Automated Translation between Lexicon and Corpora

Elisabetta Gola †
egola@unica.it

Stefano Federici †
sfederici@unica.it

Nilda Ruimy ‡
nilda.ruimy@ilc.cnr.it

John Wade †
jwade@unica.it

ABSTRACT

In this work we will show the role of lexical resources in machine translation processes, giving several examples after a brief overview of Machine Translation studies. Then we will advocate the need for a richer lexicon in MT processes and sketch a methodology to obtain it through a mix of corpus-based and machine learning approaches.

Keywords: Translation, Corpus linguistics, Natural Language Processing, Machine Learning, Lexical Knowledge.

Introduction

Machine Translation (MT) is one of the most challenging issues for Artificial Intelligence (AI) applied to language, which we here refer to as Natural Language Processing (NLP). The history of MT shows, indeed, that a

† Department of Education, Psychology, Philosophy, University of Cagliari, Italy.
‡ Istituto di Linguistica Computazionale Antonio Zampolli, National Research Council, Pisa, Italy.

translation process presupposes a good understanding of the text to be translated.

In this paper, we will argue for the relevance of the lexicon in the translation process and the need to dispose of wide coverage and high quality lexical resources. The access to a rich lexical knowledge is in fact a fundamental requirement for a computational system to correctly analyze a text and generate its translation.

To this purpose, we will present an Italian lexicon that meets the requirements of MT systems, and we will show how its lexical information can be used in a translation process.

It must, however, be emphasized that building a large coverage lexicon is a very costly and time consuming process. That is the reason why Computational Lexicography is today mostly oriented toward the development of methodologies and strategies that make the creation of lexicons easier and faster with the automatic acquisition of data from corpora, from the Web, or by induction from existing resources. In this paper, we will show a bootstrapping method, based on a machine learning technique, that allows us to build at the same time a corpus-based lexicon and a tagged corpus, that grow incrementally together in a semi-automated way.

1. Machine Translation: historical overview and state of the art

Since the beginning of Artificial Intelligence (AI) and Natural Language Processing (NLP) , studies and research were devoted to realizing the dream of Machine Translation.

During the first decades of MT research, an articulated panorama of methodologies and strategies started shaping. Classifying all the approaches is almost impossible, given that perspectives change along with the adopted parameters.

> There is now a variety of MT systems which almost defies any neat classification. It is still often legitimate to apply the labels of the 1960s: practical vs. theoretical, empirical vs. perfectionist, direct vs. indirect, interlingual and transfer. But now there are new labels and new perspectives: interactive vs. fully automatic, 'try-anything' systems vs. 'restricted language' systems, mainframe systems vs. microcomputer or word-processor systems, AI-based systems vs. linguistics-oriented systems (Hutchins, 1986, p. 19).

To our purposes, we will focus on the distinction between direct and indirect strategies that belong respectively to first and second generation MT systems.

Until the sixties, MT systems, called first generation systems, followed a so-called direct strategy, in which a direct correspondence was established between Source Language and Target Language (henceforth, SL and TL). In this strategy, the SL was only analyzed from a morphological point of view. The output of the morphological analysis constituted the access point to the bilingual lexicon. In this way, a text could only be translated word-by-word. This strategy failed therefore to cope with the translation of ambiguous sentences or sentences with different SL and TL syntactic structures, such as the Italian sentence Questo ragazzo piace a Maria (lit. this boy likes to Maria), whose English structure: Maria likes this boy is quite different.

During the sixties, the second generation systems adopted an indirect strategy, in which two approaches were followed. Firstly, a two-phase process defined as the Interlingua approach and, secondly, a three-phase process defined as the Transfer approach.

In the Interlingua approach, a formal, abstract and language-independent representation interfaces source and target languages: a SL text is analyzed into an interlingual representation which is then synthesized into a TL text. In this view, a conceptual lexicon is required, the building of which is an extremely complex and controversial task.

For this reason, more realistic strategies, based on Transfer, are adopted. In this case, the translation steps are the following:

- analysis of a SL text into a SL formal representation;
- transfer of the SL formal representation into a TL formal representation;
- generation of a TL text from the TL formal representation.

In the Transfer approach, the structural analysis of the SL text is performed in different steps and leads to the building of a formal representation of the SL structures that, in the transfer phase, is mapped onto a formal representation of the TL structures. As to the lexical transfer, the SL lexical units are translated into TL lexical units, using an electronic bilingual dictionary. During the

synthesis phase, the TL formal representation is turned, following different steps, into a TL text. In this perspective, cultural aspects of different languages are taken into account.

