Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus

Luisa BENTIVOGLI, Pamela FORNER, Emanuele PIANTA
ITC-irst
Via Sommarive, 18
38050 Povo – Trento
Italy
{bentivo, forner, pianta}@itc.it

Abstract

In this paper we illustrate and evaluate an approach to the creation of high quality linguistically annotated resources based on the exploitation of aligned parallel corpora. This approach is based on the assumption that if a text in one language has been annotated and its translation has not, annotations can be transferred from the source text to the target using word alignment as a bridge. The transfer approach has been tested in the creation of the MultiSemCor corpus, an English/Italian parallel corpus created on the basis of the English SemCor corpus. In MultiSemCor texts are aligned at the word level and semantically annotated with a shared inventory of senses. We present some experiments carried out to evaluate the different steps involved in the methodology. The results of the evaluation suggest that the cross-language annotation transfer methodology is a promising solution allowing for the exploitation of existing (mostly English) annotated resources to bootstrap the creation of annotated corpora in new (resource-poor) languages with greatly reduced human effort.

1 Introduction

Large-scale language resources play a crucial role for a steady progress in the field of Natural Language Processing (NLP), as they are essential for carrying out basic research and for building portable and robust systems with broad coverage. More specifically, given the advances of machine learning statistical methods for NLP, with supervised training methods leading the way to major improvements in performance on different tasks, a particularly valuable resource is now represented by large linguistically annotated corpora.

Up until some years ago, linguistically annotated corpora were only produced through manual annotation, or by manual check of automatically produced annotations. Unfortunately, manual annotation is a very difficult and time-consuming task, and this fact has led to a shortage of manual-quality annotated data. The scarcity of large size annotated corpora is more acute for languages different from English, for which even minimal amounts of data are still missing. This state of affairs makes it clear that any endeavour aiming at reducing the human effort needed to produce manual-quality labelled data will be highly beneficial to the field.

Recent studies have shown that a valuable opportunity for breaking the annotated resource bottleneck is represented by parallel corpora, which can be exploited in the creation of resources for new languages via projection of annotations available in another language. This paper represents our contribution to the research in this field. We present a novel methodology to create a semantically annotated corpus by exploiting information contained in an already annotated corpus, using word alignment as a bridge. The methodology has been applied in the creation of the MultiSemCor corpus. MultiSemCor is an English/Italian parallel corpus which is being created on the basis of the English SemCor corpus and where the texts are aligned at the word level and semantically annotated with a shared inventory of senses.

Given the promising results of a pilot study presented in (Bentivogli and Pianta, 2002), the MultiSemCor corpus is now under development. In this paper we focus on a thorough evaluation of the steps involved in the transfer methodology. We evaluate the performance of a new version of the word alignment system and the final quality of the annotations transferred from English to Italian. In Section 2 we lay out the annotation transfer methodology and summarize some related work. In Section 3 we discuss some problematic issues related to the methodology which will be extensively tested and evaluated in Section 4. In Section 5 we report about the state of development of the MultiSemCor corpus and, finally, in Section 6 we present conclusions and our thoughts on future work.

2 The Annotation Transfer Methodology

The MultiSemCor project (Bentivogli and Pianta, 2002) aims at building an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word sense. The parallel

corpus is created by exploiting the SemCor corpus (Landes et al., 1998), which is a subset of the English Brown corpus containing about 700,000 running words. In SemCor all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged with reference to the WordNet lexical database[1] (Fellbaum, 1998).

The main hypothesis underlying this methodology is that, given a text and its translation into another language, the semantic information is mostly preserved during the translation process. Therefore, if the texts in one language have been semantically annotated and their translations have not, annotations can be transferred from the source language to the target using word alignment as a bridge.

The first problem to be solved in the creation of MultiSemCor was the fact that the Italian translations of the SemCor texts did not exist. Our solution was to have the translations made by professional translators. Given the high costs of building semantically annotated corpora, requiring specific skills and very specialized training, we think that manually translating the annotated corpus and automatically transferring the annotations may be preferable to hand-labelling a corpus from scratch. Not only are translators more easily available than linguistic annotators, but translations may be a more flexible and durable kind of annotation. Moreover, the annotation transfer methodology has the further advantage of producing a parallel corpus.

