Description of French Idioms in a Lexical Database

Sara-Anne Leblanc

OLST – Département de linguistique et de traduction,
Université de Montréal
C.P. 6128, succ. Centre-ville,
Montréal (Québec) H3C 3J7 Canada
sara-anne.leblanc@umontreal.ca

Abstract

French expressions like ‘FRUIT DE MER’ ‘seafood’, ‘MONTER UN BATEAU’ [à qqn] ‘to fool someone’ and ‘À PROPOS’ [de qqch] ‘about something’ are idioms. Set phrases like these are considered in Meaning-Text Theory (MTT) as lexical units and are therefore described by an autonomous lexical entry in the dictionary. However, since each phrase consists of more than one word-form, these lexical units display complex behavior which has not yet been rigorously described in the Meaning-Text lexicography. In this paper, we present a MTT treatment of idioms focusing on their presentation in the dictionary, and propose tools for constructing a lexical database of idioms, which can efficiently represent form flexibility and combinatorial particularities of French idioms.

Keywords


1 Introduction

In recent years, several large-scale research projects concerning the lexicon have been launched. Lexical databases aiming to describe specific languages are being set up within various theoretical frameworks. Representative examples of English lexical database projects are WordNet (Fellbaum, 1998) and FrameNet (Fillmore & al., 2001). We can mention, for the French lexicon, the lexicon-grammar tables (Gross, 1975), the work carried out by the DAFLES (Selva & al., 2002) as well as the Fontenelle’s database (1997), which models the lexicon of several languages, including French.¹

¹ We will introduce the Meaning-Text inspired lexical projects developed at OLST-Université de Montréal in Section 3.2.
The study of fixed phrases is also a main interest for today’s researchers. These particular lexical units cause problems for NLP researchers (Sag & al., 2002) as well as for theoretical linguists and lexicographers (Mel’čuk, 1995; Čermák, 2001). In spite of these efforts, there remains a crucial need to develop a coherent set of tools designed to efficiently describe idioms. On one hand, idioms are lexical units as well as lexemes and as such should be modeled by separate lexical entries. They constitute complex lexical units which should be taken into account by major lexical databases. Still, idioms have some characteristics that wordforms do not. We therefore should study their specific characteristics in order to improve our lexical description of them.

In this paper, we will first define idioms and offer an overview of their treatment within an MTT framework. We will then present our current work on a new lexical database of idioms, the BDLoc. Finally we will describe some formal and combinatorial peculiarities of those French idioms which we believe should be included in their lexical entries.

2 What is an idiom?

An idiom is a non-free phrase considered to be a lexical unit. It is a multilexemic expression whose meaning cannot be inferred as the regular sum of the meanings of its constituent lexemes. For example, in French, the nominal idiom FRUIT DE MER means ‘seafood’, and not lit. fruit of the sea, and the verbal idiom MONTER UN BATEAU [à qqn] means ‘to fool someone’ and not lit. to assemble a boat. These idioms also have functional autonomy and some internal cohesion, the criteria most commonly accepted by linguists.

In the Explanatory and Combinatorial Lexicography (ECL), the lexical module of the MTT, the term phraseme is used to designate all fixed phrases. There is a large subset of phrasemes, collocations or semi-phrasemes (Mel’čuk, 1995; Mel’čuk, Clas & Polguère, 1995), which we will not examine in this paper because they are not lexical units and therefore should not be represented by separate lexical entries.

3 Treatment of idioms within the MTT framework

In this section we will briefly explain how idioms are represented in a Meaning-Text model and in Explanatory and Combinatorial Lexicography, and the implications of this treatment for the characterization of idioms in the context of dictionary compilation.

---

2 This is not a criterion for the definition of idioms, since a lot of idioms have a relative internal cohesion. For example, it is normal to insert adverbial modifier “inside” the verbal idioms. We will discuss the form particularities of French idioms in Section 5.2.
3.1 Linguistic representations in the Meaning-Text model

The Meaning-Text model operates with a series of linguistic representations. We are interested here in three levels of representation, that is to say DSyntR, SSyntR and DMorphR. An example sentence using the idiom MONTER UN BATEAU [à qqn] is shown in (1). Figure 1 shows the DSynts and SSyntS for example (1).

(1) Jean monta un bateau à Julie. ‘Jean fooled Julie.’

Figure 1: DSynts and SSyntS of (1)

The DSynts shows all full lexical units of the sentence. The idiom is represented by only one node, and its DSynt actants are visible. The SSyntS treats fixed phrases and free phrases in a similar way. There are no idioms in this representation, only surface lexemes. Idioms, however, differ from free phrases in their restrictions of syntactic transformations and their blockage of syntactic rules. These constraints and special syntactic properties are included in the lexical entry of the idiom.

