

# Coping with lexical gaps when building aligned multilingual wordnets

Luisa Bentivogli, Emanuele Pianta, Fabio Pianesi

ITC-irst  
via Sommarive 18, I-38050 Povo - Trento, Italy  
{bentivo,pianta,pianesi}@irst.itc.it

## Abstract

In this paper we present a methodology for automatically classifying the translation equivalents of a machine readable bilingual dictionary in three main groups: lexical units, lexical gaps (that is cases when a lexical concept of a language does not have a correspondent in the other language) and translation equivalents that need to be manually classified as lexical units or lexical gaps. This preventive classification reduces the manual work necessary to cope with lexical gaps in the construction of aligned multilingual wordnets.

## 1. Introduction

In this paper we present a general methodology for handling lexical gaps when building aligned multilingual wordnets. The procedure has been tested during the construction of Italian WordNet (IWN), an on-going project at ITC-irst aiming at producing a wordnet for the Italian language strictly aligned to the Princeton WordNet (PWN), see (Fellbaum, 1998). In its first version, IWN will contain synsets for 40,000 Italian words and information about the correspondence between Italian and English (PWN) synsets (Ciravegna et al., 1994). IWN is being developed within MultiWordNet, a methodological framework distinct from EuroWordNet.

### 1.1. MultiWordNet vs EuroWordNet

There are at least two models for building a multilingual wordnet. The first model, adopted within the EuroWordNet project (EWN), consists in building language specific wordnets independently from each other, trying in a second phase to find correspondences between them (Vossen, 1998). The second model, adopted within the MultiWordNet project (MWN), consists in building language specific wordnets keeping as much as possible of the semantic relations available in the Princeton WordNet (PWN). This is done by building the new synsets in correspondence with the PWN synsets, whenever possible, and importing semantic relations from the corresponding English synsets; i.e. we assume that, if there are two synsets in PWN and a relation holding between them, the same relation holds between the corresponding synsets in the new language.

According to (Vossen, 1996), the MWN model (or “expand model” in his words) seems less complex and guarantees the highest degree of compatibility across different wordnets. To see this, consider that building any wordnet necessarily implies a large number of subjective (and questionable) decisions. Thus, if two wordnets are built independently for two different languages, they will exhibit differences which depend only partially on differences between the languages. Some non trivial structural discrepancies will depend in fact on subjective decisions or different building criteria. The MWN model minimizes these discrepancies by strictly adhering to the PWN building criteria and subjective choices.

The MWN model has also potential drawbacks. The most serious risk is that of forcing “an excessive dependency on the lexical and conceptual structure of one of the languages involved”, as (Vossen, 1996) points out. This risk can be avoided by allowing the new wordnet to diverge, when necessary, from the PWN.

### 1.2. Building Italian WordNet

Another important advantage of the MWN model is that automatic procedures can be devised to speed up both the construction of corresponding synsets and the detection of divergences between PWN and the building wordnet. In all these procedures PWN itself can be used as a useful resource.

The construction of IWN, which is the first instantiation of the MWN model so far, is crucially based on two automatic procedures. The first is called Assign-procedure. Given an Italian word sense the procedure selects a weighed list of the most likely corresponding PWN synsets. Such a list is then used by the lexicographers to actually build the Italian synsets. The second procedure, which is the main concern of this paper, supports the detection of “lexical gaps” (LG procedure), that is cases when a lexical concept of a language does not have a correspondent in the other language.

Information about lexical gaps can be used in two ways, depending on whether we are dealing with Italian-to-English gaps or vice versa. The Italian-to-English gaps point to a set of Italian synsets that need to be added manually in IWN: we know for sure and from the beginning that such synsets cannot be build in correspondence to any English synset and thus their construction cannot be based on the results of the Assign-procedure. On the other hand, information about English-to-Italian gaps point to idiosyncratic PWN synsets that can be excluded a priori from those selected by the Assign-procedure. The rest of this paper describes the LG procedure in detail. In section 2. we define the notions of lexical unit and lexical gap. Then, in section 3. we describe the steps of a procedure that automatically classifies the translation equivalents of a machine readable bilingual dictionary in three sets: lexical units, lexical gaps and translation equivalents that need to be manually classified as lexical units or lexical gaps. Section 4. discusses the results of the procedure.

## 2. What is a lexical gap

The literature on contrastive analysis shows that, given a *source* and a *target language*, various types of idiosyncrasies can occur at lexical level. One of the most common idiosyncrasies, especially relevant given the PWN building criteria, are lexical gaps, see (Bentivogli and Pianta, 2000).