With respect to a situation in which the translation of a corpus is already available, a corpus translated on purpose presents the advantage that translations can be *"controlled"*, i.e. carried out following criteria aiming at maximizing alignment and annotation transfer. Our professional translators are asked to use, preferably, the same dictionaries used by the word aligner, and to maximize, whenever possible, the lexical correspondences between source and target texts. The translators are also told that the controlled translation criteria should never be followed to the detriment of a good Italian prose. Controlled translations cost the same as free translations, while having the advantage of enhancing the performances of the annotation transfer procedure.

Once the SemCor texts have been translated, the strategy for creating MultiSemCor consists of (i) automatically aligning Italian and English texts at the word level, and (ii) automatically transferring the word sense annotations from English to the aligned Italian words. The final result of the MultiSemCor project is an Italian corpus annotated with PoS, lemma and word sense, but also an aligned parallel corpus lexically annotated with a shared inventory of word senses. More specifically, the sense inventory used is MultiWordNet (Pianta et al., 2002), a multilingual lexical database in which the Italian component is strictly aligned with the English WordNet.

2.1 Related Work

The idea of obtaining linguistic information about a text in one language by exploiting parallel or comparable texts in another language has been explored in the field of Word Sense Disambiguation (WSD) since the early 90's, the most representative works being (Brown et al., 1991), (Gale et al., 1992), and (Dagan and Itai, 1994).

In more recent years, Ide et al. (2002) present a method to identify word meanings starting from a multilingual corpus. A by-product of applying this method is that once a word in one language is word-sense tagged, the translation equivalents in the parallel texts are also automatically annotated.

Cross-language tagging is the goal of the work by Diab and Resnik (2002), who present a method for word sense tagging both the source and target texts of parallel bilingual corpora with the WordNet sense inventory.

Parallel to the studies regarding the projection of semantic information, more recently the NLP community has also explored the possibility of exploiting translation to project more syntax-oriented annotations. Yarowsky et al. (2001) describe a successful method consisting of (i) automatic annotation of English texts, (ii) cross-language projection of annotations onto target language texts, and (iii) induction of noise-robust taggers for the target language. A further step is made in (Hwa et al., 2002) and (Cabezas et al., 2001), which address the task of acquiring a dependency treebank by bootstrapping from existing linguistic resources for English. Finally, in (Riloff et al., 2002) a method is presented for rapidly creating Information Extraction (IE) systems for new languages by exploiting existing IE systems via cross-language projection.

The results of all the above mentioned studies show how previous major investments in English

[1] WordNet is an English lexical database, developed at Princeton University, in which nouns, verbs, adjectives, and adverbs are organized into sets of synonyms (synsets) and linked to each other by means of various lexical and semantic relationships. In the last years, within the NLP community WordNet has become the reference lexicon for almost all tasks involving word sense disambiguation (see, for instance, the Senseval competition).

annotated corpora and tool development can be effectively leveraged across languages, allowing the development of accurate resources and tools in other languages without comparable human effort.

3 Quality Issues

The MultiSemCor project raises a number of theoretical and practical issues. For instance: is translational language fully representative of the general use of language in the same way as original language is? To what extent are the lexica of different languages comparable? These theoretical issues have already been presented in (Pianta and Bentivogli, 2003) and will not be discussed here. In the following, we address the issue of the quality of the annotation resulting from the application of the methodology.

As opposed to automatic word sense disambiguation tasks, the MultiSemCor project specifically aims at producing manual-quality annotated data. Therefore, a potential risk which needs to be faced is represented by the possible degradation of the Italian annotation quality through the various steps of the annotation transfer procedure. A number of factors must be taken into account. First, annotation errors can be found in the original English texts. Then, the word aligner may align words incorrectly, and finally the transfer of the semantic annotations may not be applicable to certain translation pairs.

SemCor quality. The English SemCor corpus has been manually annotated. However, some annotation errors can be found in the texts (see Fellbaum et al., 1998, for SemCor taggers' confidence ratings). As an example, the word *pocket* in the sentence "He put his hands on his pockets" was incorrectly tagged with the WordNet synset {pouch, sac, sack, pocket -- an enclosed space} instead of the correct one {pocket -- a small pouch in a garment for carrying small articles}.

Word alignment quality. The feasibility of the entire MultiSemCor project heavily depends on the availability of an English/Italian word aligner with very good performance in terms of recall and, more importantly, precision.

