Looking for lexical gaps

Luisa BENTIVOGLI and Emanuele PIANTA,
Trento, Italy

Abstract
In this paper we present the results of a quantitative evaluation of the discrepancies between the Italian and English lexica in terms of lexical gaps. This evaluation has been carried out in the context of MultiWordNet, an ongoing project that aims at building a multilingual lexical database. The quantitative evaluation of the English-to-Italian lexical gaps shows that the English and Italian lexica are highly comparable and gives empirical support to the MultiWordNet model.

1 Introduction

The literature on contrastive analysis is rich of classifications and exemplifications of the lexical divergencies that can occur between pairs of languages. It is quite difficult, instead, to find systematic quantitative analyses of the differences between lexica. This comes with no surprise given the huge amount of lexicographic work that a quantitative evaluation of lexical divergencies implies. However, thanks to the availability of machine readable dictionaries, this is now a more feasible task. In this paper we present the results of a semi- automatic quantitative evaluation of the discrepancies between the Italian and English lexica in terms of lexical gaps.

The evaluation has been carried out in the context of MultiWordNet (MWN), an ongoing project that aims at building a multilingual lexical database [Ciravegna et al. 1994]. 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 MWN project, consists in building language specific wordnets keeping as much as possible of the semantic relations available in the Princeton WordNet (PWN) [Fellbaum 1998]. 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. This strategy makes sense only if the structural differences between English and the lexicon of the other language are small, i.e. there are relatively few cases when the synset of one language has no correspondent in the other language. This motivates our interest in a quantitative evaluation of the lexical divergencies between pairs of languages. See [Bentivogli et al. 2000] for a discussion of the pros and cons of the EWN and MWN approaches.

In the rest of this paper we will present a methodology for the semi-automatic quantitative evaluation of lexical gaps mostly based on a bilingual machine readable dictionary. The methodology, applied to a contrastive analysis of English and Italian, shows that the number of English lexical concepts which do not have a lexicalized translation equivalent in Italian amounts to less than 8 per cent. This result has an autonomous lexicographic relevance and gives empirical evidence of

the feasibility of the MWM model. Moreover, with appropriate bilingual resources, this methodology can be applied to other languages. From a practical point of view, this approach reduces the human lexicographic work and speeds up the MWN building process, showing in advance where there will be problems in the mapping between concepts across the two wordnets and therefore manual intervention will be necessary.

2 Lexical idiosyncrasies

The literature on contrastive analysis shows that, given a source and a target language, various types of idiosyncrasies (or discrepancies) can occur. Here is a summary of the most common:

- *Syntactic divergencies*: the translation equivalent (TE) does not have the same syntactic ordering properties of the source language word. Ex: the man entered the room = *l'uomo entrò nella stanza*. See [Dorr 1993].

- *Lexicalization differences*: the source and target languages lexicalize the same concept with a different kind of lexical unit (word, compound or collocation) or one of the two languages has no lexicalization for a concept (lexical unit vs. free combination of words). In the latter case we have a so-called *lexical gap*. Ex: private = *soldato semplice* (collocation); to dam = *sbarrare con una diga* (gap). See [Marello 1989, Vinay and Darbelnet 1977].

- *Divergences in connotation*: the TE fails to reproduce all the nuances expressed by the source language word. Ex: fanciullo (literary) = *child*. See [Brown 1995].

- *Denotation differences*: the denotation of the the source language word only partially overlaps the denotation of the TE. Ex: convento = *monastery* (for monks), *convent* (for nuns). See [Lo Cascio et al. 1995].

Only some of these idiosyncrasies are relevant for the information coded in MWN. In MWN, two synsets belonging to two different languages are correspondent if the words of one language are cross-linguistic synonyms of the words of the other language. For this reason we focused our work on those idiosyncrasies that imply a lack of cross-linguistic synonymy. MWN adopts for both the intra- and cross-linguistic synonymy the same criteria that are used to build the PWN synsets, i.e.:

- only cognitive synonymy is required, i.e. words that have the same denotation but different connotation or syntactic behaviour are taken to be synonyms;

- idioms and restricted collocations are considered lexical units and thus can be synonymous with simple or compound words.

