Detecting Hidden Multiwords in Bilingual Dictionaries

Luisa Bentivogli and Emanuele Pianta
ITC-irst

via Sommarive 18

38050 Povo (Trento)

Italy

{bentivo, pianta}@itc.it

Abstract

Dictionaries are a valuable source of information about multiwords. Unfortunately, only few multiwords are explicitly marked as such in dictionaries: most of them are presented without being distinguished from free combinations of words. In this paper we present a methodology for detecting hidden multiwords in bilingual dictionaries, along with their translation in another language. The methodology is based on a number of automatic procedures which exploit regularities in the different kinds of expressions that can be found in the Collins English-Italian bilingual dictionary to select those phrases that are most likely to contain multiwords. The quantitative results of the experiment are provided.

1 Introduction

Multiword units are an important research field for lexicography and language teaching; they pose interesting theoretical questions on the human language faculty, but they also have a practical impact in many applications in the field of computational linguistics. Information retrieval and extraction, machine translation, text summarization, text classification, and question answering are all fields in which information about multiwords turns out to be useful. For this reason ITC-irst has started a project for extending the multilingual lexical database MultiWordNet [Pianta et al. 2002] with a large number of Italian multiwords along with their English translations.

Techniques for the recognition of multiwords include linguistic and statistical techniques for extracting multiword units from corpora [Jacquemin & Bourigault 2000], [Daille 1996]. Such techniques are often used as a source of new multiwords to be included in dictionaries. However, dictionaries themselves are a valuable source of information about multiwords [Fontenelle 1997]. Bilingual dictionaries tend to be particularly rich of multiwords in comparison with monolingual dictionaries as, for their very nature, multiwords cannot be translated on a word by word basis. For example, the English multiword *one way* cannot be translated literally in Italian as *una via*; the corresponding Italian multiword is instead *senso unico* (lit. unique direction). Bilingual dictionaries are particularly relevant to our purposes as they also provide translations of multiwords.

Unfortunately, in both monolingual and bilingual dictionaries, a consistent number of multiwords are presented without being explicitly distinguished from free combination of words. Only in few cases multiwords are presented either as headwords or as sub-headwords, or are explicitly marked *(explicit multiwords)*. This state of affairs led us to study the microstructure of the Collins English-Italian bilingual dictionary (available in electronic

format), to verify whether we could extract in a semi-automatic way both explicitly marked and *hidden multiwords*, i.e. multiwords not explicitly marked as such.

This paper is organized as follows. In Section 2 we analyze the microstructure of the Collins dictionary, paying special attention to the way multiwords are handled, while in Section 3 we present a more fine-grained classification of the expressions that can be found in the phrase-zones (containing both phraseology and examples) of a Collins dictionary entry. Then, in Section 4 we illustrate the semi-automatic procedures we have devised to extract hidden multiwords from the phrase-zone. Finally, in Section 5 the quantitative results of our research are presented before concluding in Section 6.

2 Multiwords in the Collins Dictionary

2.1 Definitions about Multiwords

Before analyzing the aspects of the microstructure of the Collins dictionary related to multiwords, let us introduce some definitions and conventions. Following part of the literature on lexical collocations [Benson et al. 1986], [Cruse 1986], [Heid 1994], we adopt a three-fold classification distinguishing idioms, (restricted) collocations, and free combinations of words.

With the term *multiword* we refer to both *idioms* and *restricted collocations*:

- an *idiom* is a relatively frozen expression whose meaning cannot be built compositionally from the meanings of its component words. Also, the component words cannot be substituted with synonyms.
- a *restricted collocation* is a sequence of words which habitually co-occur and whose meaning can be derived compositionally. Restricted collocations have a kind of semantic cohesion mainly due to use and therefore they considerably limit the substitution of their component words. Usually restricted collocations do not have a literal translation in other languages. Following [Aisenstadt 1979] and [Marello 1989], we prefer to use the term "restricted collocation" instead of "collocation" to avoid confusion with some uses of "collocation" in the wider sense of any co-occurrence of words.

Multiwords must be distinguished from *free combinations of words*:

- a *free combination of words* is a combination of words following only the general rules of syntax: the elements are not bound specifically to each other and so they occur with other lexical items freely.