Transfer quality. Even when both the original English annotations and the word alignment are correct, a number of cases still remain for which the transfer of the annotation is not applicable. An annotation is not transferable from the source language to the target when the translation equivalent does not preserve the lexical meaning of the source language. In these cases, if the alignment process puts the two expressions in correspondence, then the transfer of the sense annotation from the source to the target language is not correct.

The first main cause of incorrect transfer is represented by translation equivalents which are not cross-language synonyms of the source language words. For example, in a sentence of the corpus the English word *meaning* is translated with the Italian word *motivo* (reason, grounds) which is suitable in that specific context but is not a synonymic translation of the English word. In this case, if the two words are aligned, the transfer of the sense annotation from English is not correct as the English sense annotation is not suitable for the Italian word. A specific case of non-synonymous translation occurs when a translation equivalent does not belong to the same lexical category of the source word. For example, the English verb *to coexist* in the sentence "the possibility for man to coexist with animals" has been translated with the Italian noun *coesistenza* (coexistence) in "le possibilità di coesistenza tra gli uomini e gli animali". Even if the translation is suitable for that context, the English sense of the verb cannot be transferred to the Italian noun. Sometimes, non-synonymous translations are due to errors in the Italian translation, as in *pull* translated as *spingere* (push).

A second case which offers challenge to the sense annotation transfer is phrasal correspondence, occurring when a target phrase has globally the same meaning as the corresponding source phrase, but the single words of the phrase are not cross-language synonyms of their corresponding source words. For example, the expression *a dreamer sees* has been translated as *una persona sogna* (a person dreams). The Italian translation maintains the synonymy at the phrase level but the single component words do not. Therefore, if the single words were aligned any transfer from English to Italian would be incorrect. Another example of phrasal correspondence, in which the semantic equivalence between words in the source and target phrase is even fuzzier, is given by the English phrase *the days would get shorter and shorter* translated as *imminente fine dei tempi* (imminent end of times).

Another controversial cause of possible incorrect transfer is represented by the case in which the translation equivalent is indeed a cross-language synonym of the source expression but it is not a lexical unit. This usually happens with lexical gaps, i.e. when a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words, as for instance the English word *successfully* which can only be translated with the Italian free combination of words *con successo* (with success). However, it can also be the result of a choice made by the translator who

decides to use a free combination of words instead of a possible lexical unit, as in *empirically* translated as *in modo empirico* (in an empirical manner) instead of *empiricamente*. In these cases the problem arises because in principle if the target expression is not a lexical unit it cannot be annotated as a whole. On the contrary, each component of the free combination of words should be annotated with its respective sense.

In the next Section we will address these quality issues in order to assess the extent to which they affect the cross-language annotation transfer methodology.

4 Evaluation of the Annotation Transfer Methodology

A number of experiments have been carried out in order to test the various steps involved in the annotation transfer methodology. More precisely, we evaluated the performances of the word alignment system and the quality of the final annotation of the Italian corpus.

4.1 Word Alignment

Word alignment is the first crucial step in the methodology applied to build MultiSemCor. The word aligner used in the project is KNOWA (KNOwledge-intensive Word Aligner), an English/Italian word aligner, developed at ITC-irst, which relies mostly on information contained in the Collins bilingual dictionary, available in electronic format. KNOWA also exploits a morphological analyzer and a multiword recognizer for both English and Italian. For a detailed discussion of the characteristics of this tool, see (Pianta and Bentivogli, 2004).

Some characteristics of the MultiSemCor scenario make the alignment task easier for KNOWA. First, in SemCor all multiwords included in WordNet are explicitly marked. Thus KNOWA does not need to recognize English multiwords, although it still needs to recognize the Italian ones. Second, within MultiSemCor word alignment is done with the final aim of transferring lexical annotations from English to Italian. Since only content words have word sense annotations in SemCor, it is more important that KNOWA behaves correctly on content words, which are easier to align than functional words.

To evaluate the word aligner performance on the MultiSemCor task we created a gold standard composed of three English unseen texts (`br-f43`, `br-l10`, `br-j53`) taken randomly from the SemCor corpus. For each English text both a *controlled* and a *free* translation were made. Given the expectation that free translations are less suitable for word alignment, we decided to test KNOWA also on them in order to verify if the annotation transfer methodology can be applied to already existing parallel corpora.

