Three issues in cross-language frame information transfer

Sara Tonelli, Emanuele Pianta
FBK-Irst
Via Sommarive 18
I-38100 Povo (Trento), Italy
{*satonelli,pianta*}@*fbk.eu*

Abstract

In this paper we address the task of transferring FrameNet annotations from an English corpus to an aligned Italian corpus. Experiments were carried out on an English-Italian bitext extracted from the Europarl corpus and on a set of selected sentences from the English FrameNet corpus that have been manually translated into Italian. Our research activity is aimed at answering the following three questions: (1) What is the best annotation transfer algorithm for the English-Italian couple? (2) What kind of parallel corpus is best suitable to the annotation transfer task? (3) How should the annotation transfer be evaluated, given the final aim of the transfer?

Keywords

Frame semantics, cross-language annotation transfer, automatic development of lexical resources.

1 Introduction

In recent years, the creation of annotated lexical resources has become crucial to the development of text processing systems, especially to train supervised learning systems and evaluate unsupervised or handcrafted systems. The FrameNet database [8] clearly exemplifies this trend. This resource contains more than 135,000 annotated sentences pointing to more than 10,000 lexical units, with a rich repository of semantic roles (the *frame elements*) and almost 900 situation descriptions (the *frames*). FrameNet has proved to be useful in a number of NLP tasks, from textual entailment [3] to question answering [16], and the development of systems for frame recognition has become a topic of great interest for the NLP community, with a devoted task at the last SemEval workshop[1].

Given the success of the English FrameNet initiative, many researchers have focused on the development of FrameNet-like resources for other languages through manual annotation, for example [4] for German and [17] for Spanish. Manual annotation guarantees high accuracy but requires trained annotators and is expensive and time-consuming. For this reason, a second approach has been investigated, which is based on the automatic projection of frame information from English texts into a new language using bilingual parallel corpora and possibly carrying out automatic annotation of frame information on the English side. If no parallel corpora are available, manually translating an annotated English corpus and automatically transferring the annotations may represent a reliable alternative to hand-labeling a new corpus from scratch, given that translators are more easily available than linguistic annotators, particularly for complex tasks like frame annotation. In this paper we explore the possibility to develop a FrameNet database for Italian using transfer methodologies, and discuss the advantages of using a manually translated parallel corpus for the transfer task. Besides, we present and discuss two existing evaluation frameworks and propose a new evaluation approach. More specifically we try to answer the 3 following questions: (1) What is the best annotation transfer algorithm for the English-Italian couple? (2) What kind of parallel corpus is best suitable to the annotation transfer task? (3) How should the annotation transfer be evaluated, given the final aim of the transfer? We try to answer these questions in Section 3, 4, and 5.

[1] http://framenet.icsi.berkeley.edu/semeval/FSSE.html

2 The FrameNet projects

FrameNet [8] is a lexical resource for English based on corpus evidence, whose conceptual model comprises a set of prototypical situations called *frames*, the frame-evoking words or expressions called *lexical units* or *targets* and the roles or participants involved in these situations, called *frame elements*. They can be either *core*, i.e. typical of a given frame, and *non-core*, with more general meaning and several instantiations in different frames. All lexical units belonging to the same frame have similar semantics that is expressed by a set of *valence patterns*. i.e. patterns of grammatical realizations of the frame elements. We report in the table below an example frame from the FrameNet database. The WEARING frame is described with a definition, the list of frame-evoking lexical units and the core frame elements with an example sentence each:

A particular feature of the FrameNet resource is that it comprises a language-independent layer with the description of frame and frame elements (Table 1, *Def* row), and two language-dependent parts, namely the lexical unit set for every frame and the corresponding example sentences (*LUs* and *FEs* rows). For

International Conference RANLP 2009 - Borovets, Bulgaria, pages 441–448

Frame: WEARING		
Def.	The words in this frame refer to what CLOTHING a WEARER (or a specific BODY_PART of the WEARER) has on	
LUs	attired.a, bare-armed.a, bare-breasted.a, bare.v, braless.a clothed.a, coatless.a, costumed.a, decked out.a, dressed.a have got on.v, sport.v, swaddled.a, swathed.a, wear.v [...]	
FEs	BODY_PART	She was wearing a glove *on one hand*.
	CLOTHING	Lucy had *dark glasses* on.
	WEARER	She reached a group of costumed *dancers*.

Table 1: *Frame* WEARING

this reason, the FrameNet model is particularly suitable to cross-lingual induction and can be applied to languages other than English, keeping the theoretical framework as it is and populating the frames with language-specific lexical units and corpus instances. In some cases, new frame definitions may be required for the new language.

