Translating Idioms

Eric Wehrli*
Laboratoire d'analyse et de technologie du langage
University of Geneva
wehrli@latl.unige.ch

Abstract

This paper discusses the treatment of fixed word expressions developed for our ITS-2 French-English translation system. This treatment makes a clear distinction between compounds – i.e. multiword expressions of X^0-level in which the chunks are adjacent – and idiomatic phrases – i.e. multiword expressions of phrasal categories, where the chunks are not necessarily adjacent. In our system, compounds are handled during the lexical analysis, while idioms are treated in the syntax, where they are treated as "specialized lexemes". Once recognized, an idiom can be transfered according to the specifications of the bilingual dictionary. We will show several cases of transfer to corresponding idioms in the target language, or to simple lexemes. The complete system, including several hundreds of compounds and idioms can be consulted on the Internet (http://latl.unige.ch/itsweb.html).

1 Introduction

Multiword expressions (henceforth MWE), are known to constitute a serious problem for natural language processing (NLP)[1]. In the case of translation, a proper treatment of MWE is a fundamental requirement, as few customers would tolerate a literal translation of such common expressions as *entrer en vigueur 'to come into effect'*, *mettre en oeuvre 'to implement'*, *faire preuve 'to show'* or *faire connaissance 'to meet'*.

* I am grateful to Anne Vandeventer, Christopher Laenzlinger and Thierry Etchegoyhen for helpful comments. Part of the work described in this paper has been supported by a grant from CTI (grant no 2673.1).

[1]Cf. Abeillé & Schabes (1989), Arnold *et al.* (1995), Laporte (1988), Schenk (1995), Stock (1989), among others.

However, a simple glance at some of the current commercial translation systems shows that none of them can be said to handle MWEs in an appropriate fashion. As a matter of fact, some of them explicitely warn their users not to use multiword expressions.

In this paper, we will first stress some fundamental properties of two classes of MWEs, **compounds and idioms**, and then present the treatment of idioms developed for our French-English ITS-2 translation system (cf. Ramluckun & Wehrli, 1993).

2 Compounds and idioms

A two-way partition of MWEs in (i) compounds and (ii) idioms is both convenient and theoretically well-motivated[2]. Compounds are defined as MWEs of X^0-level (ie. word level), in which the chunks are adjacent, as exemplified in (1), while "idiomatic expressions" correspond to MWEs of phrasal level, where chunks may not be adjacent, and may undergo various syntactic operations, as exemplified in (2-3).

(1)a. pomme de terre 'potato'
b. à cause de 'because of'
c. dès lors que 'as soon as'

The compounds given in (1) function, respectively, as noun, preposition and conjunction. They correspond to a single unit, both syntactically and semantically. In contrast, idiomatic expressions do not generally constitute fixed, closed syntactic units. They do, however, behave as semantic units. For instance the complex syntactic expression *casser du sucre sur le dos de quelqu'un*, literally *break some sugar on*

[2]This distinction between compounds and idioms is also discussed in Wehrli (1997)

somebody's back is essentially synonymous with *criticize*.

(2)a. Jean a forcé la main à Luc.
Jean has forced the hand to Luc
'Jean twisted Luc's hand'

b. C'est à Luc que Jean a forcé la main.
It is to Luc that Jean has forced the hand
'It is Luc's hand that Jean has twisted'

c. C'est à Luc que Paul prétend que Jean a voulu forcer la main.
It is to Luc that Paul claims that Jean has wanted to force the hand
'It is Luc's hand that Paul claims that Jean has wanted to force'

d. La main semble lui avoir été un peu forcée.
The hand hand seems to him to have been a little forced
'His hand seems to have been somewhat twisted'

The idiom illustrated in (2) is typical of a very large class of idioms based on a verbal head. Syntactically, such idioms correspond to verb phrases, with a fixed direct object argument (*la main*, in our example) and an open indirect object argument. Notice that this verb phrase is completely regular in its syntactic behaviour. In particular, it can can undergo syntactic operations such as adverbial modification, raising, passive, dislocation, etc., as examplified in (2b-d).