In spite of this innovation, disappointment with the feasibility of MT was growing, due to the "semantic barriers" that researchers encountered and that proved difficult to overcome.

Furthermore, in 1964, the US government sponsors asked the National Science Foundation to constitute a committee in order to evaluate the progress made in NLP in general and in the MT state-of-the-art in particular. The commission produced in 1966 a "(in)famous report", as John Hutchings (1996) defined it, the ALPAC report (from the name of the committee: Automatic Language Processing Advisory Committee). The ALPAC report stated that MT systems were slower, less precise and more expensive than human translators. The verdict then was: "there is no immediate or predictable prospect of useful machine translation" (ALPAC, 1966).

It should be noticed, though, that the ALPAC committee, in its report, took into consideration only direct strategy systems, evaluating them negatively. For the next ten years, this assessment caused the U.S. Government to reduce its funding in this area dramatically. As a direct consequence, research in this field stopped in the US for over a decade, while it carried on in Canada, Germany and France.

It is only in the middle of the Seventies that we find a renewed interest for automated translation, with the emergence of third generation systems based on Artificial Intelligence.

Starting from the 1990s, a new methodological approach emerges, that makes use of large bodies of text (corpora) (Hunston, 2002). Among the corpus-based systems, the most common approaches are statistics-based systems (SBMT) and example-based systems (EBMT).

SBMT follows strategies in which SL and TL sentences are tentatively aligned on the basis of the probability that each word in the SL sentence corresponds to one or more words in the TL sentence. On the contrary, the example-based methodology, suggested by Nagao in 1984 but implemented only in the 1990s, gives a translation by analogy, comparing the input sentence with a bilingual dictionary that includes examples and matching those that are more similar to the input (Nagao, 1984; Brown, 1999, Turcato et al., 1999).

In the same years, the rule-based systems move away from syntax-based representations to more 'lexicalist' approaches. At its extreme, the essence of

the lexicalist approach in MT system design is to reduce transfer rules to simple bilingual lexical equivalences. Such a drastic reduction was first put forward in the CRITTER project (Isabelle et al., 1988). The approach has been explored in the ACQUILEX project devoted primarily to the construction of multilingual lexicons for transfer-based MT (Sanfilippo et al., 1992), and is probably best known as the ′shake-and-bake′ method described by Whitelock (1992). The requirement for structural representations - common to both transfer and interlingua approaches - is abandoned in favour of sets of semantic and syntactic constraints on lexical items. Translation involves the identification of TL lexical items which satisfy the semantic constraints attached to the SL lexical equivalents.

The ′bag′ of target lexical items is then ′shaken′ to generate an output text consistent with the syntax and semantics of the target language (Hutchins, 1993).

This 'lexicalist' turn led the MT community to an increasing interest for computational lexicons.

Today, Machine Translation systems usually follow either a corpus-based or a rule-based approach. In the first trend, we find statistical approaches and example-based approaches. In the second one, emphasis is given to lexical resources. In the following section of the paper, we will propose an integration of these two approaches.

2. Relevance of the lexicon in MT

In order to produce a good translation it is necessary to understand correctly the input text. It is precisely for this reason that Machine Translation is deemed one of the most difficult tasks in the field of AI language applications. Any translation process implies, in fact, the resolution of a whole range of problems regarding both the analysis and the generation of texts. In this context, the lexicon plays a crucial role. A robust translation system should be able to cope with a wide range of issues inherent to the complexity of natural language, such as the various types of ambiguity, non literal uses, polysemy and so on. A poor lexicon fails to support these challenging tasks.

3. Lexicon and lexical problems in MT

Table 1 illustrates some of the most typical and frequent lexical problems that are encountered during a translation process and that a lexicon tailored for an MT system should be able to deal with.

The lexicons used in MT systems must have wide coverage and provide, for each lexical entry, a large range of rich and various information spanning all levels of linguistic description.

Direct strategy MT systems used a unique, very complex bilingual lexicon containing all grammatical information concerning both the SL and the TL lexical units, as well as the conditions for selecting the appropriate translation in case there are different alternatives possible.

Transfer-based MT systems, by contrast, use different monolingual lexicons (morphological, syntactic and semantic) containing all relevant information for each level of linguistic description for both the analysis and generation phases. In the transfer phase a bilingual lexicon is used. The transfer bilingual lexicon consists of lexical rules setting i) the correspondences between the lexical units described in the SL and TL monolingual semantic lexicons and ii) the conditions imposed on those equivalences. For example, in case of a SL word translatable by different TL words, the lexical transfer rule selects the appropriate TL equivalent, on the basis of the information provided by the two translational equivalents in their respective monolingual description.