A *lexical gap* occurs whenever a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a *free combination of words* (Hutchins and Somers, 1992). Following the PWN building criteria, a *lexical unit* may be either a *simple word*, an *idiom* or a *restricted collocation* (Cowie, 1981):

- an *idiom* is a 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. Therefore they considerably limit the substitution of their component words. Usually restricted collocations do not have a literal translation in other languages. For example the Italian *senso unico* (lit. unique direction) corresponds to the English *one way*.
- a *free combination* 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.

Figure 1 summarizes the relations among the relevant lexicographic notions.

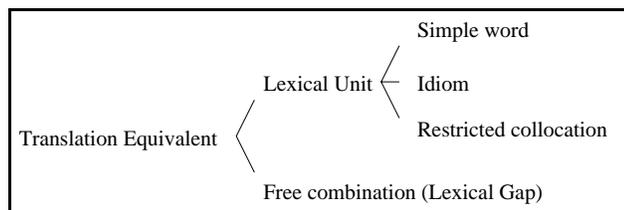


Figure 1: A classification of Translation Equivalents

In practice, the boundaries between idioms, restricted collocations and free combinations are not clear-cut. However, in many cases a distinction can be drawn relying on knowledge contained in dictionaries that explicitly mark idioms and collocations. Also, the three groups exhibit certain structural regularities that can be exploited to automatically distinguish them from each other with a certain degree of confidence. In the following, we will refer to both idioms and restricted collocations with the generic term of *collocations*.

## 3. Finding lexical gaps

To make the detection of lexical gaps easier, we have devised a procedure relying mainly on the Collins machine

readable bilingual dictionary. This is a medium size dictionary including 37,727 headwords and 55,777 translation groups in the English section and 32,602 headwords and 46,545 translation groups in the Italian section. By translation group (TGR) we mean a group of Italian synonyms translating a sense of an English word. In bilingual dictionaries, translation groups are usually separated by semicolons. In Example 1 the English word *wood*, as a noun, has 5 translation groups<sup>1</sup>

- (1) **wood** 1. *n a.* (material) legno; (timber) legname (m) **b.** (forest) bosco **c.** (Golf) mazza di legno; (Bowls) boccia 2. *adj a.* (made of wood) di legno **b.** (living etc. in a wood) di bosco, silvestre.

We take translation groups as the lexicographic counterpart of the notion of word sense as they have the same sense granularity as the sense distinction of the PWN synsets. More precisely we expect that each translation group corresponds to a PWN synset. Besides the bilingual Collins, a number of electronic monolingual resources have been used as well; among them, the DISC Italian dictionary and PWN itself.

The purpose of the LG procedure is to reduce as much as possible the word senses that need to be manually identified as lexical gaps. More specifically the procedure classifies all TGRs of a section of a bilingual dictionary in three main groups: (1) lexical units, (2) lexical gaps, (3) TGRs that need to be manually classified. To this extent we singled out a number of TGR support classes. These classes can be established either by resorting to available linguistic resources, or in a formal way (that is without relying on human semantic intuitions). The TGR support classes are defined on the basis of the kind of translation equivalents that compose them, that is:

1. at least one *simple word*. Obviously, for all TGRs of this class we can exclude the existence of a lexical gap: *cielo = sky*;
2. at least one phrase listed as a main entry or as a collocation in the available resources. Given the definition of lexical unit presented in section 2., we can safely assume that also TGRs in this class exclude the existence of lexical gaps: to acknowledge = *prendere atto di* (listed as collocation in the Italian dictionary)
3. at least a so-called *support verb construction*. These expressions represent a typical structural pattern in which collocations can be found (Heid, 1994): to brief = *dare istruzioni a* (lit. to give instructions to)
4. at least one multi-word phrase with a number of non-functional words corresponding to that of the source language phrase. This class is relevant for our purposes because we noticed that, when both the source and target languages express a meaning with a multi-word phrase containing the same number of content words, if the source language expression is a restricted

<sup>1</sup>In the Collins entries, translation groups having semantic aspects in common are grouped in sections introduced by letters, whereas different parts of speech are introduced by numbers.

collocation, very often so is the translation equivalent: roller coaster = *montagne russe*; agony column = *posta (del) cuore*;