The first step towards the creation of a FrameNet database for a new language should be the annotation of frame information on a corpus of sentences in the new language. Since manual annotation is time-consuming and requires relevant financial efforts, several approaches have been proposed in the past to automatically carry out the annotation process. The most convenient alternative to manual annotation seems to be the import of English FrameNet annotation into another language exploiting a parallel corpus. [13] proposed a method to transfer frame annotation from English to German starting from parallel texts with the English side annotated with frame information. They proposed a model based on alignment at constituent level obtained through word overlap similarity. [14] tested a similar approach on a parallel English-French corpus, showing that the transfer framework can get promising results also if applied to Romance languages. [9] applied the transfer method to English-Swedish parallel texts with the English side being automatically annotated with a semantic role labeller trained on the English FrameNet database.

As for Italian, a few projects are currently aimed at developing FrameNet for Italian and at exploring new approaches to speed up manual annotation or convey fully automatic annotation. [1] have proposed a methodology to automatically transfer frame information on an English-Italian parallel corpus based on a statistical machine translation step augmented with a rule-based post-processing. [5] have trained and tested a system for automatic frame element detection using a corpus of Italian dialogs manually annotated with frame information.

3 Transfer algorithm selection

The task of frame annotation transfer is two-folded as it implies transferring the annotation of the target, which is always a lexical unit, and of frame elements, which are more complex syntactic constituents (up to full clause). Also, the annotation can be carried out at the level of strings of words or at the level of syntactic constituents, as in the work of [13] and [14]. As a consequence, the transfer algorithm can be based only on word alignment or also, when available, on syntactic structure information. [13] carried out experiments with both approaches and proved that exploiting constituent information yields substantial improvements over relying on word alignment alone. The methodology was then further optimized by [12] and applied to English-German and English-French corpora in order to transfer FE information via constituent alignment. We explored two variants of the constituent-based strategy applied to frame information transfer from English to Italian. The first variant, which was presented in [18], requires full parsing on both source and target corpus. Given an English constituent, annotated as FE, the algorithm extracts its head, aligns it with the corresponding Italian head, then looks for the maximal syntactic projection of the Italian semantic head, and transfers the English FE annotation to such constituent. In this approach, the correct alignment of the head is enough to carry out the FE transfer. However, this feature may also turn in a disadvantage, because if the semantic head is not aligned, there will be no transfer.

We present here a second version of the transfer algorithm which is more similar to [12] in that the alignment between constituents is not based on the semantic head but on the best percentage of aligned words. However, unlike [12] who considers all possible constituents in the parse tree, we take into account only constituents that are syntactically connected to the target in the Italian sentence. Note that in this approach no parsing information on the English side is required. The algorithm description is reported below:

Given two aligned sentences s_{en} and s_{it}
take $lexunit_{en} \in s_{en}$
if exists $alignment_{lexunit_{en}}$
    take aligned $lexunit_{it} \in s_{it}$
    transfer $info_{frame}$ *from* $lexunit_{en}$ *to* $lexunit_{it}$
    return $lexunit_{it+infoframe}$
    extract D_{it} *from* s_{it}
// D_{it} = set of syntactic dependents of $lexunit_{it}$ in s_{it}
    for each $fe_{en} \in FE_{en}$
        $Score_{best} = 0$
        $Cand_{best} = empty$
        for each $d_{it} \in D_{it}$
            calculate $Score_{it}$
// $Score_{it}$ = n. of aligned words between fe_{en} and d_{it}
            if $Score_{it} > Score_{best}$
              $Score_{best} = Score_{it}$
              $Cand_{best} = d_{it}$
            end if
        end for
        return $Score_{best}$
        return $Cand_{best}$
    end for
else
    return false

We take the English corpus annotated with frame information C_{en} and align it at word level to the Italian corpus C_{it}, whose sentences have been previously parsed. For each sentence $s_{en} \in C_{en}$, we take the annotated lexical unit $lexunit_{en}$ and find the Italian aligned word, that we assume to be the target lexical unit $lexunit_{it}$. If no alignment is available, the transfer fails, otherwise the English frame label is assigned

to the Italian $lexunit_{it}$. Then, for every English frame element fe_{en}, we take all syntactic dependents D_{it} of $lexunit_{it}$ and compute the number of aligned words between fe_{en} and $d_{it} \in D_{it}$. We consider the Italian dependent with most aligned words $Cand_{best}$ as the best candidate for annotation projection.