In the domain of computational lexicography, a significant number of electronic lexical resources are now available, even though not all languages are equally represented. Most lexicons deal with a single level of linguistic description; some describe a unique part of speech or are strictly theory-dependent. Some are created in order to describe the vocabulary of a particular domain; others in order to meet the requirements of a specific application.

Level	Phenomenon	Example
Phonology	Homography	▪ it. *pésca* = en. *fishing* ▪ it. *pèsca* = en. *peach*
Morphology	Homonymy	▪ it. *legge, porta, sbarra*: N & V ▪ it. *appunto*: N. & ADV
Syntax	Syntagmatic realization	▪ en. *know* + NP = it. *conoscere* ▪ en. *know* + WH-clause = *sapere*
Semantics	Homonymy	▪ fr. *louer* = en. *to praise* ▪ fr. *louer* = en. *to rent*
	Polysemy	▪ en. *set up* = it. *piantare, erigere, mettere su, causare, installare, allestire, formare, etc.*
	Conceptual division	▪ en. *corner* = sp. *rincón* (internal), *esquina* (external)
	Lexical gaps	▪ it. *fuoricorso, consuocero*: not lexicalized in English and in French

Table 1.

Very few lexical resources, however, have the required features to be used in an MT system. As a matter of fact, besides providing a rich and various amount of information, a lexicon must guarantee completeness and coherence of the encoded lexical data. Moreover, it must be conceived as a dynamic resource, and not as a static and crystallized repertory of lexical information. Such a resource should be simple to update and expand not only manually but essentially through the automatic acquisition of information from textual resources, so as to reflect the continuous evolution of languages and to meet the new needs and answer the problematic issues which might emerge from the translation process. In this perspective, a generic lexical model and a modular architecture are essential for an electronic lexicon to be profitably exploitable.

A large computational lexical resource for the Italian language was developed at the *Istituto di Linguistica Computazionale* of the National

Research Council in Pisa from 1996 to 2003, which presents these characteristics.

4. The Lexical Resource

The computational lexicon PAROLE-SIMPLE-CLIPS (Ruimy *et al.*, 1998; 2002; 2003), elaborated in the framework of three different projects[1], provides a wide-coverage, four-level description of the Italian language. This lexical resource was built according to a multifunctional and multilingual perspective and in compliance with the international standards set out in the PAROLE-SIMPLE lexical model (Ruimy et al., 1998; Lenci et al., 2000).

This model, based on the EAGLES recommendations (San Filippo et al., 1998) and on an extended version of the GENELEX model (Antoni-Lay et al., 1994), is at the forefront of the field of Computational Lexicography for some outstanding and innovative features. The flexible architecture of the model as well as the building methodology allow the coherent encoding of a wide range of highly structured information, at the desired granularity level. Consensually adopted at a European level for the building of twelve harmonized monolingual electronic lexicons, the PAROLE-SIMPLE lexical model became a *de facto* standard and subsequently strongly inspired the ISO standard for NLP lexicons, the metamodel *Lexical Markup Framework*[2].

The PAROLE-SIMPLE-CLIPS lexicon offers, therefore, the outstanding advantage of being compatible with eleven other lexicons developed for European languages, with which it shares the theoretical and representational model, the working methodology as well as a kernel of entries.

The lexicon is articulated in four independent but interrelated modules, which correspond respectively to the phonological, morphological, syntactic and semantic levels of linguistic representation. The complete description of a lexical unit consists therefore in a minimum of four interconnected entries, each one providing a structured set of information relevant to the description level that hosts it.

A phonological entry accounts for the phonetic and phonological features of a lexical unit while a morphological entry informs on its grammatical category and inflectional paradigm. A syntactic entry describes both the

[1] The European projects LE-PAROLE and LE-SIMPLE and the Italian project *Corpora e Lessici dell'italiano Parlato e Scritto* (CLIPS)
[2] ISO-24613:2008

intrinsic and contextual properties of a lexical unit in *one* specific syntactic structure. The subcategorization frame is modelled in terms of syntactic category, grammatical function, optionality and morphosyntactic, syntactic and lexical restrictions of the governed elements. Systematic frame alternations, such as the causative-inchoative variation, are represented in a complex entry whereby the correspondence between the constituents of the two structures is specified.

