French frozen verbal expressions: from lexicon-grammar tables to NLP applications

Laurence Danlos, Benoît Sagot, Susanne Salmon-Alt

Frozen verbal expressions are rarely accounted for in NLP systems, both because of their linguistic complexity and because of the rareness of appropriate lexical resources. In this paper, we aim to indicate how these problems can be tackled. Indeed, we show that an NLP-oriented representation of the information contained in lexicon-grammar tables for frozen verbal expressions allows to deal with this phenomenon in TAG or LFG grammars and parsers.

1 Maurice Gross' lexicon-grammar for frozen verbal expressions

Following the principles of the lexicon-grammar for simplex French verbs, Maurice Gross [1] has classified around 25.000 frozen verbal expressions[1]. He notes N_i a position for a verbal complement (e.g., N_0, $N_1 \ldots$). Frozen verbal expressions are characterized by the fact that at least one of these N_i is replaced by C_i, which denotes a constant (frozen complement)[2]. Other (free) complements remain denoted as N_i. For example, the verbal expression *prendre en compte que P* ("to take into account the fact that S", e.g., *Jo prend en compte que Luc dort*) corresponds to the structure $N_0\ V\ (Qu\ P)_1\ Prep\ C_2$. The free positions of this structure correspond to the simplex verb structure $N_0\ V\ (Qu\ P)_1$, which is the basic structure for table 6 (*Jo pense que Luc dort*). For this reason, the table for frozen verbal expressions with a $N_0\ V\ (Qu\ P)_1\ Prep\ C_2$ structure is named C6.

Hence, this classification principle builds a parallel between a simplex verb (such as *penser*) and a frozen verbal expression with a non-trivial internal structure (in *prendre en compte*, a verb followed by a prepositional phrase in which the noun has no determiner). In some expressions, such as *prendre en compte*, this internal structure may be realized in a non-continuous way, which is a source of difficulties for parsing:

(1a) Jean prend en compte cette information (1b) Jean prend cette information en compte (Table C8)
(2a) Jean prend acte de ce que Marie est partie (2b) *Jean prend de ce que Marie est partie acte (Table C6)

In some tables, the frozen complement is not exactly a constant, since it might include:

- a variable possessive determiner which refers to another complement N_i (this reference is indicated by an exponent i): *casser (*sa^0 *pipe)*$_1$, *dicter (*sa^2 *conduite)*$_1$ *à* N_2,
- a variable reflexive pronoun: *avoir* N_1 *(sur* soi^0*)*$_2$,
- a mandatory variable modifier[3], denoted by $(C\ de\ N)_i$: *battre (le rappel de [ses amis])*$_1$.

2 NLP lexical encoding

In order to make available those fine-grained linguistic descriptions of frozen expressions to NLP analysis, the current database has to be converted into a standardized NLP lexicon. The main idea behind the current lexicographic practice — as well as the recent normalisation proposals of LMF [2] — is to reuse as much as possible lexicographical components for the description of simplex forms, and to combine them in order to characterize appropriatly complex linguistic expressions. In our case, we propose to adapt and evaluate the LMF proposals to the description of (1) the internal structure and (2) the subcategorization properties of frozen verbal expressions.

[1] These expressions should not be confused with light verb expressions or frozen adverbials which have been studied in other works.

[2] Since these C_i are frozen N_i, the same structural constraints on syntactic positions as for simplex verbs do apply.

[3] The modifier is generally introduced by the preposition *de*. However, if the (frozen) head denotes a part of the body, it should be noticed that the cliticization is possible, and realized with the dative clitic pronoun (*lui*).

With regard to the current LMF proposal for encoding the internal structure of multi-word expressions — i.e., embedded /lemmatizedForms/ — the fine-grainedness of the linguistic data in the lexicon-grammar tables (sec. 1), but also the requirements of an optimized lexical representation as input to a robust parser (sec. 3), lead us to introduce the following additional lexicographical features:

- a replacement of /lemmatizedForm/ by an underspecified /form/ element, since parts of expressions might be inflected forms as well. In general, any inflectional feature should be available for the description of the components. Furthermore, the extensional description of the /form/ component might be replaced by using simply a pointer to simplex entries,
- a possible hierarchization by means of embedded /form/ elements, in order to express syntactic dependency between the components,
- the possibility to replace a concrete /form/ with an extensional or intensional paradigm (i.e. lists or reference to classes) of lexical or morphosyntactic variants, as the cases mentioned in section 1.

The second aspect, the description of subcategorization properties, relies on the parallel between constructions containing one or more frozen constants (i.e., table C6) and corresponding "free" constructions (table 6). Indeed, since the form and variation of the frozen constants are described as parts of the lexical entry (cf. supra), the encoding of subcategorization properties are shrinked down to regular and standard representations of syntactic frames, such as proposed in LMF.

3 Interface with grammars and parsers

Once the lexical information on frozen verbal expressions has been formalized such as presented above, it needs to be interfaced with the grammars used in parsers. We will outline how this can be achieved in two different formalisms: LTAG and LFG. Both mechanisms are currently being implemented in large-scale parsing systems for French.

In LTAG, which is a lexicalized formalism, frozen verbal expressions can be encoded, as other multi-words expressions, thanks to elementary trees which have a main anchor and several co-anchors. The internal structure of the expression is hence encoded in the structure of the elementary tree, whereas its external subcategorization frame, as for simple verbs, corresponds to substitution or adjunction nodes (see Figure 1). Of course, the whole expression will correspond to a single node in the derivation tree, which is semantically satisfying.

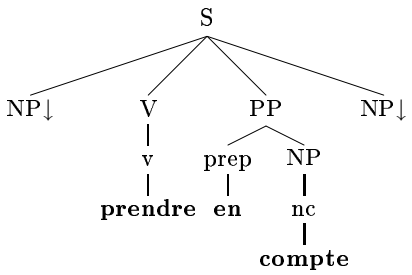

Figure 1

In LFG, the situation is not so easy, because it is not possible to encode directly structures in the lexicon. However, we intend to implement the same idea as for LTAG in our LFG parser SxLfg, thanks to a two-step mechanism:

- In the standard parsing step, frozen complements are considered as multi-word units, and checks are performed in the functional structure to guarantee that the elements of a given expression are used in a consistent way[4],
- In a second step, after parsing, a substitution is performed that replaces these multi-word units by structures describing their internal structures of the expression.

References

1. Gross, M.: Les phrases figées en français. L'information grammaticale **59** (1993) 36–41
2. ISO/TC 37/SC 4 N130 Rev.9: Language resource management — Lexical Markup Framework (LMF). (2006)

[4] This involes special rules, that deal with these special multi-word tokens. Hence, part of the internal structure of the expression is encoded in the rule, as is the external sub-categorization frame