E. Cartier

CORPUS FOR LINGUISTIC RESOURCES BUILDING AND MAINTENANCE (CLRBM): SYSTEM ARCHITECTURE AND EXPERIMENTS

In this paper we present a system whose main goal is to link linguistic resources to corpora, so as to build and maintain their description. By linguistic resources we mean dictionary entries and their morpho-syntactic features, as well as more complex linguistic description such as “syntactico-semantic patterns”. First we present the overall obectives of this project and its linguistic background. Then we detail the architecture of the CLRBM system and its main components. The third part explains the linguistic theoretical model used in the project. The last part is devoted to detailing the procedures to follow the life cycle of words and meanings in discourses.

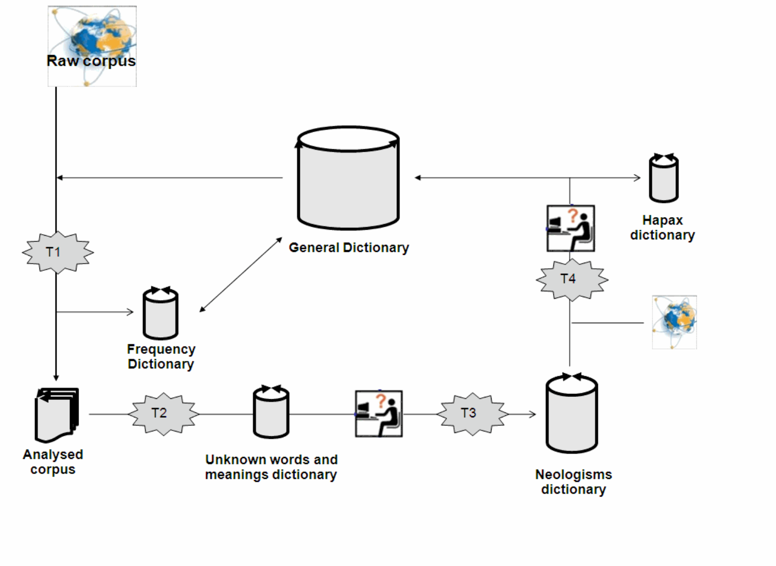
1. Objectives of the project

With the advent and availability of electronic corpora, we can now achieve one of the key goals of linguistics: to make linguistic resources descriptions stick to discourse utterances through Natural language Processing techniques. The overall idea is to build and maintain linguistic resources through a two-ways interaction with corpora: corpora are the sole linguistic data available (performance), whereas linguistic resources are built to represent the linguistic competence; as a result, an ideal system should continuously feed linguistic resources with corpora, to provide neologisms but also to follow the diachrony of linguistic usage that linguistic resources are meant to represent. Our system aims therefore at informing on the life of words and meanings.

Up to now, this linking is limited to identifying morphological neologisms: a morpho-syntactic analyser or concordancer or a statistical engine parses corpus from a prebuilt dictionary and identifies unknown words that are potential neologisms. Linguistic experts then decide on their integration in the dictionary. Such a system is one of the main tools used by lexicographers. But our goal is far more ambitious as it aims at identifying not only new words, but also at tracking the life of existing meanings.

2. Architecture of the system

The architecture (see figure) is composed of four main steps occuring continuously as a new corpus feeds the system.



Architecture of the CLRBM system.

The first step (T1) consists in analysing the feed corpus. This feed corpus consists of new feeds automatically and periodically retrieved with RSS Corpus Builder, a RSS reader that downloads RSS feeds and the textual content of newspaper articles. A morpho-syntactic analysis is then performed with a Linguistic Analyser called TextBox.

This software 1) completely externalizes linguistic resources; 2) enables linguistic resources to be freely defined as dictionary entries with features, but also as sequences of entries and patterns, i.e. grammars. In addition to the linguistic analysis, frequency counts are triggered for every word sequences and patterns in the dictionary.

The result of the first stage is:

- an XML analysed corpus notably indicating unknwon words;
- a frequency table where 0 frequency words are emphasized;

The second step (T2) is the human analysis of the annotated corpus. This concerns two types of words and meanings: first, unknown words and meanings; a user-interface presents to the user unknown words and unknown structures, and the expert can either add them to the dictionary, either, for the meanings, modify an existing structure to include the new one; secondly, the analyser can and will fail in several ways: bad segmentation, bad syntactical analysis essentially. All these problems are presented to the user who can, from the examples, correct the linguistic rules and entries interactively; as for unknown words, he can directly validate some tokens and make them part of the dictionary for the next analysis. He can also save some named entities and domain-related vocabulary into specific dictionaries.

The system is iterative and enables to build linguistic resources and then maintain them.

The third and fourth steps (T3, T4) concern exclusively neologisms. They permit the neologisms to be registered in a specific dictionary and then their life cycle to be followed through periodic queries on the web. After a given time frame, the user can include neologisms into general or specific dictionaries, or remove them from the neologisms list if he considers them as hapaxes.

3. Linguistic resources

In this part, we present the dictionaries and their structures used in the project. They consist of two main dictionaries: in the first one, words and frozen phrases are described from the morpho-syntactical point of view; the current project focuses on the French language and we use a dictionary called Morfetik which is the most comprehensive dictionary for NLP at the moment; in the second one we find three different linguistic units described: arguments, predicates and actualisateurs. Arguments are semantic units denoting an object, and we give for them their superclass, their class, as well as indication of the domain(s) to which they apply; Predicates are the core linguistic units forming sentences, whether represented by verbs, adjectives or nouns. Predicates are described following a complex scheme[1]. Actualisateurs are composed of determiners, support verbs and schemes enabling predicates and arguments to be included in discourses.

4. Neologisms, words and meaning life cycle

In this part, we will first focus on the neologism life cycle: so as to follow the life of neologisms and their intergration into the lexicon, we set up a procedure retrieving their frequency from web pages, through search engines. We present the procedures implemented and give first results. We also present a more ambitous program whose goal is to detail the general profile of meanings, through a combination of statistical measures and linguistic analysis. Theses procedures make it possible to retrieve from discourses instances of a given meaning (expressed as a syntactico-semantical scheme) and to follow the evolution of the stereotypical usage. This last project, yet in the modelisation phase, will be implemented into the system in the near future.

[1] See for details: *Buvet P.-A., Grezka A.* Les dictonnaires électroniques du modèle des classes d'objets // Langages. 2009. N. 176. P. 63–79.