The adopted theoretical framework for the representation of semantic information is based on the fundamental principles of the Generative Lexicon theory (Pustejovsky, 1995). In a generative lexicon, a semantic unit is modelled through four different levels of representation[3] that account for the componential aspect of meaning, define the type of event denoted, describe its semantic context and set its hierarchical position with respect to other lexicon units.

The semantic lexicon is structured in terms of an ontology of semantic types (the SIMPLE ontology). In a semantic entry, which encodes a single meaning of a lexeme, the membership in an ontological type represents the primary and most relevant information. Besides the ontological classification, the semantic unit is endowed with information concerning its domain of use; the type of event it denotes, where relevant; some distinctive semantic features; its links with other lexical units - among which synonymy and morphological derivation links - and membership in a class of regular polysemy. The semantic frame of predicative units is also described in terms of semantic role and selectional restrictions of the arguments.

To express the links holding among sense units, the SIMPLE lexicographers benefited from a remarkably efficient expressive means, the *Extended Qualia Structure*. This representational tool was derived from the *Qualia Structure*, a four-role[4] structure which is considered a mainstay in the Generative Lexicon theory for representing the multidimensionality of a word's meaning. The extended structure was created by defining, for each of the four Qualia roles, a subset of semantic relations. Such relations obviously allowed a much sharper expression of both the multidimensional aspect of a word sense and the nature of its syntagmatic and paradigmatic links to other lexical units. To give but one example, considering the telic role that informs

[3] Namely Qualia structure, Event Structure Argument Structure and Lexical Typing Structure.
[4] Formal, constitutive, agentive and telic.

about the function or purpose of an entity, the most appropriate relation may be selected among the following ones: 'used_for', 'used_by', 'used_as', 'used_against', 'is_the_activity_of', 'object_of_the_activity' and so on.

Moreover, in a new and revised version of the lexical-semantic database, called *Simple_PLUS*, the semantic representation has been enriched with significant information concerning the relationships holding between events and their participants and among co-participants in events (Ruimy, 2010).

This lexicon offers, therefore, a wide range of very rich and interesting information, especially at the semantic level. It is our deep conviction that an MT system could greatly benefit from such a wealth of lexical data, for both the granularity of the information provided and its explicit formulation.

5. Lexical Semantics for the resolution of some MT problems

A translation process presupposes the understanding of the many and various aspects that characterize the input text. Besides the morphological and syntactic aspects, it is necessary to disambiguate the logical form of the sentence, checking the coherence among semantic restrictions and preferences of words. To establish an equivalence between a source and a target text a translator should also understand other semantic and pragmatic aspects (for example conversational implicatures, metaphors, ironic contexts, etc.), that are not easily detectable. In the following, we will briefly show how Lexical Semantics plays a central role in the resolution of problems that typically emerge in Machine Translation.

Word sense ambiguity is a pervasive characteristic of natural language. It is one of the main reasons for poor performance of Information Retrieval systems. In MT, lexical ambiguity may occur both in the analysis and the transfer phases. Its resolution, which is therefore considered a major problem, requires a large amount of rich lexical knowledge.

5.1.1. Polysemy / homonymy and domain knowledge

A SL polysemic word or two SL homonyms may translate in two different ways according to their usage domain (see Table 2). Matching the information concerning the topic of the source text and the indication, in the monolingual lexicon, of the different domains of use of the ambiguous word enables the selection, in the bilingual lexicon, of the appropriate translation.

en.	*mouse*	➔	it.	(gen.) (inform.)	*topo* *mouse*
it.	*borsa*	➔	en.	(gen.) (econ.)	*bag* *stock exchange*
it.	*calcolo*	➔	en.	(gen.) (med.)	*calculation* *gallstone*

Table 2.

5.1.2. Polysemy / homonymy and ontological classification

The semantic classification of a word sense is generally sufficient to discriminate among its different meanings or among homonyms and therefore to enable the selection of the relevant one from its different possible translational equivalents, as shown in Table 3 for Italian-English and Italian-French translations.

Italian	➔	English	French
ala: [PART]	➔	*wing*	*aile*
ala: [BODY_PART]	➔	*wing*	*aile*
ala: [ROLE]	➔	*winger*	*ailier*
espresso [ARTIFACT_DRINK]	➔	*espresso*	*express*
espresso [VEHICLE]	➔	*express (train)*	*express*
espresso [SEMIOTIC_ARTIFACT]	➔	*express (letter)*	*exprès*

Table 3.

