An Experimental Constructional Database: The MorTAL Project

Nabil Hathout,[1] Fiammetta Namer,[2] and Georgette Dal[3]
[1] *ERSS, CNRS, and University of Toulouse–Le Mirail*
[2] *LANDISCO, University of Nancy 2*
[3] *SILEX, CNRS, and University of Lille 3*

1. Introduction

1.1. MorTAL or: Why a constructional database?

The constructional database[1] which is the object of this presentation is part of the MorTAL project, a research project funded for a period of 3 years by the French Ministry of Research.[2]

The French language lacks specifically constructional databases, certainly due to the alleged irregularity of the constructional morphology of the language. However, natural language processing (NLP) and information retrieval (IR), in particular, can benefit from this type of resource, for the following usages, among others:

- in document retrieval, to improve the filtering of documents resulting from a query
- in automatic analysis systems based on using unification-based grammars (TAG, LFG, HPSG), in order to automatically and dynamically (according to need) obtain the lexical entry of a constructed word C from its base B, and reduce the size of the lexicon used by the grammar

[1] As in the theoretical model developed in Lille (France) by D. Corbin and her team, we prefer the term “constructional” to “derivational,” which does not always imply a single notion.

[2] MorTAL brings together the co-authors of this article, as well as Christian Jacquemin who deserves our deepest thanks for his careful rereading of this presentation.

- in information retrieval when the knowledge of the constructional links between lexical units partially solves the problems related to term variants (cf. especially Jacquemin 1999)
- in text comprehension, when the semantic relationship that exists between the constructed words and their bases can be used

Therefore, from an NLP and IR point of view, MorTAL fills a gap.

From a theoretical linguistics point of view, it is an ambitious enterprise in that developing a computer program capable of automatically analyzing constructed French words in a way that is linguistically motivated proves that the construction of those words is governed by rules, and thus that the reputation from which French constructional morphology suffers is unjustified. In this way, MorTAL is part of the new scientific paradigm of corpus linguistics in which corpora can validate linguistic hypotheses when they are applied to massive quantities of data, or if necessary, allow the hypotheses to be amended.

1.2. A lexicon enriched by constructional and semantic features: Subject description

The database under development[3] is designed as a large lexicon of approximately 70,000 lexical units, essentially combining the major lexical entries that appear in the *TLFnome*[4] and in the *Robert électronique* (1994). At this time, the suffixes *-able*, *-ité*, *-et(te)*, *-is(er)* and *-ifi(er)*, as well as the prefixes *dé-* and *in-* have been (almost) completely studied and implemented, creating a database of approximately 5,000 lexical entries.

Eventually, our database will take the form of a French language lexicon enriched by constructional and semantic features (hereafter LECSF). Each entry will include the following information:

1. the lemma (i.e., the form that conventionally subsumes all its inflectional variants)
2. its grammatical category
3. its constructional analysis presented as a tagged and bracketed schema
4. a repetition of the information from the previous field, presented more clearly; i.e., for affixed and converted words, the entry, then its base, and

[3] For the first stages, see Dal et al. 1999.

[4] The *TLFnome* is a lexicon of inflected forms compiled at the INaLF and based on the nomenclature of the *Trésor de la Langue Française*. It presently contains 63,000 lemmas, 390,000 forms, and 500,000 entries. It has been complemented by a second lexicon of 36,400 additional lemmas taken from the index of the *TLF*.

the base's base when necessary, etc. until it reaches the entry's primitive – a unit that cannot be decomposed

5. a gloss that illustrates the semantic results of the application of the most peripheral constructional operation

Some of these fields can be left blank. Indeed, none of the entries that appear in the two reference corpora are rejected *a priori*; however, among those entries, not all are constructed.

For example, all fields of an adjective like *inarticulable* 'inarticulatable' are filled because the adjective is constructed:

(1) inarticulable/ADJ: [in [[articul(er) $_{\text{VERBE}}$] able $_{\text{ADJ}}$] $_{\text{ADJ}}$][5], (inarticulable, articulable, articuler), "qui n'est pas articulable" 'that which is not articulatable'

However, the fifth field is left blank for a non-constructed verb such as *articul*(*er*) 'to articulate,' because the verb does not have a constructed meaning:

(2) articul(er)/VBE: [articul(er) $_{\text{VERBE}}$], (articuler)

Eventually, all of the information described will be linguistically verified. It is for this purpose that the DériF system was developed. However, given the time-cost of such a verification, we will use a second computer program, DéCor, which was also developed for this purpose. DériF is a system that implements linguistic hypotheses, excluding the restrictions involved in the implementation. By the end of 2002, DériF will have provided a complete analysis of approximately 20,000 French lexical units.[6] DéCor is a learning-based analysis system that uses the pairing of formally similar lexical units that belong to the same reference. By 2002, DéCor will allow us to offer a preliminary analysis for an as-yet uncalculated number of the 50,000 units in our corpus that are not treated by DériF.

The rest of this paper will be devoted to a detailed presentation of these applications. After giving a progress report on the various points of view that

[5] ADJ: adjective; VERBE/VBE: verb; NOM: noun; FWD: foreign word.

[6] This number corresponds to an estimate of the number of derivatives produced by the constructional operators that we have decided to implement following linguistic rules. Together, they give good coverage of the constructed lexicon. These suffixes are *-(a)tion* (*liberation*), *-(at)eur* (*voyageur* 'voyager'), *-able* (*abordable* 'accessible'), *-age* (*lavage* 'washing'), *-aire* (*bancaire* 'banking'), *-al* (*adjectival*), *-et(te)* (*fillette* 'little girl'), *-eux* (*neigeux* 'snowy'), *-ifi(er)* (*acidifier* 'to acidify'), *-is(er)* (*budgétiser* 'to budgetize'), *-ité* (*sévérité* 'severity'), *-oir(e)* (*arrosoir* 'watering can'), and the prefixes *dé-* (*décapsuler* 'to take a cap off') and *in-* (*inabordable* 'inaccessible').

The complete article appears in *Many Morphologies*, edited by Paul Boucher, published in 2002 by Cascadilla Press. To purchase *Many Morphologies* or to find out more about the book, please visit our web site at http://www.cascadilla.com/manym.html or contact us:

Cascadilla Press
P.O. Box 440355
Somerville, MA 02144
USA

phone: 1-617-776-2370
fax: 1-617-776-2271
e-mail: sales@cascadilla.com
http://www.cascadilla.com