In the DMorphS we can see the linearization of the sentence, as in the example in Figure 2. Information specifically linked to this level of representation, namely the blockage of grammemes, grammatical constraints and linearization of the dependents (inside or outside the “frame” of the idiom) are also included in the lexical entry of the idiom.

Figure 2: DMorphS of (1)

---

3 Since we aim to emphasize properties specific to idioms, we will not include other levels of representation. Although the SemR of an idiom is of course fundamental, it has roughly the same characteristics as the SemR of a lexeme, since it models the meaning of a lexical unit (and all its paraphrases) by its minimal decomposition in simpler semantemes.

4 Only the core structure of the representation, i.e. the DSyntS, is represented here. We will apply the same procedure to the other levels of the model.
3.2 The ECD and the DiCo

The *Explanatory and Combinatorial Dictionary (ECD)* is a formal theoretical dictionary of a language, elaborated according to principles of ECL (for French, see Mel’čuk & al., 1984, 1988, 1992, 1999). The *DiCo (Dictionnaire de Combinatoire)* is a computerized version of the *ECD*, built in the form of a lexical database (Polguère, 2000). It resembles to some extent a simplified *ECD* which concentrates on restricted lexical co-occurrence and semantic derivatives. Igor Mel’čuk and Alain Polguère have been developing this project for several years at the Meaning-Text Linguistics Observatory (*OLST*).

As already mentioned, an idiom necessarily has a lexical entry of its own, of which it is the headword. Therefore, the *ECD* and the *DiCo* have lexical entries that describe idioms. We find 181 vocables in the *ECD* and 28 vocables in the *DiCo* that are idioms.

4 Building a lexical database of French idioms

The *BDLoc (base de données de locutions)* is a database we have developed to complement the *DiCo*, and with it we aim to rigorously describe all the idioms cited in the main database. We recovered them from all fields of the *DiCo*, for a total of 1452 vocables. The database, which is now under construction, encompasses nearly 1800 lexical units (taking polysemy into account).

A lexical entry in the *DiCo* has a field tagged “ph” (phraseme), where all the idioms having the headword in their form are encoded (as in the general dictionaries). These idioms constitute about half of the *BDLoc*. We also extracted all idioms encoded in the *DiCo* as lexical function values, from the fields “syn” (synonym) and “fl” (lexical function). These idioms can be semantic derivatives or collocates of the headword. Finally, we recovered all idioms that are government pattern values. These are found in the *DiCo* field “tr” (*tableau de régime* ‘government pattern’) when they are related to the expression of the headword actants. They may also be found in the field “fl”, when they correspond to the government pattern value of a collocate.

Table 1 illustrates a sample of the idioms recovered from the *DiCo* entry of *BATEAU* (‘boat’), and the location from which they were extracted. When relevant, we also mention their encoding (i.e. the lexical function that link the collocate to *BATEAU* or the government pattern encoding).

---

5 The lexical database can now be consulted directly on the web, at http://olst.ling.umontreal.ca/dicouebe.

6 We just took into account the lexical entries that are completely described and verified. There are a lot more of idioms partially described in the *DiCo*.
Description of French Idioms in a Lexical Database

<table>
<thead>
<tr>
<th>French idiom and meaning</th>
<th>DiCo field</th>
<th>Encoding (when applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MONTER UN BATEAU ³ ‘to fool someone’ lit. to build a boat</td>
<td>ph</td>
<td></td>
</tr>
<tr>
<td>COQUILLE DE NOIX ³ ‘small boat’ lit. nutshell</td>
<td>fl</td>
<td>{AntiMagn} //coquille de noix_</td>
</tr>
<tr>
<td>LARGUER LES AMARRES ³ ‘[a boat] to leave the port and begin to move’ lit. to release the mooring ropes</td>
<td>fl</td>
<td>{IncepFact0} <em>larguer les amarres</em></td>
</tr>
<tr>
<td>À BORD ³ ‘on board’</td>
<td>tr</td>
<td>Y = II = avec N <em>à bord</em></td>
</tr>
<tr>
<td>EN DIRECTION ³ ‘in direction of’</td>
<td>fl</td>
<td>{Fact0} voguer ([<em>en direction</em> de N])</td>
</tr>
</tbody>
</table>