5.1.3. Polysemy / homonymy and contextual links

More complex situations emerge when two readings of a lemma cannot be disambiguated through their semantic classification or other paradigmatic information. In this case, syntagmatic and therefore contextual links may be used. In the following example reported in Table 4, means for selecting the appropriate translation are provided by the domain of use, but also by semantic relations linking each ambiguous term to the predicate denoting its function.

	Italian	➔	English
ferri_1	[INSTRUMENT] used_for *sferruzzare* (to knit)	➔	*knitting needles*
ferri_2	[INSTRUMENT] used_for *operare* (to operate)	➔	*surgical instruments*

Table 4.

5.1.4. Polysemy / homonymy and semantic frame

The semantic frame description may also provide clues for solving lexical ambiguities. Two homonym predicates may be distinguished by a different argument structure, either by the number of arguments they require (Table 5, first example) or by the semantic restrictions imposed on those arguments (Table 5, second example).

Italian	➔	English
avvertire1: arg0, arg1, arg2	➔	*to inform, to warn*
avvertire2: arg0, arg1	➔	*to feel, to notice*

Italian	➔	English
camminare1: arg0 = + animate	➔	*walk*
camminare2: arg0 = - animate	➔	*work*

Table 5.

It is worth noting that the whole range of lexical semantic information used for solving the above cases of ambiguity is encoded in the lexicon presented in the previous section.

6. The Corpus-based Approach

In order to briefly illustrate how a corpus-based approach may work, we have decided to focus our attention on one specific example, the translation of the English phrasal verb 'set up' into Italian, gathering our samples from electronic texts on the Web and analyzing them with a KeyWord in Context (KWIC) tool.

The experiment outlined here was carried out using the following procedure. A lexical item was selected, for the purposes of this analysis the

English phrasal verb 'set up' (Wade & Federici, 2006), since this item is problematic from a semantic point of view. It provides an interesting example of the highly polysemic nature of the English language, characterised by "remarkable range, flexibility and adaptability" (Crystal, 1988, p. 39). In this case the translator, for example, is required to consider the context specific nature of the lexical item (Eco, 2003, p. 29) and where areas of "inherent fuzziness" (Bell, 1991, p. 102) are found in establishing equivalence between one language and another. Indeed, 'set' alone has about 120 different meanings (cf. Collins Cobuild English Dictionary, 1995). With regard to 'set up', it was decided to first examine its meanings using a traditional bi-lingual dictionary Ragazzini-Zanichelli (2009). Secondly, a small sample of examples was collected from the web with a specifically designed search tool, followed by the manual examination and analysis of the gathered data and comparison with the information provided in the dictionary. The analysis was then extended through the analogical comparison of the initial manual analysis, allowing the further extraction of a wider sample of data.

To perform the kind of analysis described above, a tool was developed to acquire word-concordances directly from the web. The tool is a combination of several web/linguistic tools:

- a web spider that acquires a predefined number of web pages;
- a segmenter that splits acquired web pages;
- a rule-based lemmatiser;
- a KWIC (*KeyWord In Context*) tool;
- a self-learning analogy-based engine.

The web spider (cf. Federici, Wade, 2007) extracts web pages starting from a given web address, thus providing "a random snapshot of the current state of the Internet in a given language" (Sharoff, 2006, p. 437). The spider filters out all unneeded web overstructure (see Figure 1).

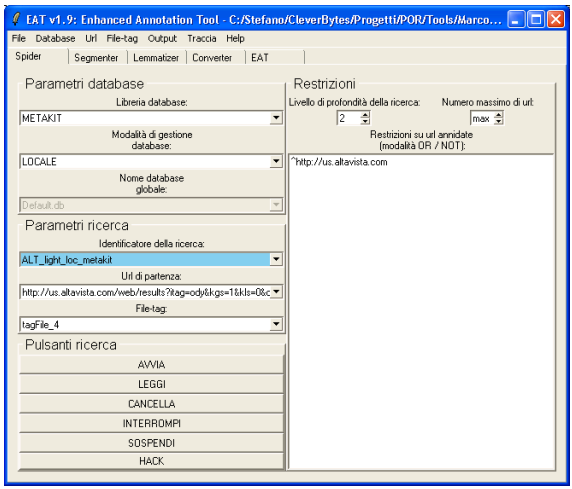

Figure 1

Then the lemmatiser associates each word form contained in the extracted web pages to the corresponding lemma. After the corpus has been cleaned and lemmatised, the KWIC will read the corpus by indexing all the lemmas. This is illustrated in Figure 2, where the word forms or lemmas are in the keyword area on the left (2a), and clicking on the keyword which is the focus of our interest the concordances are created (2b).