As an example, we report in Figure 1 the output of the first transfer algorithm applied to two parallel sentences from the *Europarl* corpus [10]. Dotted arrows connect aligned tokens.

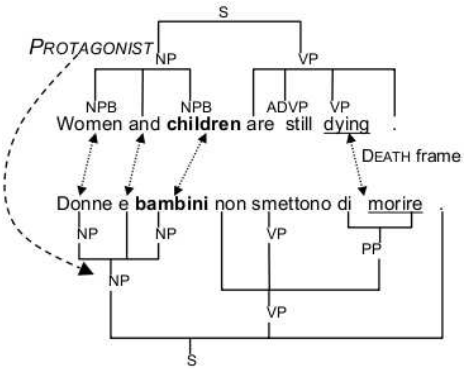

Fig. 1: *Correct transfer with Algorithm 1*

Since *"morire"* is correctly aligned with the target *"dying"*, it becomes the Italian lexical unit of the DEATH frame. As for the PROTAGONIST frame element, first *"children"* is identified as the semantic head of the constituent, then it is connected to *"bambini"*, and finally the NP node dominating *"Donne e bambini"* is selected as the best Italian constituent because it represents the highest syntactic projection of the Italian head compatible with the annotated English constituent. Algorithm 2 would not deliver any FE transfer on the same couple of sentences, as it cannot identify *"Donne e bambini"* as dependent of *"morire"*, due to the different syntactic structure of the Italian sentence. In Figure 2 we report the output of the second transfer algorithm applied to two parallel sentences from the *Europarl* corpus. Note that, unlike [12], we do not exploit any syntactic information on the English side and that the FE labels point to flat chunks, whereas in Figure 1 the sentences have been parsed on both sides.

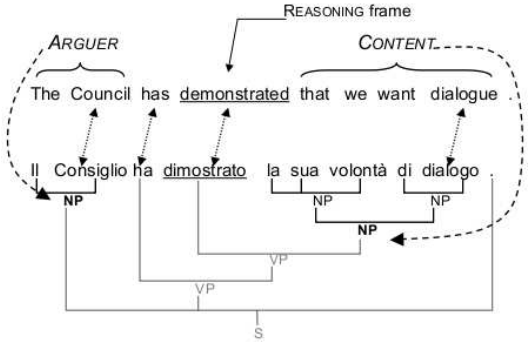

Fig. 2: *Correct transfer with Algorithm 2*

In this example, *"demonstrated"* is the target of the REASONING frame, and two frame elements are present, namely CONTENT and ARGUER. Both frame elements point to the correct constituent nodes in Italian, that are the syntactic dependents of the target *"dimostrato"*. The CONTENT frame element is correctly transferred even if only one word (*dialogue* - *dialogo*), which is not the semantic head of the constituent, has been aligned. This algorithm can cope with a different syntactic structure of the sentence in Italian, where the English secondary clause *"that we want dialogue"* is translated as *"la sua volontà di dialogo"* (i.e. *its will to dialogue*). With algorithm 1 the transfer of the CONTENT label would have failed because the semantic head of the constituent, *"want"*, has no alignment in Italian.

4 *Europarl* and *MultiBerkeley*

In order to investigate the influence of the corpus characteristics on the transfer quality, we took into account two different parallel corpora.

The first corpus was an excerpt of 987 English and Italian sentences taken from the *Europarl* multilanguage parallel corpus [10]. The English side of the corpus has been automatically annotated with part of speech and syntactic information and manually enriched with frame-semantic information as described in [13] in the context of transfer experiments between English and German. The same sentences were used also for the English-Italian transfer. The Italian sentences were parsed with Bikel's phrase-based statistical parser trained for Italian [6], which obtained the best score in the EVALITA evaluation campaign for Italian NLP tools with 70.79 f-measure. Then the English-Italian corpus was aligned at word level with KNOWA (KNowledge-intensive Word Aligner) [15]. The *coverage* of the *word alignment* process reached 65.1 coverage on the whole corpus. The Italian side of the corpus was manually annotated with frame information in order to build a gold standard to assess transfer quality. The gold standard turns out to include instances of 158 frames, mainly connected to the communication and the political scenarios, with the great majority of lexical units being verbs. This means that the variability of frames was limited, with about 6 instances for every frame. Another characteristic of the *Europarl* corpus is the presence of extremely free translations. This is due not only to the translation style, but also to the corpus structure. In fact, if we consider a set of parallel sentences from this corpus, it may include translations of the same sentence from a third language. For instance, a pair of English-Italian parallel sentences may have been translated from a French source sentence. This makes the corpus less suitable for the task of transfer annotation.