2.2 Explicit and Hidden Multiwords in the Collins Dictionary Entries

The Collins machine-readable bilingual dictionary is a medium size dictionary including 37,727 headwords in the English Section and 32,602 headwords in the Italian Section. In the Collins dictionary a certain amount of multiwords are presented in canonical form and explicitly marked.

In the English-to-Italian Section 9,054 multiwords are included as headwords. See for instance:

(1) **roller coaster** *n* montagne russe *(fpl)*

In the Italian-to-English Section only very few multiwords are included as headwords (93, mostly frozen expressions):

(2) **usa e getta** *agg inv (rasoio, siringa)* disposable, throwaway

However, 2,874 multiwords are included as sub-headwords (introduced by a bullet):

(3) **soldato** *sm* soldier; ... • **soldato di leva** conscript • **soldato semplice** private

Explicit multiwords are particularly interesting for us, as they are immediately available for inclusion in MultiWordNet. However the entries of the Collins dictionary also contain many hidden multiwords. Consider a typical Collins entry:

	TE-zone	*phrase-zone*
(4) **shelf** *n* **a.**	*(in cupboard, oven)* ripiano; *(fixed to wall)* mensola;	*to be on the shelf, (fig:*

fam: woman) essere zitella **b.** *(in rock face, underwater)* piattaforma

A Collins entry contains sense sub-divisions introduced by a letter (**a** and **b** in the example). Each sub-division can be made up of two translation zones: in the first zone, translation equivalents (TEs) for the headword are provided (let's call this the *TE-zone*); in the second zone, phrases containing the headword with their translations are provided (*phrase-zone*). Both the TE-zone and the phrase-zone are sources of hidden multiwords.

Let us consider the phrase-zone first. The phrase in Example 4 above (*to be on the shelf*) is a multiword in canonical form. Unfortunately, a phrase-zone does not contain only multiwords in canonical form. See:

(5) **source** *n (of river)* sorgente *(f) (fig: of problem, epidemic)* fonte *(f)* origine *(f)* oranges
are a source of vitamin C, le arance sono ricche di vitamina C; *I have it from a reliable source that ...*, ho saputo da fonte sicura che...

Here, the two phrases are examples of usage in inflected form. As a matter of fact, such phrases do contain multiwords (*vitamin C, to have it from, reliable source*), but they are neither isolated nor in canonical form. There is no extrinsic way to distinguish a phrase-zone containing multiwords in canonical form from a phrase-zone containing examples of usage. Moreover, the two kinds of phrases can occur in the same phrase-zone.

Also the TE-zone may contain multiwords in canonical form:

(6) **peanut** n arachide *(f)*, nocciolina americana; *to work for peanuts, (fam)* lavorare per una miseria.

In this entry there are no explicit multiwords. However, one of the two TEs of the headword *(nocciolina americana)* is a restricted collocation in Italian.

Summing up, hidden multiwords come from two sources: (a) translation equivalents from the TE-zone; (b) phrases from the phrase-zone.

In Bentivogli and Pianta [2000], in the context of a study on lexical gaps between English and Italian, we presented a methodology, based partly on automatic procedures and partly on manual check, to select multiwords from TEs made up of more than one word. By applying the methodology we extracted 9,792 multiwords out of 14,530 complex Italian TEs.

In the following sections we will illustrate techniques to extract hidden multiwords from the phrase-zone of a dictionary entry.

3 The Phrase-zone of the Collins Dictionary

In this section we present a more fine-grained classification of the expressions that can be found in the phrase-zones of a Collins entry, with the final aim of finding a procedure that selects those phrases which are most likely to contain multiwords, preferably in canonical form (examples of translation pairs are presented as: <*source item* = target item>). Remember that these groups of different expressions are not explicitly distinguished in the microstructure of the dictionary.

1) *Examples of usage*
   For instance: <*the book is divided into 5 chapters* = il libro si divide in 5 capitoli>. Examples of usage are specific phrases, mostly finite verb sentences or noun phrases with a determiner, exemplifying how the headword *(divide*, in the example) is used in a free combination of words. Very often examples are chosen so as to implicitly provide additional linguistic information about the usage of the headword. For instance from the above example we may learn about the subcategorization pattern of *to divide (into* + NP). Examples of usage may contain multiwords in inflected form. For example, under the Italian headword *accusare* one finds: <*ha accusato il colpo* = you could see that he had felt the blow>. The Italian phrase contains a multiword *(accusare il colpo)*, but in inflected form.