Given such criteria only two of the above mentioned idiosyncrasies imply lack of cross-linguistic synonymy and thus are relevant for our purposes. These are:

- lexical gaps: a language expresses in a lexical unit what the other language expresses with a free combination of words (borrower = *chi prende in prestito*);

- denotation differences: the TE of a source language exists but it is more general (generalization) or more specific (specification). In the former case the TE is a sort of cross-linguistic hypernym of the source language word and in the latter case it is a cross-linguistic hyponym (*bell* ≅ (small/electric bell) *campanello* + (church bell) *campana* + (on cats) *sonaglio*).

In this paper we take into consideration only lexical gaps from English to Italian and show how it is possible to develop a procedure for identifying them in a semi-automatic way, using knowledge available in electronic dictionaries.

3 What is a lexical gap?

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*. For a suitable evaluation of lexical gaps we first need to distinguish *idioms* and *restricted collocations* from free combinations of words. While idioms and restricted collocations can be considered as composite lexical units, free combinations do not belong to a language lexicon and imply a lexical gap. We adopt the following definitions [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 built compositionally. They allow only a limited substitution of its component words, which have a kind of semantic cohesion mainly due to use. Collocations spring to mind readily, are psychologically salient, and do not usually have a literal translation in other languages.

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

In practice, the boundaries between idioms, restricted collocations and free combinations are not clear-cut. In many cases a distinction can be drawn relying on knowledge contained in dictionaries that explicitly mark idioms and collocations. Also, the three sets 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*.

4 Finding lexical gaps

To make a quantitative evaluation of the English-to-Italian lexical gaps we used the electronic version of the Collins bilingual dictionary and of the DISC monolingual Italian dictionary. The bilingual Collins is a medium size dictionary, including in the English section 40,959 headwords and 60,901 translation groups. By *translation group* (TGR) we mean a group of Italian synonyms translating a sense of an English word. In bilingual dictionaries, TGRs are usually separated by semicolons. We take them as the relevant sense unit as they correspond to WordNet senses. In the following example "wood", as a noun, has 5 TGRs.

wood [wUd] **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 .

Given this information, we tried to determine in how many cases the Italian TE is a free combination, which implies a gap. Of course the Collins does not contain all English lexical units. However the 60,901 TGRs listed in the Collins can be considered a significant sample to estimate the percentage of English-to-Italian lexical gaps.

To this extent we developed a procedure that aims at classifying all Italian TGRs according to the kind of TEs which constitute them: simple words, collocations or free combinations. As a preliminary step, we singled out a number of subclasses whose members can be enumerated automatically. These subclasses are relevant for our purposes but may also be significant from a general lexicographic point of view.

- The TGR includes one *simple word*. If an English word has a TGR in this subclass, we can exclude the existence of a lexical gap. Grammatical collocations, i.e. collocations formed by a word followed by a preposition introducing an argument (answer = *rispondere a*), have been considered equivalent to simple words. See [Benson 1986].

- The TGR includes a phrase explicitly listed as *collocation in the Italian monolingual dictionary*. Given the MWN criteria for deciding whether a phrase is a lexical unit or not, also TGRs in this subclass allow us to exclude the existence of a lexical gap. Note that, if a phrase is not listed as collocation, we cannot infer that it is a free combination. It is known in fact that dictionaries list only a part of existing collocations [Cowie 1981].

- The TGR includes only a phrase explicitly marked as free combination of words in the Collins. The Collins marks some TEs with a special tag (TRGL = translation gloss) to signal that the TE is an explanation of the meaning of the English word rather than a true translation (butterscotch = *caramella dura a base di burro e zucchero di canna)*. Note that TEs tagged as translation glosses by the Collins cover only a part of TEs that are free combinations of words.

- The TGR includes a multiword adverb with the structure *in* + *modo*/*maniera* + ADJ or *con* + N, or a multiword verb with structure *fare* + V. This class includes a number of gaps corresponding in a *systematic* way to English simple words with specific morphological or

syntactic features (alarmingly = *in modo allarmante*, admiringly = *con ammirazione*). The Italian adverbial pattern corresponds to a productive English mechanism deriving adverbs from adjectives by adding the "ly" suffix. As for the verbal pattern, this is explained by the fact that a considerable number of verbs can occur as either inchoatives or causatives in English but can only be used as inchoatives in Italian. Italian expresses the causative sense through a paraphrase (start = *far partire*).