With example (3), we have a much less common pattern, since the subject argument of the verb constitutes a chunk of the expression. Here, again, various operations are possible, including passive and raising [3]

(3)a. Quelle mouche a piqué Paul?
'What has gotten to Paul?'

b. Quelle mouche semble l'avoir piqué?
'What seems to have gotten to him'

c. Je me demande par quelle mouche Paul a été piqué.
'I wonder what's gotten to him'

[3] Another interesting example of idiom with fixed subject is *la moutarde monte au nez de NP ("NP looses his temper")*, discussed in Abeille and Schabes (1989).

The extent to which expressions can undergo modifications and other syntactic operations can vary tremendously from one expression to the next, and in the absence of a general explanation for this fact, each expression must be recorded with the list of its particular properties and constraints[4].

Given the categorial distinction (X^0 vs. XP) and other fundamental differences sketched above, compounds and idioms are treated very differently in our system. Compounds are simply listed in the lexicon as complex lexical units. As such, their identification belongs to the lexical analysis component. Once a compound has been recognized, its treatment in the ITS-2 system does not differ in any interesting way from the treatment of simple words.

While idiomatic expressions must also be listed in the lexicon, their entries are far more complex than the ones of simple or compound words (cf. section 3.2). As for their identification, it turns out to be a rather complex operation, which cannot be reliably carried out at a superficial level of representation. As we saw in the above examples, idiom chunks can be found far away from the (verbal) head with which they constitute an expression; they can also be modified in various ways, and so on. Preprocessing idioms, for instance during the lexical analysis, might therefore lead to lengthy, inefficient or unreliable treatments. We will argue that in order to drastically simplify the task of identifying idioms, it is necessary to undo whatever syntactic operations they might have undergone. To put it differently, idioms can best be recognized on the basis of a normalized structure, a structure in which constituents occur in their canonical position. In a generative grammar framework, normalized structures correspond to D-structure representations. At that level, for instance, the four sentences in (2), share the common structure in (4).

(4) ... [$_{VP}$ forcer [$_{DP}$ la main] [$_{PP}$ à X]]

As we will show in the next section, our treatment of idiomatic expression takes advantage of

[4] See for instance Nunberg *et al.* (1994), Ruwet (1983), Schenk (1995) or Segond and Breidt (1996) for a discussion on the degree of flexibility of idioms and (in the first two) interesting attempts to connect syntactic flexibility to semantic transparency

the drastic normalization process that our GB-based parser carries out.

3 A sketch of the translation process

In this section, we will show how idioms are handled in the French-to-English ITS-2 translation system, a transfer-based translation system which uses GB-style D-structure representations as interface structures. The general architecture of the system is given in figure 1 below.

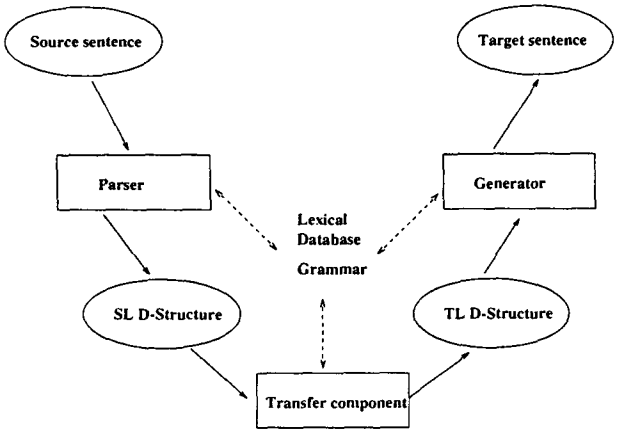

Figure 1. Architecture of ITS-2

For concreteness, we shall first focus on the epinonymous idiom given in (5):

(5)a. Paul a cassé sa pipe.
lit. 'Paul has broken his pipe'

b. Paul kicked the bucket.