5. at least a two-word phrase labeled with a gloss specifying a technical semantic field. In specialized technical terminologies, most multi-word expressions can be classified as idioms or restricted collocations: armour (MIL) = *mezzi blindati*.
6. only one phrase following a syntactic pattern that can be systematically put in correspondence with a derivational morphological pattern in the source language. This happens when the source language expresses a meaning through morphological means whereas the target language expresses the same meaning through syntax. As phrases expressing systematically a certain meaning through general syntactic mechanisms cannot be classified as restricted collocations, TGRs in this class should be classified as systematic gaps: alarming-ly = *in modo allarmante* (where alarming=*allarmante*)

The 6 classes mentioned above are interesting for our purposes because their members can be automatically enumerated. Here follows a brief descriptions of the procedure used for each class:

- 1 The presence of simple words in TGRs is detected by simply counting the number of words composing each translation equivalent. For Italian, we have included in this class also collocations formed by a verb followed by a preposition. In Italian such preposition is never part of the verb but introduces an argument (grammatical collocations). For English it is not possible to distinguish automatically between grammatical collocations and phrasal verbs. For this reason we have considered all of them in the second class (collocations).
- 2 The class of TGRs including at least a collocation is defined on the basis of the available resources. English collocations can be found as lexical units making up the synsets of the PWN or as headwords in the English section of the Collins. Italian collocations are found in the DISC on-line monolingual dictionary. To check if a phrase is a collocation in the DISC, a procedure computes the quotation form for each of the words of the phrase, and looks up the corresponding entries to find if the phrase is listed in the collocation section of the entry.
- 3 We defined a list of support verb constructions based on the relevant literature. See (Renzi, 1986), (der Wouden, 1994), (Heid, 1994). A procedure looks for a matching between a complex translation equivalent and one of the following two patterns: SupportVerb (Prep) (Art) Noun, SupportVerb Adjective. For English we considered as SupportVerb 7 items, i.e: *do, make, have, take, put, give, go*. For Italian we selected 5 verbs: *fare, prendere, mettere, avere, dare*.

4,5 These classes simply require counting the content words which are part of the headword and of its translation equivalents. Content words are all those not included in a closed list of functional words.

- 6 To detect systematic gaps a procedure looks for a morphological pattern in the headword coupled with a syntactic pattern in the translation equivalent. Here are the relevant couplings. For English to Italian:

ADV(-ly) => *in + modo/maniera + ADJ*

ADV(-ly) => *con + N*.

V => *fare + V*

The third pattern is explained by the fact that in Italian causative alternations are very rare: most often Italian expresses the causative sense through a paraphrase (start = *far partire*).

For Italian to English:

ADV(-mente) => *in a + ADJ + way/fashion/manner*

ADV(-mente) => *with + N*.

REFLEXIVE-V => *become + V*

V starting with the *re-* prefix => *V + again*

The LG procedure draws a number of inferences to derive the three main groups from the six support classes.

- If a TGR falls into the classes 1 or 2, we know for sure that it is a lexical unit.
- If a TGR belongs to the 6th class we know for sure that it is a lexical gap.
- If a TGR belongs to the classes 3, 4 or 5 we can infer with a reasonable degree of confidence that it is a lexical unit. The degree of confidence of the inference has been measured by manually checking a statistically relevant sample of the classes. The margin of error is always lower than 10%.
- All TGRs that cannot be included in one of the support classes need to be manually classified.

The mapping from the support classes to the main groups can be represented through the decision tree in Figure 2. Decisions in the tree are ordered following a decreasing degree of certainty (which is never lesser than 90%).

Here follows some examples, showing how the LG procedure works.

- (2) **abduction** *n* rapimento, sequestro di persona.

In the Collins, the English noun *abduction* has one TGR made up of two translation equivalents: *rapimento* and *sequestro di persona*. The first translation equivalent is a simple word, while the second one is listed as a collocation in the Italian monolingual dictionaries. Thus the above TGR belongs to both the first and second support class. The decision tree classifies the TGR among the lexical unit in the first step.

- (3) **affannosamente** *avv* (*respirare*) with difficulty; (*freneticamente*) anxiously.