Table 1: Idioms extracted from the DiCo entry for BATEAU

5 Lexical Characteristics of Idioms

Mel’čuk (1995, 168) raises a fundamental question of “what should be stated about the given idiom [in the dictionary article for it] to be correctly selected and used in speech”. Since our focus is always on their synthesis or production by the speaker, is it truly necessary to describe the behavior of an idiom, knowing that we will have to look it up in corpora to validate our intuitions? We would argue that it is. An ECD is supposed to be exhaustive on the level of each individual lexical unit, and our lexical database is directly inspired by the French ECD. Instead of postulating that idioms are indivisible lexical units that cannot be subject to syntactic transformation, internal modification or paradigmatic lexical permutations, we prefer to verify their behavior as we construct their lexical entries. Our lexicographic methodology, perhaps more empirical than theoretical, allows us to capture the relevant lexical and grammatical facts about idioms.

The description of idioms involves lexical particularities that lexemes do not have, due to their multiword character. It will be necessary not only to have the headword, but also the partial SSynt tree of each idiom in the lexical entry (Mel’čuk, 1995; Kahane, 2001) in order to allow a good description of these particularities. As any lexical unit, idioms are linguistic signs, thus we will present some issues in the description of their meaning, their form and their combinatorial properties in the next sections.⁷

5.1 Meaning description

In principle, we should provide a proper definition for each idiom (see Mel’čuk, Clas & Polguère, 1995, for the ECL principles of definition’s construction). However, since the DiCo does not furnish a full definition but merely a semantic label and propositional form for each

⁷ We would like to point out one general dictionary that seems to present a fairly good description of English idioms, taking into account variations and syntactic behavior, the Collins Cobuild Idioms Dictionary (Seaton & Macaulay (ed), 2002).
lexical unit (Polguère, 2000), the representation of the meaning is similarly limited in the BDLoc. The improvement of this aspect of the database is currently under evaluation.

5.1.1 Illustrations

The verbal idiom "MENER EN BATEAU" (‘to fool someone’ lit. carry out in a boat) has the propositional form (translated) ‘individual X ~ individual Y (concerning Z)’ and receives the semantic label <agir négativement vis-à-vis de qqn.> (‘to act negatively towards someone’).

A thornier case is "LE TORCHON BRÛLE" [entre N et N] (‘there is a quarrel between N and N’ lit. the cloth burns) which is more difficult to treat with existing tools. The semantic label is supposed to be of the same part of speech as the lexical unit, but this idiom is a complete clause. The semantic label should therefore be a paraphrase of the idiom. We choose the propositional form (translated) ‘~ between individual X and individual Y (over Z)’ and the semantic label <il y a désaccord> (‘there is dissension’).

5.2 Description of the Form

The basic form of a lexical unit is represented by the headword of the entry. The partial SSynt tree also presents the form of the idiom, with the dependency relations between its constituent surface lexemes. When there is formal variation (not related to semantic or combinatorial properties), it should be indicated in the partial SSynt tree, with the headword as the most common form. The variant forms would also be given lexical entries consisting simply of cross-references to the main lexical entry of the idiom.

5.2.1 Illustrations

The example idiom "MENER EN BATEAU" has no variant forms. Its lexical entry, shown in Figure 3, should contain the following partial SSynt tree with no further indications.

Figure 3: Partial SSynt tree of "MENER EN BATEAU"

---

8 This is problematic, since semantic labels are supposed to describe multiple lexical units. We thus created some prototypical clausal semantic labels, after the model <il y a X> (‘there is X’).

9 In the partial SSynt tree, the syntactic places of lexical unit actants are indicated by white, non-polarised nodes (Kahane, 2001).
Other French idioms, however, do show formal variation. The clausal idiom “Quand on parle du loup on en voit la queue” (lit. when one speaks about the wolf, one sees its tail), which refers to a person’s entry into a room while someone else is talking about him or her, has an omissible part: on en voit la queue. Its headword will be the complete form, but the partial SSynt tree will show its possible modifications, with nodes that could be omitted in dotted lines (see Figure 4).

It is sometimes possible to substitute a constituent lexeme of the idiom by a near synonym, and this should be noted in the lexical entry. For example, for “Remuer le couteau dans la plaie” (‘to mentally torture someone by bringing up a sensible subject’ lit. to stir the knife in the wound), we found all the following variations with same meaning: “Retourner le couteau dans la plaie” (lit. to turn again the knife in the wound), “Remuer le fer dans la plaie” (lit. to stir the steel in the wound), “Tourner le fer dans la plaie” (lit. to turn the steel in the wound). In the BDLoc, as said before, we choose the most frequent form as the headword of the main lexical entry. The partial SSynt tree of this idiom, as shown in Figure 4, also contains all the possible variations.