- The TGR includes a verb phrase with the following structure: Support-V + (Art) + N + (Prep) or Support-V + Prep + (Art) + N [Renzi 1988]. The so called *support verb constructions* are interesting for us because they represent a typical structural pattern in which collocations can be found [Heid 1994]. For instance: to brief = *dare istruzioni a*.

- The TGR includes a multiword phrase with a *number of words comparable to that of the English headword*. This happens when the number is the same (roller coaster = *montagne russe*) or the English entry is formed by two words and the TE by three. The latter situation occurs when a nominal specifier is translated in Italian with a prepositional phrase, see "agony column = *posta del cuore*". By manually checking a significative sample of this class, we noticed that, as a matter of fact, this class includes collocations in 90% of cases.

- The TGR includes a two word TE and is labelled with a gloss specifying a *semantic field*. In specialized domains, many fixed expressions can be classified as idioms or collocations (armour (MIL) = *mezzi blindati*). In a significative sample of this class, TGRs included collocations in more than 90% of all cases.

On the basis of these seven subclasses we designed a procedure that classifies all TGRs in three main classes: lexical units, lexical gaps and TGRs that need to be manually checked. This procedure can be represented as a decision tree as illustrated in figure 1. The subclasses are ordered by first taking into account those allowing a more certain decision. The degree of certainty is never lesser than 90%.

The TGRs requiring human check have been manually classified as collocations or gaps following the criteria illustrated in Section 3. Table 1 summarizes the quantitative evaluation of English-to-Italian lexical gaps summing up the results of the application of the decision tree and the manual check.

Translation Groups		Simple ws.	%	Collocations	%	Gaps	%
Nouns	(31,978)	23,800	74.4	6,394	20.0	1,784	5.6
Verbs	(12,939)	10,226	79.0	1,755	13,6	958	7.4
Adjectives	(13,113)	10,455	79.7	1,217	9.3	1,441	11.0
Adverbs	(2,871)	1,890	65.8	426	14.9	555	19.3
Total	(60,901)	*46,371*	*76.1*	*9,792*	*16.1*	*4,738*	**7.8**