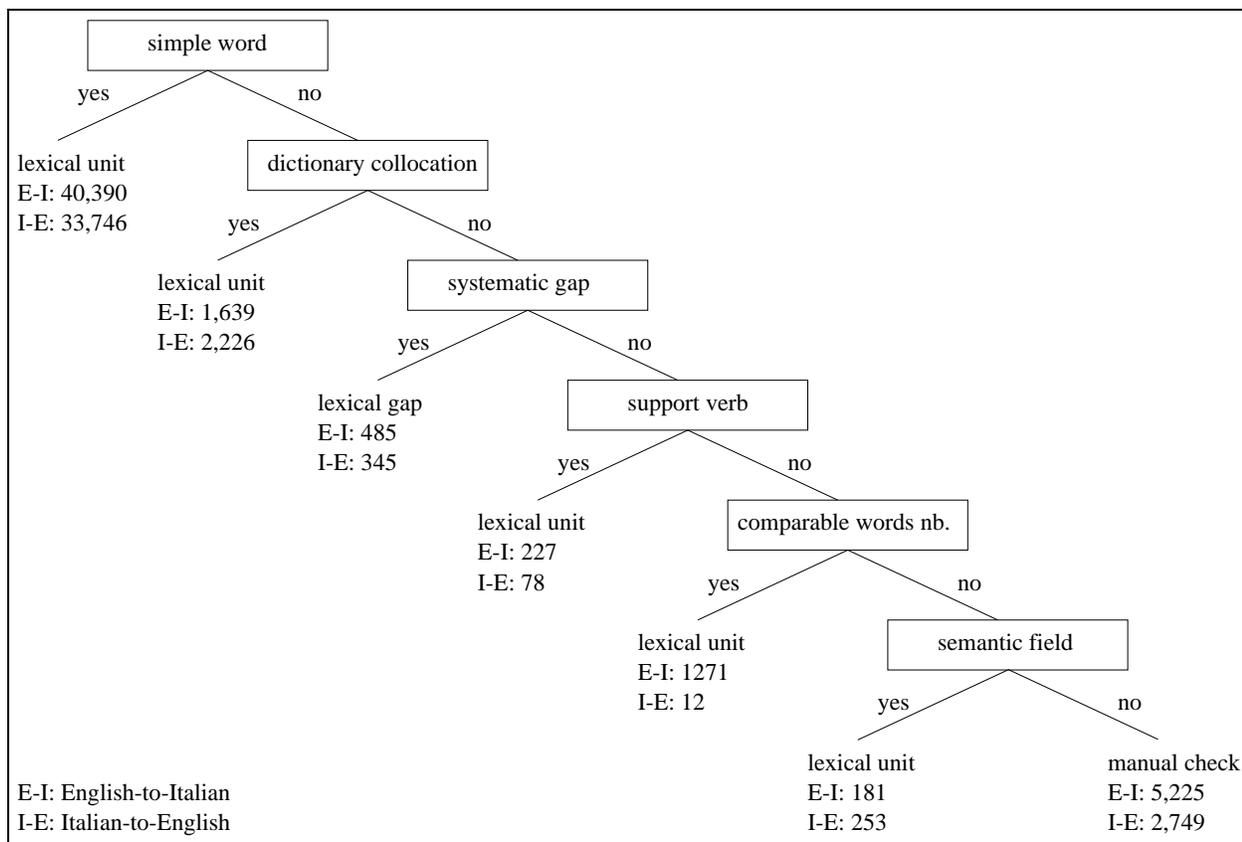


Figure 2: Decision tree.

The Collins entry for the the Italian adverb *affannosamente* lists two TGRs. The LG procedure analyses the translation equivalent contained in the first TGR and verifies that the couple made up of the Italian adverb (derived from the adjective *affannoso* and the suffix *-mente*) and the English adverbial phrase (*with + N*) matches the pattern of a systematic gap. As the adverbial phrase is the only translation equivalent given in the TGR, the decision tree classifies the first sense *affannosamente* (said of breathing) among lexical gaps. The second TGR is classified instead as a lexical unit.

- (4) **dam** 1. *n* ... 2. *vt* (*dam up: river*) sbarrare con una diga; (: *lake*) costruire una diga su; (: *fig*) arginare, frenare.

The Collins lists three TGRs for the English verb *to dam*. The LG procedure cannot include the first and second TGRs into any of the 6 support classes and thus assigns them to the third main group (TGRs that need human intervention). In a second phase, the lexicographer has classified them as lexical gaps. Also Example 5 below is put in the third main class by the LG procedure. However in this case the lexicographer has classified the translation equivalent as a restricted collocation even if it does not occur in current monolingual Italian dictionaries. The decision is justified on the grounds that the same concept could be expressed as *campo aereo* but only the reported expression is actually used, see section 2.

- (5) **airfield** *n* campo d'aviazione.

#### 4. Discussion and conclusions

The LG procedure has been applied to 39,405 TGRs of the Italian-to-English section (senses of Italian headwords belonging to the 40,000 words list) and to 49,418 TGRs of the English-to-Italian section (senses of English headwords belonging to PWN). The results of the procedure show that only 7.0% of all TGRs need manual control in the Italian-to-English section and 10,6% in the English-to-Italian section. Tables 1. and 2. illustrate the results of the LG procedure in more detail. Given the different size of the two starting sets of TGRs, for a comparison of the results of the two sections of the dictionary it is more appropriate to take into consideration percentages rather than absolute figures. For example, a comparison of the percentage of adverb gaps in the two sections of the dictionary shows that English tends to lexicalize more often than Italian.