2) *Proverbs and stereotypical sentences*
   Proverbs are explicitly marked through a gloss: <*chi si accontenta gode (Proverbio)* = well pleased is well served>. As regards stereotypical sentences, some fixed expressions are explicitly signaled by double quotes (<*"keep off the grass"* = "vietato calpestare l'erba">); some others are not (<*the exception proves the rule* = l'eccezione conferma la regola>).

3) *Grammatical collocations*
   These kinds of phrases (occurring only in dictionaries) are composed of a lexical unit *(head)* and a pattern specifying the subcategorization properties of the head. The literature on collocations [Benson et al. 1986], often assumes that the head of a grammatical collocation is a simple word, as in <*to convey to sb that* = comunicare a

qualcuno che>. However, in the grammatical collocations of the Collins dictionary the head of a grammatical collocation is very often a complex unit, and in most cases this complex unit is a multiword. For example: *<to enter into a contract with sb to do sth/for sth* = stipulare un contratto con qn per fare qc/per qc>. This phrase contains a multiword in canonical form: *to enter into a contract* (and *stipulare un contratto* in Italian), while the rest of the phrase specifies the subcategorization of the verb.

4) *Free combinations of words in canonical form*
There are two kinds of such phrases. The most common are examples of usage in canonical form, for instance *<abitare al terzo piano* = to live on the third floor>. Other free combinations of words are instead explanations of lexical gaps. For example, under the Italian headword *abusivamente* one finds: *<occupare abusivamente una casa* = to squat>. The Italian expression is used to paraphrase a lexical concept which does not exist in Italian.   ·

5) *Multiwords in canonical form*
For instance, under the English headword *drop* we find: *<to drop anchor* = gettare l'ancora>

Given the aims of our project, we are especially interested in multiwords in canonical form (class 5) as they are ready for insertion in MultiWordNet. Also grammatical collocations (class 3) and examples of usage (class 1) are relevant to our work as they may contain multiwords, although immersed in larger expressions or not in canonical form. Additional work is needed to isolate them and to put them in canonical form. On the contrary, we are not interested in proverbs and stereotypical sentences (class 2) as MultiWordNet does not contain such kind of expressions. Neither are we interested in free combinations of words (class 4) because they do not belong to a language lexicon.

4 Semi-automatic Extraction of Hidden Multiwords from the Phrase-zone

By analyzing the expressions in the phrase-zone, we found that some classes exhibit structural regularities that can be exploited to automatically enumerate their members. Unfortunately, it is not possible to devise an automatic procedure that directly enumerates all and only the *translation pairs* (TPs) belonging to class 5. On the other hand, it is relatively easy to implement procedures that enumerate the TPs of classes 1, 2 and 3. Moreover, some members of class 5 can be enumerated with acceptable precision but low recall. Thus, we try to select multiwords in canonical form in an indirect way.

First, we single out as many TPs as possible not belonging to class 5. A first procedure selects TPs in inflected form (classes 1 and 2), and a second procedure selects grammatical collocations (class 3). Note that class 3, given some additional processing for stripping away the subcategorization patterns, becomes indeed a valuable source of multiwords. Then we apply a third procedure that selects multiwords with high precision but low recall. What remains after the application of the three procedures is another group of TPs which contains a high number of multiwords in canonical form which have to be manually distinguished from free combinations of words.

The procedures for detecting hidden Italian multiwords have been applied to both sections of the Collins dictionary. In the Italian-to-English Section, Italian phrases are found as source phrases, while in the English-to-Italian Section they are found as target phrases. However, the procedures operate not only on Italian phrases but on TPs, as we are interested in Italian multiwords along with their English translations.

4.1 Selecting TPs in Inflected Form

The first procedure selects TPs including phrases that belong to class 1 (examples of usage) or class 2 (proverbs and stereotypical sentences). In order to automatically select TPs in inflected form we devised a procedure that is simple but guarantees a very high precision:

> if the first word of the Italian phrase is not a preposition, noun, verb or adjective in canonical form or if you can find in either the Italian or English phrase any first or second person pronoun or possessive, then the TP is in inflected form.