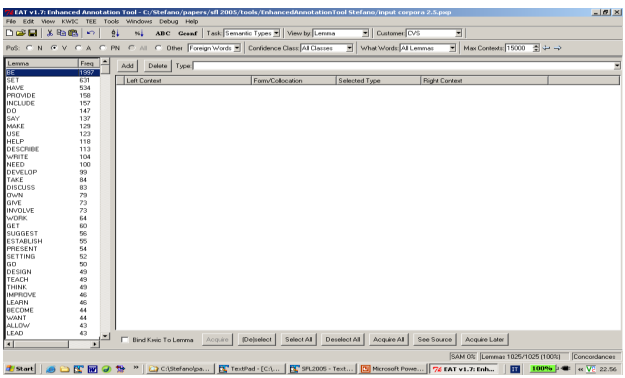

Figure 2a

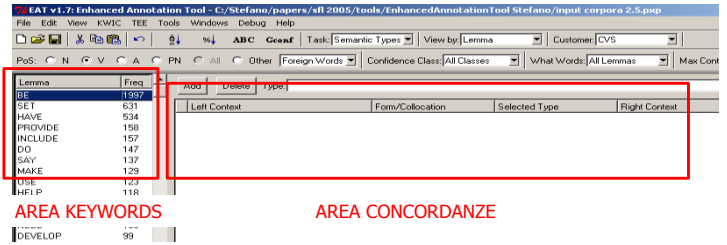

Figure 2b

While this approach is certainly useful as it enables the linguist to capture the real usage of a given word, it also suffers from a number of limitations:

- the manual analysis of data is extremely time-consuming;
- it is often not practicable to analyse all the examples, especially in large corpora, so only a selected number of examples are chosen as representative;
- there is the risk of human error and inconsistency in manual analysis.

7. Corpus vs. Dictionary

Our starting point was an analysis of the word senses provided in bi-lingual English-Italian dictionary Ragazzini-Zanichelli (2009). The result is illustrated in Figure 3.

to set up (verbo transitivo)

1. mettere su; alzare; erigere; piantare; montare; installare; allestire

2. mettere su; montare; installare; allestire

3. mettere su; mettere in piedi; istituire; fondare; costituire; formare; aprire (un ufficio); avviare (un'azienda)

4. sistemare; mettere (q.) in affari (o politica ecc.); aiutare (q.) finanziariamente (politicamente ecc.)

5. lanciare (un grido)

6. causare; provocare; dare l'avvio (o il via) a

7. stabilire (sport)

8. comporre (tipog.)

9. tesare, arridare (naut.)

10. (fam.) rimettere in salute (o in forze; in sesto); tirare su

11. (fam.) montare un'accusa contro (q.); incastrare; mettere contro; mettersi a fare; fornire; essere forte; essere ben fornito

Figure 3

It is to be noted that only a restricted number of entries provide contextualised examples of usage.

An initial analysis of the corpus created with the tools described above, on the other hand, reveals significantly richer contextualised source. In fact, it becomes immediately apparent that there are cases which are not included in the dictionary, such as the meaning 'creare', which is the appropriate translation of 'set up' in the case illustrated below:

> [...] useful information on the British Council's website, which was *set up* specifically for assistants to use in their placement countries. *("Foreign assistance"* by Katie Phipps, «Education Guardian online», August 23th 2005)

In an experiment that analysed 600 contexts of 'set up', only 8 out of 17 translations were attested in the dictionary. While it may be argued that the entries in the dictionary could be the most frequent usages of 'set up', it does not seem to be the case if we consider that the dictionary covers only about 47% of the translations of 'set up' occurring in our corpus (see Figure 4).

Translations present in the dictionary	Translations not present in the dictionary
ALLESTIRE	APPRONTARE
AVVIARE	ATTUARE
COSTITUIRE	CREARE
FONDARE	DEFINIRE
FORMARE	DIPINGERE
INSTALLARE	IMPOSTARE
ISTITUIRE	ORGANIZZARE
STABILIRE	PREPARARE
	REALIZZARE

Figure 4

From these analyses it emerges that examples extracted from real texts may be useful (i) to extend coverage of the lexicon and (ii) to refine semantic entries.

8. Extending the study: the 'bootstrapping' process

In order to extend the study and refine the data gathered, we need to use some type of Artificial Intelligence engine that (semi-)automatically carries out the annotation task. The procedure applied for the purposes of this study is called 'bootstrapping'.