For this reason, we take into account also a second corpus called *MultiBerkeley*, which is built by manually translating in a controlled way a number of sentences from the Berkeley FrameNet corpus. The selection of sentences was guided by the desire to include in the resulting Italian corpus frames that were not already present in *Europarl*. Also, we wanted to acquire targets that are not verbs. Besides, as past experi-

ments on annotation transfer have shown (see [2]), the automatic projection of annotation between two parallel corpora in different languages can benefit from a translation that minimizes syntactic differences from source and target language. For this reason (i) we selected 400 frames that are not represented in the *Europarl* gold standard and (ii) for each of them, we chose the target with the largest set of example sentences in the English FrameNet database. Even if the information in the FrameNet database is not statistically significant w.r.t. the frequency of the occurrence of the different targets, we assumed that a target with several attestations in the Berkeley corpus and a complete annotation should be considered significant of the frame it belongs to. Among the extracted sentences for every target, (iii) we selected the shortest one, discarding the instances where all frame elements are expressed by a personal pronoun (e.g. *"He took it"*). In the end, we obtained an English corpus composed of 400 sentences with one example per frame. The sentences are taken from the English FNet database, thus they are PoS tagged and annotated with frame information. All frame elements are also labeled with phrase type (NP, PP, VP, etc.) and grammatical function (Ext, Dep, Head, etc.). We manually translated the English corpus into Italian trying to limit "free" translations in order to enhance the correspondence between source and target texts. If possible, we preferred Italian translations minimizing divergences with English. However, priority was always given to good Italian prose. Once we created the Italian version of the corpus, the rest of the pre-processing step remained the same as for the *Europarl* corpus, with the Italian sentences being parsed with Bikel's parser and the bitext aligned at word level with KNOWA. Finally, we manually annotated all Italian sentences with frame information in order to create a second gold standard for evaluation. We call the resulting corpus *MultiBerkeley*.

We report in Table 2 some statistics about the two corpora. Note that FE parallelism is computed over the subset of sentences with frame parallelism.

	Europarl	*MBerk.*
Avg. sent. length (tokens)	23±9	10±4
Frame parallelism	0.61	0.98
FE parallelism	0.82	0.91

Table 2: *Corpus comparison*

The average sentence length in the *Europarl* corpus is more than double than that in the *MultiBerkeley* corpus due to the different selection strategy of the sentences. Different values of frame and FE parallelism depend partly on the fact that the English side had been annotated with FrameNet v. 1.1 for previous experiments [13], while we used version 1.3 for the Italian gold standard. Nonetheless, the main reason for lacking parallelism are free translations, that are particularly frequent in the *Europarl* corpus. If we apply the framework proposed by [7] to the parallel sentences with diverging frame annotation, we notice that they are mainly caused by a subset of *translation shifts* called *semantic shifts* , showing a variation of the meaning in the source and the target sentence. On the contrary, *grammatical shifts* (f.e. change of category) tend to preserve the frame label in the translated sentence. As an example, we report two pairs of sentences from *Europarl*. In (1), the change of category between *pay.v* and *pagamento.n (payment.n)* does not affect the assigned frame COMMERCE_PAY. In (2), instead, the target word *say* was translated as *sottolineare (underline)*, that led to a frame change from STATEMENT to CONVEY_IMPORTANCE[2]:

(1) I do not believe that we can solve the problem by **paying** fees. [COMMERCE_PAY]
Non credo che la soluzione consista nel **pagamento** di nuove spese. [COMMERCE_PAY]

(2) Let me **say** it again quite clearly, we have not brought up the question. [STATEMENT]
Desidero ancora una volta **sottolineare** che non abbiamo affrontato la questione. [CONVEY_IMPORTANCE]

We expect the different semantic parallelism and the different complexity of the two corpora to impact on the transfer performance.

5 Evaluation framework

In different research works about frame annotation transfer, several evaluation criteria have been applied. The common feature among them is the choice to include in the testset only sentences that present a certain degree of semantic parallelism in the parallel gold standards. We believe that this approach is not suitable for our goal. Since we aim at producing an annotated corpus with near manual annotation quality, we need to evaluate all the annotations resulting from the transfer. In the following subsections, we will illustrate two existing evaluation approaches and add our proposal for a more general and effective evaluation framework. Moreover, we will evaluate the output of our algorithms applying the presented metrics.