To verify whether a word is in canonical form, the procedure checks if the word is a headword in the relevant section of the dictionary. The procedure takes into account the fact that in English the canonical form of verbs follows the *to + infinite* verb pattern. Here is an example of TP selected by the procedure: <*abita al numero 10* = she lives at number 10>.

By manually checking the 5% of the Italian phrases selected by this procedure, we found that only 16.7% of them contain multiwords (in inflected form).

4.2 Selecting Grammatical Collocations

All the TPs that are not selected from the previous procedure are very likely to be in canonical form. Among these TPs, the second procedure selects grammatical collocations (class 3), i.e. phrases containing subcategorization patterns. In the Collins dictionary, subcategorization information is specified by different patterns:

- for English: "(PREP) (sb|sth|o.s.)" and "(COMP) (do|doing) (sth)"
- for Italian: "(PREP) (qc|qn|qualcosa|qualcuno)" and "(COMP) (fare qc)"

We used these patterns to run the following procedure:

> If the source or target phrase of the TP contains a subcategorization pattern, then classify the TP as a grammatical collocation. If what remains after stripping away the subcategorization pattern from the Italian phrase, i.e. the head of the grammatical collocation, is formed by more than one word, then take the head as a possible multiword in canonical form.

For example, the procedure selects the following TP: <*to take sb to court over sth* = citare in tribunale qn per qc>. Stripping away the subcategorization pattern we get a multiword both in English and in Italian: *to take to court* and *citare in tribunale*.

After an exhaustive manual check of this group, it turned out that in the 60.2% of the cases the complex head of the grammatical collocation is a multiword.

4.3 Selecting TPs that are Likely to Include Multiwords in Canonical Form

At this point we are left with TPs which are in canonical form and are not grammatical collocations. The third procedure selects among them two groups that are very likely to contain multiwords (belonging to class 5, according to the classification in Section 3).

Italian Support Verbs. These expressions represent a typical structural pattern in which restricted collocations can be found [Heid 1994]: *<to brief* = dare istruzioni a> (lit. to give instructions to). A list of support verb constructions has been defined on the basis of the relevant literature [Renzi 1988]. Five Italian verbs have been selected as support verbs: *fare, prendere, mettere, avere, dare*. Then the procedure looks for a matching between an Italian phrase and one of the following two patterns:

- SupportVerb (Prep) (Art) Noun;
- SupportVerb Adjective.

The manual check of this group showed that the 94.4% of the phrases selected by the procedure are indeed multiwords.

Back Translations. If a multiword is present in both sections of the dictionary, then it is highly probable that the multiword is found exactly in the same form as source phrase in one section and as target phrase in the other. This is due to the fact that the items of a multiword have low variability.

The procedure exploits this fact to select a group of TPs that are very likely to contain Italian multiwords. More precisely we select from the Italian Section of the dictionary all TPs *<Ita-source* = Eng-target> such that we can find in the English Section a pair *<Eng-source* = Ita-target> with *Ita-source* coinciding with Ita-target.

For instance, the Italian Section contains the following TP: *<mettersi in ghingheri* = to dress up to the nines>. In the English Section we find: *<to get (all) dolled up* = mettersi in ghingheri>. In some cases also Eng-target and *Eng-source* coincide. For instance, the Italian Section contains *<tagliarsi le vene* = to slash one's wrists> and in the English section we find: *<to slash one's wrists* = tagliarsi le vene>.

We verified by manual check that, if this condition holds, 85% of the TPs contain an Italian multiword.

4.4 Remaining Phrases

After the sequential application of the three procedures described above we are left with a group of TPs in which multiwords must be manually distinguished from free combinations of words. As all phrases in inflected form, which contain a very low number of multiwords, has been previously eliminated, this group is likely to contain a high number of multiwords in canonical form. After a manual check of a sample of this group, we found that the 44.8% are indeed multiwords.