![Figure 4: Partial SSynt trees of “Quand on parle du loup on en voit la queue” and of “Remuer le couteau dans la plaie” and its variants](image)

**5.3 Combinatorial properties description**

**5.3.1 Part of speech (Mel’čuk, 2006)**

Lexical units have two parts of speech, one in DSynt and one in SSynt. We indicate the SSynt part of speech of the lexemes in the DiCo. However, idioms do not exist at the SSynt level (where, as indicated above, they are represented like free phrases). Since idioms have no true SSynt part of speech, they receive their head’s SSynt part of speech, with the exception of clausal idioms such as “Quand on parle du loup on en voit la queue” and “Le torchon brûle” above. These are considered as complete clauses and tagged with a special part of speech “proposition” (‘clause’).
We also specify cases in which the syntactic behavior of an idiom does not match that of its head. For example, "TAMBOUR BATTANT" (lit. beating drum), formally a nominal idiom, functions as an adverb meaning ‘with energy and easiness’. We have accordingly assigned it the part of speech locution nominale, emploi adverbia (‘nominal idiom, adverbial use’).

5.3.2 Government pattern

As mentioned above, we show the syntactic place of the Sem actants of the idiom, or the government pattern of each idiom, in the partial SSynt tree. But, since a great part of the studied idioms are typical syntactic dependents, we also describe their passive valence, highly relevant when a Sem actant is syntactically expressed as the governor of the lexical unit. Although not part of the government pattern, the idiom’s syntactic governor will be shown in the partial SSynt tree with a specification of its part of speech, using a modified square-shaped node, as in Figure 5.

![Figure 5: Partial SSynt tree of "TAMBOUR BATTANT"](image)

We should also present the behavior of actants of the idiom. They will sometimes be expressed “inside” the idiom. For example, "CASSER SA PIPE" (‘to die’ lit. to break his/her pipe) syntactically expresses its first Sem actant twice by both subject and possessive determiner. In "REMETTRE À SA PLACE" (‘to tell a person he/she is out of line’ lit. to put back to his/her place) the second Sem actant is syntactically expressed by both the object and the possessive determiner. This will be shown in the partial SSyntS as well, with an anaphoric relation (Figure 6).

![Figure 6: Partial SSynt trees of "CASSER SA PIPE" and "REMETTRE À SA PLACE"](image)
5.3.3 Blocking of Inflection of Idiom and Constituents

Additionally, we must describe all inflectional irregularities, which can be associated with the entire expression, or with single constituents. The “normal” inflectional properties of the constituents (except the head of the expression) are generally at least partially blocked, and one specific grammeme is needed. We will, for each constituent lexeme of the expression, specify the grammemes required in the idiom’s partial SSynt tree. Figure 4 shows blocked inflectional properties of the constituent lexemes of \textit{Quand on parle du loup on en voit la queue} indicated by the necessary grammemes.

5.3.4 Restricted Lexical Co-occurrence

Idioms may also have their own collocates and semantic derivatives. These are described by lexical functions in the field “fl” (lexical functions) of the lexical entry. Modifiers are sometimes dependents of a constituent that isn’t the head. For example, the entire idiom \textit{Monter un bateau} [à qqn.] (“to fool someone” lit. to assemble a boat) can be intensified by modifying the internal noun \textit{bateau}: \textit{Jean a monté un gros bateau à Julie}. This must be indicated in the condition of \textit{gros} as a value of the lexical function \texttt{Magn} for \textit{Monter un bateau}. We use the following encoding, reflecting passive valence of the modifier, as shown in example (2).

(2) \texttt{Magn} (monter un bateau) = \{bateau\} →\texttt{gros}

Conclusions

We have presented our observations drawn to date in the course of our development of a lexical database of French idioms. While idioms provide certain challenges for adequate treatment within a MTT framework as well as in the dictionary, the most challenging aspect of their lexical treatment is the description of their combinatorial properties and of their form constraints and variants.

The lexical description of an idiom must contain its partial SSynt tree. We have demonstrated that the syntactic representation could be used to describe a significant part of idiom’s unique behavior. We must now develop an exhaustive set of formal tools for the description of lexical particularities of French idioms.

Acknowledgements

I would like to express my gratitude to my readers, especially Igor Mel’čuk for his generous comments and merciful omission of \textit{Tête de mort} (skull and crossbones symbol) on my draft, even when I was completely wrong. Sylvain Kahane gave me many valuable advices for the final version. Thanks also to Josh Holden for his English-language editing of this paper and useful comments.
References