Translation of (5a) is a three-step process:

- Identification of source idiom
- Transfer of idiom
- Generation of target idiom

3.1 Idiom identification

As we argued in the previous section, the task of identifying an idiom is best accomplished at the abstract level of representation (D-structure). ITS-2 uses the IPS parser (*cf.* Wehrli, 1992, 1997), which produces the structure (6) for the input (5a) [5]:

(6) [$_{TP}$ [$_{DP}$ Paul] [$_{\overline{T}}$ a [$_{VP}$ cassé [$_{DP}$ sa [$_{NP}$ pipe [$_{PP}$ e]]]]]]

At this point, the structure is completely general, and does not contain any specification of idioms. The idiom recognition procedure is triggered by the "head of idiom" lexical feature associated with the head *casser*. This feature is associated with all lexical items which are heads of idioms in the lexical database.

The task of the recognition procedure is (i) to retrieve the proper idiom, if any (*casser* might be the head of several idioms), and (ii) to verify that all the constraints associated with that idiom are satisfied. Idioms are listed in the lexical database as roughly illustrated in (6)[6]:

(7)a. casser sa pipe
'to kick the bucket'

b. 1: [$_{DP}$] 2: [$_{V}$ casser] 3: [$_{DP}$ POSS pipe]

c. 1. [+human]
2. [−passive]
3. [+literal, −extraposition]

Idiom entries specify (a) the canonical form of the idiom (mostly for reference purposes), (b) the syntactic frame with an ordered list of constituents, and (c) the list of constraints associated with each of the constituents.

In our (rather simple) example, the lexical constraints associated with the idiom (7) state that the head is a transitive lexeme whose direct object has the fixed form "POSS *pipe*", where POSS stands for a possessive determiner coreferential with the external argument of the head (i.e. the subject). Furthermore, the subject constituant bears the feature [+human], the head is marked as [-passive], meaning that this particular idiom cannot be passivized. Finally, the object is also marked [+literal, -extraposition], which means that the direct object constituent cannot be modified in any way (not even pluralized), and cannot be extraposed.

The structure in (7) satisfies all those constraints, provided that the possessive *sa* refers

[5] In example 6, we use the following syntactic labels : **TP** (Tense Phrase) for sentences, **VP** for verb phrases, **DP** for Determiner Phrases, **NP** for Noun Phrases, and **PP** for Prepositional Phrases.

[6] See Walther & Wehrli (1996) for a discussion of the structure of the lexical database underlying the ITS-2 project

uniquely to *Paul*[7]. It should be noticed that even though an idiom has been recognized in sentence (6), it also has a semantically well-formed literal meaning. Running ITS-2 in interactive mode, the user would be asked whether the sentence should be taken literaly or as an expression. In automatic mode, the idiom reading takes precedence over the literal interpretation[8].

3.2 Transfer and generation of idioms

Once properly identified, an idiom will be transfered as any other abstract lexical unit. In other words, an entry in our bilingual lexicon has exactly the same form no matter whether the correspondance concerns simple lexemes or idioms. The corresponding target language lexeme might be a simple or a complex abstract lexical unit. For instance, our bilingual lexical database contains, among many others, the following correspondances:

French	English
avoir besoin de X	need X
casser sa pipe	kick the bucket
faire la connaissance de X	meet X
avoir envie	feel like
quelle mouche a piqué	what has gotten

The generation of target language idioms follows essentially the same pattern as the generation of simple lexemes. The general pattern of generation in ITS-2 is the following: first, a maximal projection structure (XP) is projected on the basis of a lexical head and of the lexical specification associated with it. Second, syntactic operations apply on the resulting structure (extraposition, passive, etc.) triggered either by lexical properties or general features transfered from the source sentence. For instance, the lexical feature [+raising] associated with a predicate would trigger a raising transformation (NP movement from the embedded subject position to the relevant subject position). Subject-Auxiliary inversion, topicalization, auxiliary verb insertion are all examples of syntactic transformations triggered by general features, derived from the source sentence.