	Italian source phrases from the Italian Section	Italian target phrases from the English Section	Distinct Italian phrases from the union of Sections	% of Italian multiwords (after manual check)	
				Canonical form	Inflected form
In Inflected form	11,283	13,251	23,875	-	16.7*
Gramm. colloc. (complex head)	2,996 (2,285)	3,617 (2,647)	6,248 (4,521)	60.2	-
Support Verbs	534	288	729	94.4	-
Back translations	1,222	-	1,222	85.0	-
Remaining phrases	8,639	7,818	16,457	44.8*	15.3*
Total	24,674	24,974	48,531	23.3*	8.5*

Table 1: Results of the application of the automatic procedures to the Collins dictionary

5 Quantitative Results

Table 1 shows the results obtained from the application of the described procedures. In the first and second column the data refer to Italian phrases in both Italian and English Sections of the dictionary. The data of the third column refer to the number of distinct Italian phrases obtained by the union of the two Sections. The union has been made in order to avoid the possible overlapping between the Italian source phrases and the Italian target phrases in the two sections of the dictionary. The last two columns report the percentage of multiwords actually contained in each group, obtained by manual check. Percentages marked by an asterisk refer to a manual check of a significant sample.

Summing up we can estimate that at least 15,796 Italian expressions out of 48,531 from the phrase-zone of the Collins dictionary contain hidden multiwords. Among them, 11,569 (73.2%) are already in canonical form. The automatic procedures devised allowed the lexicographers to focus on the most productive sources of Italian multiwords with English translation to be added to the MultiWordNet database. The groups containing support verbs and back translations are the first to be taken into account as they include the highest number of multiwords in canonical form (respectively 94.4% and 85.0%), which are ready for insertion in MultiWordNet. Then the groups of grammatical collocations (after having stripped away the subcategorization patterns) and of "remaining phrases", which contains a lower number of multiwords in canonical form (60.2% and 44%). Finally, the group of phrases in inflected form can be taken into account as last because it contains only few multiwords (16.7%) which, moreover, need to be isolated and put in canonical form before the insertion in MultiWordNet.

792

6 Conclusions

In this paper we analyzed a number of techniques to detect in a semi-automatic way multiwords from a bilingual dictionary. The experiment run on both sections of the Collins English-Italian dictionary confirms that bilingual dictionaries are indeed valuable sources of information about multiwords. Working on the identification of Italian multiwords we found that the Collins dictionary contains about 3,000 explicit multiwords and almost 15,800 hidden multiwords.

References

[Aisenstadt 1979] Aisenstadt, E., 1979. Collocability restrictions in dictionaries, in: R.R.K. Hartmann (ed.) *Dictionaries and their users*. University of Exeter, Exeter.

[Bentivogli and Pianta 2000] Bentivogli, L. & E. Pianta, 2000. Looking For Lexical Gaps, in: U. Heid et al. (eds.) *Proceedings of the Ninth Euralex International Congress, Euralex 2000*, Stuttgart, Germany.

[Benson et al. 1986] Benson, M., E. Benson & R. Ilson, 1986. *The BBI combinatory dictionary of English: a guide to word combinations*. John Benjamins Publishing Company, Philadelphia.

[Cowie 1981] Cowie, A.P., 1981. The treatment of collocations and idioms in learner's dictionaries, in: *Applied Linguistics*, 2(3), Oxford University Press, Oxford.

[Cruse 1986] Cruse, D.A., 1986. *Lexical semantics*. Cambridge University Press, Cambridge.

[Daille 1996] Daille, B., 1996. Study and implementation of combined techniques for automatic extraction of terminology, in: Klavans J. L., & P. Resnik (eds) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, Cambridge (Mass).

[Fontenelle 1997] Fontenelle, T., 1997. *Turning a bilingual dictionary into a lexical semantic database*. Max Niemeyer Verlag, Tubingen.

[Heid 1994] Heid, U., 1994. On ways words work together: research topics in lexical combinatorics, in: W. Martin et al. (eds) *Proceedings of the Sixth Euralex International Congress, Euralex-94*, Amsterdam, Holland.

[Jacquemin & Bourigault 2000] Jacquemin, C. & D. Bourigault, 2000. Term extraction and automatic indexing, in: R. Mitkov (ed) *Handbook of computational linguistics*. Oxford University Press, Oxford.

[Marello 1989] Marello, C., 1989. *Dizionari bilingui*. Zanichelli, Bologna.

[Pianta et al. 2002] Pianta, E., L. Bentivogli & C. Girardi, 2002. MultiWordNet: developing an aligned multilingual database, in: Udaya Narayana Singh (ed) *Proceedings of the First Global WordNet Conference*, Central Institute of Indian Languages, Mysore, India.

[Renzi 1988] Renzi, L., 1988. *Grande grammatica italiana di consultazione*, Vol. 1. Il Mulino, Bologna.