In the first step a small portion of the corpus was annotated manually, assigning a translation to each sample (see Figure 5):

Manually annotated concordances:

1. [...] websites have also been *set up/CREARE* by the LSC [...]
2. [...] Websites have also been *set up/CREARE* and open days organised [...]
3. [...] an appeal panel has been *set up/COSTITUIRE* by the Dept. [...]
4. [...] a panel, task force, *set up/COSTITUIRE* by Harvard [...]

Figure 5

In the second step, the annotation is extended automatically to the remaining concordances for 'set up' in the corpus. At this stage it is found that not all of the translations assigned are correct (see Figure 6).

Manually annotated concordances:

1. [...] websites have also been *set up/CREARE* by the LSC [...]
2. [...] Websites have also been *set up/CREARE* and open days organised [...]
3. [...] an appeal panel has been *set up/COSTITUIRE* by the Dept. [...]
4. [...] a panel, task force, *set up/COSTITUIRE* by Harvard [...]

New concordances (Automatic annotation)

1. There are [...] much more useful information on the British Council's website, which was *set up/CREARE* [...] for assistants [...] (CORRECT)
2. [...] a committee was *set up/ISTITUIRE* to arrange [...] (WRONG)

Figure 6

During this automatic annotation step the first occurrence of 'set up' is automatically annotated as 'CREARE', which is correct, while the second is automatically annotated as 'ISTITUIRE', which is incorrect. This is because the algorithm in this case failed to provide the appropriate translation for lack of evidence.

In the third step, therefore, further manual revision is necessary. During this last phase the correct interpretation is manually assigned to those keywords that have been wrongly annotated.

9. Practical application of the procedure

We tested this procedure by setting up an experiment in which 600 contexts from a 1.5 million word corpus were manually annotated by assigning a translation to each concordance with 'set up', 400 new contexts were then automatically annotated and finally revised manually.

The results were encouraging, since the correctness of the automatically assigned translations was about 49%. That is, almost half of the time the

procedure assigns the correct translation, even starting from a relatively small set of training samples. This is acceptable when compared to the high number of possible translations (17) and less thorough baselines, such as the ones that could be obtained by assigning a random interpretation (1/17=6%) or just the most frequent one (that is "avviare", that accounts for only 12% of the cases).

Conclusions

In our hypothesis, the corpus-based process outlined above might prove to be very useful in enhancing lexical resources. This study aimed to create a dynamic cyclical process, in which the lexicon, in the case of our web-based experiment, is enhanced by a corpus-based analysis, and the corpus-based analysis can then be automatized thanks to the availability of richer and more precise lexical knowledge. This would appear to be necessary when dealing with a dynamic process as opposing a static lexicon which fails to provide a complete descriptive picture of current language use. With the application of automated methods, a wide set of new lexical data and knowledge can be collected and analyzed.

There is the need, therefore, for the implementation of systems which are able to dynamically extend/enhance/update lexicons with information acquired from large corpora and from the web. Our objective should be to set up a new generation of large-size, dynamic lexical resources that fully capture current language usage (how language is materially manifested) and use (the way in which language forms are used as a means of communication) (Widdowson, 1978, pp. 18-19).

ACKNOWLEDGEMENTS

This work is the outcome of a collaborative effort. However, for the specific concerns of Italian academy, Elisabetta Gola is responsible for paragraphs 1-3, Stefano Federici is responsible for 8-9, Nilda Ruimy is responsible for 4-5 and John Wade is responsible for 6-7.

REFERENCES

Antoni-Lay, M.-H., Francopoulo, G., Zaysser, L. (1994). Generic Model for Reuseable Lexicons: The Genelex Project, *Literary and Linguistic Computing*, *9*(1), 47-54.

Bell, R.T. (1991). *Translation and Translating: Theory and Practice.* Harlow: Longman.

Brown, P.F., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P. (1990). A statistical approach to language translation, *Computational Linguistics*, *16*, 79-85.

Brown, R.D. (1999). *Example-based machine translation*, accessed on http://www.cs.cmu.edu/afs/cs.cmu.edu/user/ralf/pub/WWW/ebmt/ebmt.html [2000, January 6].

Collins Cobuild English Dictionary (1995). London: Harper Collins Publisher.

Crystal, D. (1988). *The English Language*, London: Penguin.

Dizionario Inglese-Italiano, Italiano Inglese Ragazzini-Zanichelli. Zanichelli Editore (2009).

Eco, U. (2003). *Dire quasi la stessa cosa: Esperienze di traduzione*. Milano: Bompiani.

