.
- The basic signs are **words**, which combine to form **sentences** and sentences make up **discourses**.
- In a natural language, different dimensions interact together:
  - its specialized components: **phonetics**, **phonology**, **morphology**, **syntax**, **semantics** and **pragmatics**,
  - the **grammar** and the **lexicon**,
  - **world knowledge,** which is external to the language and expressed in the conceptual system.

1.1 - Generalities about natural languages



1.2 - Formal languages and natural languages

- The mathematical tool that comes to mind first for modeling natural languages is the **theory of formal languages**.

- A **formal language** L over a finite **alphabet** Σ of symbols is a part of the monoid Σ* of words built from Σ elements.

- The class of languages defined over Σ is equipped with **operations**: intersection, union (disjunction), concatenation, complementation, Kleene closure...

1.2 - Formal languages and natural languages

- If L is infinite, it is important to have a computation procedure for **recognizing** L , that is for deciding if any string from Σ* belongs to L. If such a procedure exists, L is said to be **recursive**.

- If there exists only a computation procedure for **enumerating** L, L is said to be **recursively enumerable**.

1.2 - Formal languages and natural languages

- In a formal language, symbols are concatenated to build the words of the language in a potentially infinite way. In a natural language, words are concatenated to build the utterances of the language in a potentially infinite way too.

- In a formal language, words are double side objects: (form, meaning) pairs. In a natural language, utterances are also double side objects; they are (sound, meaning) pairs.

- Chomsky’s results on the formalization of grammars for natural languages (1956) played a great part in the development of the theory of formal languages.

1.2 - Formal languages and natural languages

- **Ambiguity** is excluded from formal languages whereas it is continually present in natural languages.
    - An utterance is ambiguous if it has multiple alternative meanings.
    - According to the source of this multiplicity, we distinguish lexical, phonological, syntactic and semantic ambiguity.
    - Ambiguity is relative to the extent of the utterance and can be eliminated with the use of the context.

1.2 - Formal languages and natural languages

- Formal languages are **rigorously delimited** whereas the border between acceptable and non acceptable utterances in a natural language is **fuzzy** and **mobile**.
  - The border is related to the society using the language. It is difficult to determine if some very specific constructions are integrated in the grammar of the language by the society or if they are accidents of the language.
  - The border of a natural language evolves because its grammar evolves with the society.
  - The border is also related to the relevance or not of the distinction between the ability of **competence** of the mind and its ability of **performance**.

1.3 - Symbolic and statistical approaches

- For modeling natural languages and for computing with them, two approaches are possible: the **symbolic** approach and the **statistical** approach.

- The **symbolic** approach aims at modeling pre-existent linguistic knowledge (**grammars**) with symbolic systems, **formalisms** in other words.

- The **statistical** approach aims at **learning** quantitative linguistic information from **corpora** with statistical methods. For this, it uses more or less pre-existent linguistic knowledge attached to the learning corpora.

1.3 - Symbolic and statistical approaches

- Specific methods are designed for modeling and computing with the specific components of the language.

- In this way, natural language analysis can be decomposed in partial tasks: tokenization, segmentation, part of speech tagging, morphological analysis, parsing, lexical disambiguation, semantic analysis, discourse analysis, anaphora resolution, named entity recognition.

- Each task can be implemented according to the symbolic or to the statistical approach.

1.3 - Symbolic and statistical approaches

- For the analysis of utterances with the symbolic approach, it is easier to understand the output of analyses than with the statistical approach.

- The advantage of the statistical approach with respect to the symbolic approach is its robustness but its drawback is the dependency to the learning corpora.

- The construction of grammatical and lexical resources in the symbolic approach is very costly but the annotation of the learning corpora in the statistical approach is also very costly.

- A promising approach consists in combining symbolic and statistical methods.

- Bibliography: *The interaction between linguistics and computational linguistics* - T. Baldwin and V. Kordoni eds - Linguistic Issues in Language Technology - vol. 6, 2011