The six resulting pairs of texts were manually aligned following a set of alignment guidelines which have been defined taking into account the work done in similar word alignment projects (Melamed, 2001). Annotators were asked to align different kinds of units (simple words, segments of more than one word, parts of words) and to mark different kinds of semantic correspondence between the aligned units, e.g. full correspondence (synonymic), non synonymic, changes in lexical category, phrasal correspondence. Inter-annotator agreement was measured with the Dice coefficient proposed in (Véronis and Langlais, 2000) and can be considered satisfactory as it turned out to be 87% for free translations and 92% for controlled translations. As expected, controlled translations produced a better agreement rate between annotators.

For assessing the performance of KNOWA, the standard notions of Precision, Recall, and Coverage have been used following (Véronis and Langlais, 2000). See (Och and Ney, 2003) and Arenberg et al., 2000) for different evaluation metrics. The performance of KNOWA applied to the MultiSemCor gold standard in a full-text alignment task is shown in Table 1. These results, which compare well with those reported in the literature (Véronis, 2000) show that, as expected, controlled translations allow for a better alignment but also that free translations may be satisfactorily aligned.

The evaluation of KNOWA with respect to the English content words which have a semantic tag in SemCor is reported in Tables 2 and 3, for both free and controlled translations and broken down by Part of Speech.

	Precision	Recall	Coverage
Free	83.5	57.9	60.0
Controlled	88.4	67.5	74.9

Table 1: KNOWA on Full-text

	Precision	Recall	Coverage
Nouns	93.7	81.1	86.5
Verbs	85.6	70.3	82.1
Adjectives	95.6	64.7	67.7
Adverbs	88.4	38.5	43.5
Total	**91.2**	**68.2**	**74.8**

Table 2: KNOWA on sense-tagged words only (Free translations)

	Precision	Recall	Coverage
Nouns	95.9	82.5	86.1
Verbs	90.7	76.8	84.7
Adjectives	95.2	69.9	73.5
Adverbs	90.4	51.6	57.1
Total	**93.9**	**74.6**	**79.5**

Table 3: KNOWA on sense-tagged words only (Controlled translations)

We can see that ignoring function words the performance of the word aligner improves in both precision and recall.

4.2 Italian Annotation Quality

As pointed out in Section 3, even in the case of a perfect word alignment the transfer of the annotations from English to the correctly aligned Italian words can still be a source of errors in the resulting Italian annotations. In order to evaluate the quality of the annotations automatically transferred to Italian, a new gold standard was created starting from SemCor text `br-g11`. The English text, containing 2,153 tokens and 1,054 semantic annotations, was translated into Italian in a *controlled* modality. The resulting Italian text is composed of 2,351 tokens, among which 1,085 are content words to be annotated. The English text and its Italian translation were manually aligned and the Italian text was manually semantically annotated taking into account the annotations of the English words. Each time an English annotation was appropriate for the Italian corresponding word, the annotator used it also for Italian. Otherwise, the annotator did not use the original English annotation for the Italian word and looked in WordNet for a suitable annotation.

Moreover, when the English annotations were not suitable for annotating the Italian words, the annotator explicitly distinguished between wrong English annotations and English annotations that could not be transferred to the Italian translation equivalents. The errors in the English annotations amount to 24 cases. Non-transferable annotations amount to 155, among which 143 are due to lack of synonymy at lexical level and 12 to translation equivalents which are not lexical units.

The differences between the English and Italian text with respect to the number of tokens and annotations have also been analysed. The Italian text has about 200 tokens and 31 annotated words more than the English text. The difference in the number of tokens is due to various factors. First, there are grammatical characteristics specific to the Italian language, such as a different usage of articles, or a greater usage of reflexive verbs which leads to a higher number of clitics. For example, the English sentence “as cells coalesced” must be translated into Italian as “quando *le* cellule *si* unirono”. Then, we have single English words translated into Italian with free combinations of words (ex: *down* translated as *verso il basso*) and multiwords which are recognized in English and not recognized in Italian (e.g. one token for *nucleic_acid* in the English text and two tokens in the Italian text, one for *acido* and one for *nucleico*). As regards content words to be annotated, we would have expected that their number was the same both in English and Italian. In fact, the difference we found is much lower than the difference between tokens. This difference is explained by the fact that some English content words have not been annotated. For example, modal and auxiliary verbs (to have, to be, can, may, to have to, etc.) and partitives (some, any) where systematically left unannotated in the English text whereas they have been annotated for Italian.