Thanks to the results of the procedure, the lexicographers dealing with lexical gaps during the construction of IWN were able to focus on a restricted set of TRGs pointing with high probability to lexical gaps.

The LG procedure has currently some limitations with regard to both coverage and precision. Coverage problems are due to the Collins size. In the English section, the Collins contains 55,777 translation groups while PWN contains 99,642 synsets. On the other hand, the Italian section the Collins contains only the senses of 26,078 words out of the 40,000 IWN words. The coverage however may be even lower than that suggested by these figures, because not all the TGRs of the English section of Collins correspond to

Translation Groups	Lexical Units	%	Gaps	%	Manual Check	%
Nouns (26,266)	24,118	91.8	0	0.0	2,148	8.2
Verbs (9,239)	7,994	86.5	126	1.4	1,119	12.1
Adjectives (11,251)	9,653	85.8	0	0.0	1,598	14.2
Adverbs (2,662)	1,943	73.0	359	13.5	360	13.5
Total (49,418)	43,708	88.4	485	1.0	5,225	10.6

Table 1: English-to-Italian main groups

Translation Groups	Lexical Units	%	Gaps	%	Manual Check	%
Nouns (22,068)	20,571	93.2	0	0.0	1,497	6.8
Verbs (9,438)	8,143	86.3	338	3.5	961	10.2
Adjectives (7,264)	7,027	96.7	0	0.0	237	3.3
Adverbs (635)	574	90.4	7	1.1	54	8.5
Total (39,405)	36,315	92.1	345	0.9	2,749	7.0

Table 2: Italian-to-English main groups

PWN synsets. For instance the English entry *arterial* has two TGRs:

- (6) **arterial** *adj* (*Anat*) arterioso/a; (*road etc*) di grande comunicazione.

The second TGR has no correspondent in PWN. Similar considerations can be done for the Italian section of Collins with respect to the Italian senses.

Precision can be affected by two kinds of factors. One is implicit in the procedure: some of the support classes (from 3 to 5) are defined on a probabilistic basis. Given the degree of certainty of the classes described in section 3 and given the results of the classification procedure we estimate a margin of error lesser than 10% for the support classes 3, 4, 5 which account for only 8% of the translation equivalents. The other factor is related to the Collins, which sometimes gives incomplete translation equivalents, i.e. it gives a phrase as translation equivalent when a simple word exists. For instance the English adjective *unreliable* is translated only with the free combination *su cui non si può contare o fare affidamento*; however there does exist an alternative translation equivalent (*inaffidabile*) which is a simple word. These inaccuracies in the Collins produce false gaps.

Before concluding let us notice that our work provides as a further result an approximate quantitative evaluation of lexical gaps, showing that the English and Italian lexica are highly comparable. To the extent that our results can be confirmed, it can be concluded that our work provided a strong empirical support to the MultiWordNet model. Moreover, with appropriate bilingual resources, this methodology could be applied to other languages.

## 5. References

- Bentivogli, L. and E. Pianta, 2000. Looking for lexical gaps. In *Proceedings of Euralex-2000 International Congress*.
- Ciravegna, F., B. Magnini, E. Pianta, and C. Strapparava, 1994. A project for the construction of an Italian lexical knowledge base in the framework of wordnet. Technical Report 9406-15, ITC-irst.
- Cowie, A. P., 1981. The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, 2 (3):223–235.
- der Wouden, T. Van, 1994. *Negative Contexts*. Ph.D. thesis, Groningen University.
- Fellbaum, C. (ed.), 1998. *WordNet: An electronic lexical database*. Cambridge (Massachusetts): The MIT Press.
- Heid, U., 1994. On ways words work together: Research topics in lexical combinatorics. In *Proceedings of Euralex-94 International Congress*.
- Hutchins, W. J. and H. L. Somers, 1992. *An introduction to Machine Translation*. London: Academic Press.
- Renzi, L., 1986. *Grande grammatica italiana di consultazione*, volume 1. Bologna: Il Mulino.
- Vossen, P., 1996. Right or wrong: Combining lexical resources in the eurowordnet project. In *Proceedings of Euralex-96 International Congress*.
- Vossen, P. (ed.), 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Kluwer Academic.