[7]Given a proper context, the sentence could be construed with *sa* referring to some other person, say Bill.

[8]Such a heuristic seems to correspond to normal usage, which would avoid formulation (5a) to state that 'Paul has broken someone's pipe'.

The first step of the generation process produces a target language D-structure, while the second step derives S-structure representations. Finally, a morphological component will determine the precise orthographical/phonological form of each lexical head.

In the case of target language idioms, the general pattern applies with few modifications. Step 1 (projection of D-structure) is based on the lexical representation of the idiom (which specifies the complete syntactic pattern of the idiom, as we have pointed out earlier), and produces structure (8a). Step 2, which only concerns the insertion of perfective auxiliary in position T^0, derives the S-structure (8b). Finally, the morphological component derives sentence (8c).

(8)a. [$_{TP}$ [$_{DP}$ Paul] [$_{VP}$ kick [$_{DP}$ the [$_{NP}$ bucket]]]]

b. [$_{TP}$ [$_{DP}$ Paul] [$_{\overline{T}}$ have [$_{VP}$ kick [$_{DP}$ the [$_{NP}$ bucket]]]]]

c. Paul has kicked the bucket.

4 Conclusion

In this paper, we have argued for a distinct treatment of compounds, viewed as complex lexical units of X^0-level category, and of idioms, which are phrasal constructs. While compounds can be easily processed during the lexical analysis, idiomatic expressions are best handled at a more abstract level of representation, in our case, the D-structure level produced by the parser. The task of recognition must be based on a detailed formal description of each idiom, a lengthy, sometimes tedious but unavoidable task. We have then shown that, once properly identified, idioms can be transfered like any other abstract lexical unit. Finally, given the fully-specified lexical description of idioms, generation of idiomatic expressions can be achieved without ad hoc machinery.

5 References

Abeillé, A. and Schabes, Y. (1989). "Parsing Idioms in lexicalized TAGs", *Proceedings of EACL-89*, Manchester, 1-9.

Arnold, D., Balkan, L., Lee Humphrey, R., Meijer, S., Sadler, L. (1995). *Machine Translation: An Introductory Guide*, HTML document (http://clwww.essex.ac.uk).

Laporte, E. (1988). "Reconnaissance des expressions figées lors de l'analyse automatique", *Langages* 90, Larousse, Paris.

Nunberg, G., Sag, I., Wasow, T. (1994). "Idioms", *Language*, 70:3, 491-538.

Ramluckun, M. and Wehrli, E. (1993). "ITS-2 : an interactive personal translation system" *Actes du colloque de l'EACL*, 476-477.

Ruwet, N. (1983). "Du bon Usage des Expressions Idiomatiques dans l'argumentation en syntaxe générative". In *Revue québécoise de linguistique*. 13:1.

Schenk, A. (1995). 'The Syntactic Behavior of Idioms'. In Everaert M., van der Linden E., Schenk, A., Schreuder, R. *Idioms: Structural and Psychological Perspectives*, Lawrence Erlbaum Associates, Hove.

Segond, D., and E. Breidt (1996). "IDAREX : description formelle des expressions à mots multiples en français et en allemand" in A. Clas, Ph. Thoiron and H. Béjoint (eds.) *Lexicomatique et dictionnairiques*, Montreal, Aupelf-Uref.

Stock, O. (1989). "Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind", *Computational Linguistics*, 15.1. 1-18.

Wehrli, E. (1992)"The IPS system", in C. Boitet (ed.) *COLING-92*, 870-874.

Wehrli, E. (1997) *L'analyse syntaxique des langues naturelles : problèmes et méthodes*, Paris, Masson.

Walther, C., and E. Wehrli (1996) "Une base de donnees lexicale multilingue interactive" in A. Clas, P. Thoiron et H. Béjoint (eds.) *Lexicomatique et dictionnairiques*, Montreal, Aupelf-Uref, 327-336.