The automatic procedures for word alignment and annotation transfer were run on text `br-g11` and evaluated against the gold standard. The total number of transferred senses amounts to 879. Among them, 756 are correct and 123 are incorrect for the Italian words. Table 4 summarizes the results in terms of precision, recall and coverage with respect to both English annotations available (1,054) and Italian words to be annotated (1,085).

We can see that the final quality of the Italian annotations is acceptable, the precision amounting to 86.0%. The annotation error rate of 14.0% has been analyzed in order to classify the different factors affecting the transfer methodology. Table 5 reports the data about the composition of the incorrect transfer.

Comparing the number of annotation errors in the English source, as marked up during the creation of the gold standard (24), with the number of errors in the Italian annotation due to errors in the original annotation (22), we can see that almost all of the source errors have been transferred, contributing in a consistent way to the overall Italian annotation error rate.

As regards word alignment, `br-g11` was a relatively easy text as the performance of KNOWA (i.e. 96.5%) is higher than that obtained with the test set (see Table 3).

	Precision	Recall	Coverage
Wrt English	86.0	71.7	83.4
Wrt Italian	**86.0**	**69.7**	**81.0**

Table 4: Annotation evaluation on text br-g11

	#	%
English annotation errors	22	2.5
Word alignment errors	31	3.5
Non-transferable annotations	70	8.0
Total incorrect transfers	123	14.0

Table 5: Composition of the incorrect transfer

The last source of annotation errors is represented by words which have been correctly aligned but whose word sense annotation cannot be transferred. This happens with (i) translation equivalents which are lexical units but are not cross-language synonyms, and (ii) translation equivalents which are cross-language synonyms but are not lexical units. In practice, given the difficulty in deciding what is a lexical unit and what is not, we decided to accept the transfer of a word sense from an English lexical unit to an Italian free combination of words (see for instance *occhiali da sole* annotated with the sense of *sunglasses*). Therefore, only the lack of synonymy at lexical level has been considered an annotation error.

The obtained results are encouraging. Among the 143 non-synonymous translations marked in the gold standard, only 70 have been aligned by the word alignment system, showing that KNOWA is well suited to the MultiSemCor task. The reason is that it relies on bilingual dictionaries where non-synonymous translations are quite rare. This can be an advantage with respect to statistics-based word aligners, which are expected to be able to align a great number of non-synonymous translations, thus introducing more errors in the transfer procedure.

A final remark about the evaluation concerns the proportion of non-transferable word senses with respect to errors in the original English annotations. It is sometimes very difficult to distinguish between annotation errors and non-transferable word senses, also because we are not English native speakers. Thus, we preferred to be conservative in marking English annotations as errors unless in very clear cases. This approach may have reduced the number of the errors in the original English corpus and augmented the number of non-transferable word senses, thus penalizing the transfer procedure itself.

Summing up, the cross-language annotation transfer methodology produces an Italian corpus which is tagged with a final precision of 86.0%. After the application of the methodology 19.0% of the Italian words still need to be annotated (see the annotation coverage of 81.0%). We think that, given the precision and coverage rates obtained from the evaluation, the corpus as it results from the automatic procedure can be profitably used. However, even in the case that a manual revision is envisaged, we think that hand-checking the automatically tagged corpus and manually annotating the remaining 19% still results to be cost effective with respect to annotating the corpus from scratch.

5 The MultiSemCor Corpus Up to Now

We are currently working at the extensive application of the annotation transfer methodology for the creation of the MultiSemCor corpus. Up to now, MultiSemCor is composed of 29 English texts aligned at the word level with their corresponding Italian translations. Both source and target texts are annotated with POS, lemma, and word sense. More specifically, as regards English we have 55,935 running words among which 29,655 words are semantically annotated (from SemCor). As for Italian, the corpus amounts to 59,726 running words among which 23,095 words are annotated with word senses that have been automatically transferred from English.