Table 1: Distribution of lexical gaps

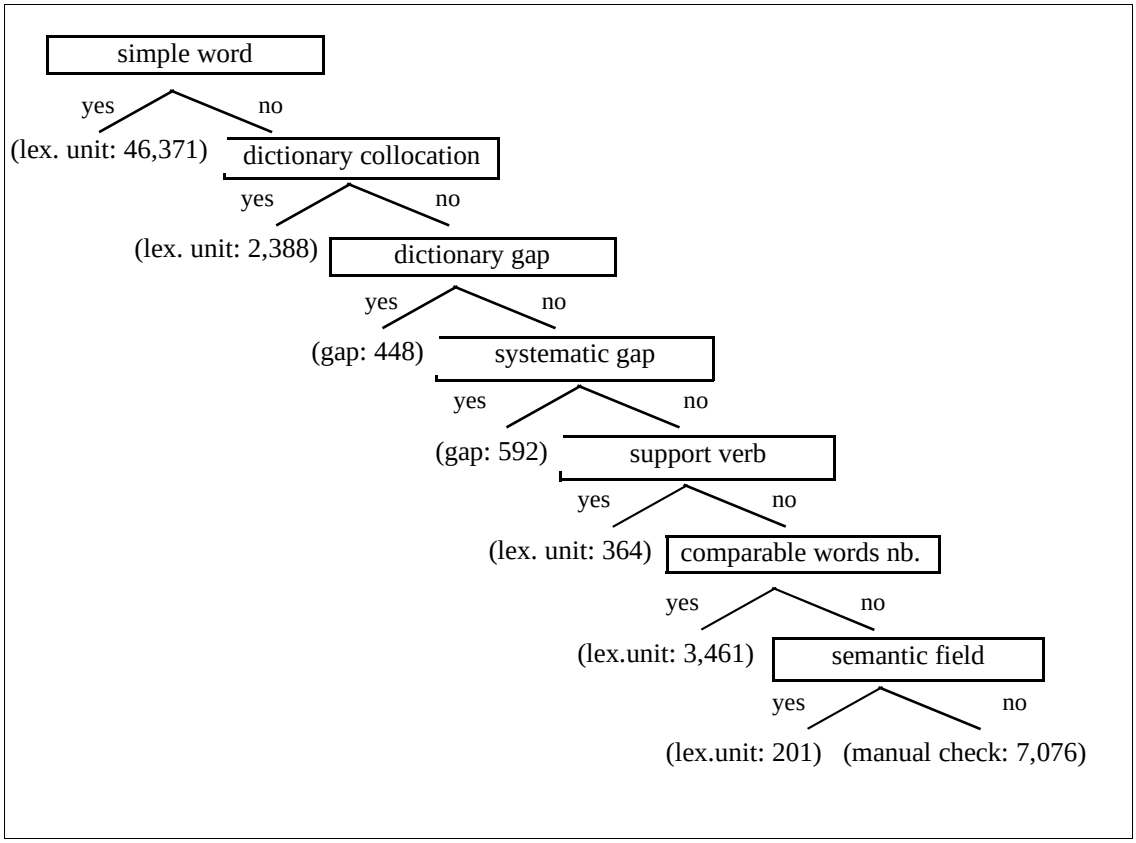

Figure 1: A decision tree to identify translation groups that require manual check

5 Conclusion and future work

In this paper we have presented a general methodology to find lexical gaps in a semi-automatic way, using knowledge available in electronic dictionaries. The results of the research give a quantification of the English-to-Italian lexical gaps on a large scale and therefore an indication about the overlapping of the lexical structures of the two languages. The procedure identified 4,738 lexical gaps out of 60,901 senses taken into consideration (7.8% of the total), showing a good overlapping between the two lexical structures. These results provide empirical support to the MultiWordNet model. Moreover, the procedure can reduce the lexicographic work needed to build aligned wordnets, by showing in advance where there will be problems in the mapping between the lexical concepts of two languages [Bentivogli et al. 2000].
We are currently investigating Italian-to-English lexical gaps and we plan to develop an automatic procedure to identify another type of idiosyncrasy, i.e. the denotation differences. Also, we plan to apply the proposed methodology to a large size bilingual dictionary.

References

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

[2] Bentivogli, L., Pianta, E. and Pianesi F. (2000). "Coping with lexical gaps when building aligned multilingual wordnets", in Proceedings of LREC 2000, Athens, Greece.

[3] Brown, V., Mendes, E. and Natali, G. (1995). *False Friends and Bugs and Bugbears*. Bologna: Zanichelli.

[4] Ciravegna, F., Magnini, B., Pianta, E. and Strapparava, C. (1994). *A project for the construction of an Italian Lexical Knowledge Base in the framework of WordNet,* Technical Report 9406-15, ITC-irst.

[5] Cowie, A. P. (1981). "The treatment of collocations and idioms in learner's dictionaries", in *Applied Linguistics*, 2(3), pp. 223-235.

[6] Dorr, B. J. (1993). "The use of lexical semantics in interlingual machine translation", in *Machine Translation*, 7, 3, pp. 135-193.

[7] Fellbaum, C. (ed.) (1998). *WordNet : An electronic lexical database.* Cambridge (Mass.): The MIT Press.

[8] Heid, U. (1994). "On ways words work together: research topics in lexical combinatorics", in *Proceedings of Euralex-94 International Congress*.

[9] Lo Cascio, V. and Boraschi, P. and Corda, A. (1995). "Correspondence between senses and translation equivalents: automatic reversal of a bilingual dictionary", in Thelen, M. e Lewandowska-Tomaszczyk (eds), *Translation and meaning*, part 3, pp. 221-231. Maastricht: Universitaire Pers.

[10] Marello, C. (1989). *Dizionari bilingui.* Bologna: Zanichelli.

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

[12] Vinay, J. P. and Darbelnet, J. (1977). *Stylistique comparée du français et de l'anglais.* Montréal: Didier.

[13] Vossen, P. (ed.) (1998). *EuroWordNet :A multilingual database with lexical semantic networks.* Dordrecht: Kluwer Academic.