Federici, S., Wade J.C. (2007). Letting in the light and working with the Web: A dynamic corpus development approach to interpreting metaphor. In M. Davis, P. Rayson, S. Hunston P. Danielsson, eds. *Proceedings of Corpus Linguistics Conference 2007*, University of Birmingham (UK), http://corpus.bham.ac.uk/corplingproceedings07/paper/207_Paper.pdf.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Hutchins, J. (1996). ALPAC: the (in)famous report, *MT News International* 14, June 1996, 9-12.

Hutchins, J. (1986). *Machine Translation: Past, Present, Future.* Chichester: Ellis Horwood Series in Computers and their Applications.

Hutchins, J. (1993). Latest Developments in Machine Translation technology: Beginning a New Era in MT research. In *Proceedings MT Summit IV.:*

International cooperation for global communication, July 20-22, 1993, Kobe, Japan, 11-34.

Hutchins, J. (2003). Machine translation: general overview. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 501-511.

Isabelle, P., Dymetman, M., Macklovitch, E. (1988). CRITTER: a translation system for agricultural market reports", *Proceedings of the 12th conference on Computational linguistics - Volume 1*, Budapest, 261-266.

Lenci, A., Bel, N., Busa, F., Calzolari N., Gola, E., Monachini, M., Ogonowski, A., Peters I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicon, *International Journal of Lexicography, special issue, Dictionaries, Thesauri and Lexical-Semantic Relations*, *13*(4), 249-263.

Nagao, M. (1984). "A framework of a mechanical translation between Japanese and English by analogy principle", *Artificial and Human Intelligence: edited review papers at the International NATO Symposium on Artificial and Human Intelligence sponsored by the Special Programme Panel held in Lyon*, France, October, 1981, Amsterdam: Elsevier Science Publishers, 173-180.

Pustejovsky, J. (1995). *The Generative Lexicon*, , Cambridge, MA: The MIT Press.

Ruimy, N., Corazzari, O., Gola, E., Spanu, A., Calzolari, N., Zampolli, A. (1998). The European LE-PAROLE Project: The Italian Syntactic Lexicon, LREC (1998) First International Conference on Language Resources and Evaluation Proceedings, I, Granada, Spain, 241-248.

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., Zampolli, A. (2002). CLIPS, a Multi-level Italian Computational Lexicon, *LREC (2002) Third International Conference on language resources and evaluation proceedings*, *III*, Las Palmas de Gran Canaria, 792-799.

Ruimy, N., Monachini, M., Gola, E., Calzolari, N., Del Fiorentino, M.C., Ulivieri, M., Rossi, S. (2003). A computational semantic lexicon of Italian: SIMPLE,. In A. Zampolli, N. Calzolari, L. Cignoni (Eds.), *Computational Linguistics in Pisa. Linguistica Computazionale*, *Special Issue, XVIII-XIX* (II), Pisa-Roma: IEPI, 821-864.

Ruimy, N. (2010). Simple_PLUS: a network of lexical semantic relations Simple_PLUS: una red de relaciones léxico-semánticas. In: *Procesamiento del*

Lenguaje Natural, *44*, Sociedad Española para el Procesamiento del Lenguaje Natural, 99-106.

Sanfilippo, A., *et al.* (1998) *EAGLES Preliminary recommendations on semantic encoding*, The EAGLES Lexicon Interest Group, http://www.ilc.cnr.it/EAGLES/EAGLESLE.PDF.

Sharoff, S. (2006). Open-source corpora: using the net to fish for linguistic data, *The International Journal of Corpus Linguistics, 11*(4), 435-462.

Turcato, D., Mcfetridge, P., Popowich, F., Toole, J. (1999). A unified example-based and lexicalist approach to machine translation. In *Proceedings of the 8th International Conference on theoretical and methodological issues in Machine Translation (TMI '99)*, Chester.

Wade, J.C., Federici, S. (2006). Struttura-significato. Il processo di traduzione. In R. Pititto, S. Venezia (Eds.), *Tradurre e comprendere: pluralità dei linguaggi e delle culture* (Atti del XII Congresso Nazionale della Società di Filosofia del Linguaggio, Piano di Sorrento 2005), Roma: Aracne, 307-332.

Whitelock, P. (1992). Shake-and-bake translation, *Proceedings of the 14th conference on Computational linguistics, 2*, Nantes, 784-791.

Widdowson H.G. (1978) *Teaching language as communication*. Oxford: Oxford University Press.