MultiSemCor can be a useful resource for a variety of tasks, both as a monolingual semantically annotated corpus and as a parallel aligned corpus. As an example, we are already using it to automatically enrich the Italian component of MultiWordNet, the reference lexicon of MultiSemCor. As a matter of fact, out of the 23,095 Italian words automatically sense-tagged, 5,292 are not yet present in MultiWordNet and will be added to it. Moreover, the Italian component of MultiSemCor is being used as a gold standard for the evaluation of Word Sense Disambiguation systems working on Italian. Besides NLP applications, MultiSemCor is also suitable to be consulted by humans through a Web interface (Ranieri et al., 2004) which is available at: http://tcc.itc.it/projects/multisemcor.

6 Conclusion and future directions

We have presented and evaluated an approach to the creation of high quality semantically annotated resources based on the exploitation of aligned parallel corpora. The results obtained from the thorough evaluation of the different steps involved in the methodology confirm the feasibility of the MultiSemCor project. The cross-lingual annotation transfer methodology is going to be applied also to the remaining 157 SemCor texts, which are currently being translated into Italian.

As regards future research directions within the transfer annotation paradigm, it would be interesting to extend the methodology to other languages, e.g. Spanish, for which a WordNet

exists and can be aligned with MultiWordNet. Moreover, as the Brown Corpus, used to create SemCor, has been syntactically annotated within the English Penn Treebank, the syntactic annotations of the SemCor texts are also available. We are planning to explore the possibility of transferring the syntactic annotations from the English to the Italian texts of MultiSemCor.

References

L. Ahrenberg, M. Merkel, H. Sagvall and A. J. Tiedemann. 2000. Evaluation of word alignment systems. In *Proceedings of LREC 2000*, Athens, Greece.

L. Bentivogli and E. Pianta. 2002. Opportunistic Semantic Tagging. In *Proceedings of LREC-2002*, Las Palmas, Canary Islands, Spain.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1991. Word-Sense Disambiguation using Statistical Methods. In *Proceedings of ACL'91*, Berkeley, California, USA.

C. Cabezas, B. Dorr and P. Resnik. 2001. Spanish Language Processing at University of Maryland: Building Infrastructure for Multilingual Applications. In *Proceedings of the 2nd International Workshop on Spanish Language Processing and Language Technologies*, Jaen, Spain.

I. Dagan and A. Itai. 1994. Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*: 20(4):563-596.

M. Diab and P. Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *In Proceedings of ACL 2002*, Philadelphia, USA .

C. Fellbaum, J. Grabowski and S. Landes. 1998. Performance and confidence in a semantic annotation task. In Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge (Mass.).

C. Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge (Mass.).

W. A. Gale, K. W. Church and D. Yarowsky. 1992. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montreal, Canada.

R. Hwa, P. Resnik and A. Weinberg. 2002. Breaking the Resource Bottleneck for Multilingual Parsing. In *Proceedings of the LREC-2002 Workshop on "Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data"*, Las Palmas, Canary Islands, Spain.

N. Ide, T. Erjavec, and D. Tufis. 2002. Sense Discrimination with Parallel Corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, USA.

S. Landes C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge (Mass.).

I. D. Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts*. The MIT Press, Cambridge (Mass.).

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-53.

E. Pianta and L. Bentivogli. 2004. Knowledge intensive word alignment with KNOWA. In *Proceedings of Coling 2004*, Geneva, Switzerland.

E. Pianta and L. Bentivogli. 2003. Translation as Annotation. In *Proceedings of the AI*IA 2003 Workshop "Topics and Perspectives of Natural Language Processing in Italy"*, Pisa, Italy.

E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First Global WordNet Conference*, Mysore, India.

M. Ranieri, E. Pianta and L. Bentivogli. 2004. Browsing Multilingual Information with the MultiSemCor Web Interface. In *Proceedings of the LREC-2004 Workshop "The amazing utility of parallel and comparable corpora"*, Lisbon, Portugal.

E. Riloff, C. Schafer and D. Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of Coling 2002*, Taipei, Taiwan.

J. Véronis and P. Langlais. 2000. Evaluation of parallel text alignment systems. In Véronis, J. (ed.). 2000. *Parallel Text Processing*, Kluwer Academic Publishers, Dordrecht.

J. Véronis (ed.). 2000. *Parallel Text Processing*. Kluwer Academic Publishers, Dordrecht.

D. Yarowsky, G. Ngai and R. Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001*, San Diego, California, USA.