In order to carry out the evaluation, we divided both corpora into a development set and a testset. The former was used to tune the transfer algorithms, while the latter was employed to run the algorithms and carry out evaluation, comparing the output to the Italian gold standard. The *Europarl* corpus was split into a devset of 300 sentences and a testset of 687 sentences. The *MultiBerkeley* corpus comprised a development set of 100 sentences and a testset of 300 sentences.

5.1 Evaluation 1

In the evaluation of frame information transfer between English and German and English and French, [12] and [14] proposed to evaluate the task following three main criteria: first, they do not consider target transfer because they focus only on FE transfer. Second, they consider for evaluation only the subset of parallel sentences in the source and target gold standard having the same frame, in order to focus on the alignment and transfer quality and exclude free translations from evaluation. Third, they propose to measure performance only on frame elements using the

[2] Even if the general meaning of the first sentence might be related to the CONVEY_IMPORTANCE frame, the *say* alone is considered as a lexical unit of STATEMENT, while *clearly* should be assigned to the OBVIOUSNESS frame.

"Exact match condition", i.e. both the label and the span of the projected role have to match the gold standard annotation for the target language to count as a true positive. We first apply the same evaluation framework and compare the results obtained with algorithm 1 and 2 on *Europarl* to the results obtained by [14] for the English-French pair, given that they worked on the same subset of sentences taken from *Europarl* and used the same English gold standard. Since Italian and French are both romance languages, we assume that they should show the same degree of syntactic and semantic similarity to English. Results are reported in Table 3.

Europarl	**Precision**	**Recall**	**F1**
Algorithm 1	0.48	0.39	0.43
Algorithm 2	0.66	0.40	0.50
MultiBerkeley	Precision	Recall	F1
Algorithm 2	0.75	0.49	0.59

Table 3: *FE transfer evaluation 1 on MultiB.*

The second algorithm improves on the first for every measure. The constituent alignment strategy based on word overlap outperforms the head alignment approach, especially in precision, while recall seems to remain a weak point of both approaches. [14] report that the best full constituent-based model on the French testset, with filters for non-aligned words and arguments, achieves 63.1 as best f-measure (0.66 precision, 0.60 recall). Our best results on *Europarl* scored the same precision but a lower recall. This discrepancy may depend on different algorithm strategies (see Section 3) but also on different characteristics of the two corpora. In fact, frame instance parallelism between English and French gold standards is higher than between English and Italian, with 0.69 frame parallelism and 0.88 FE parallelism (vs. 0.61 and 0.82 on English-Italian Europarl, see Section 4). Besides, the French parser used in the pre-processing phase scores 76.3 f-measure, whereas the Bikel parser trained on Italian has 70.79 f-measure. In order to verify the impact of wrong parse trees on the algorithm performance, we applied the transfer algorithm also to the parsed Italian sentences after a manual correction of the major nodes. The corresponding evaluation on *Europarl* highlighted that for algorithm 1 the correction step enhances precision of 0.14 and recall of 0.12. With the second algorithm, the values improved respectively of 0.14 and 0.9. This proves that parsing problems are a relevant source of error.

As for *MultiBerkeley*, we could not apply algorithm 1 because it requires the source sentences to be represented as syntactic trees, whereas the English FrameNet corpus has annotation pointing to flat chunks without parsing information. Also for this second corpus, we evaluated the improvement of the algorithm on manually corrected parse trees on the Italian side. Precision scores an enhancement of 0.16, and recall of 0.11. The improvement via correction step is greater for *MultiBerkeley* than for *Europarl*. This means that in *MultiBerkeley* parsing problems are the main source of error, whereas in the *Europarl* corpus also other factors have a significant impact on the algorithm performance, for instance free translations. In general, we notice that the transfer approach performs better on a corpus like *MultiBerkeley*, where syntactic complexity is limited by the sentence length and the faithful translation of the parallel sentences enhances the performance of the aligner.

5.2 Evaluation 2

[1] presented a fully automatic transfer process based on alignment with *Moses* [11] at chunk level between English and Italian parallel sentences and a selection of the best candidate segment for semantic transfer according to some ranking and post-processing criteria. The algorithm was evaluated on the same subset of *Europarl* corpus that we used. However, they apply an evaluation framework that is different from that of [12] presented in the previous section. In fact, they consider each FE and target annotation as independent and include in the testset only those FEs having the same label both in the Italian and in the English gold standard. In order to compare this approach to ours, we decided to adopt the same evaluation measures. Accuracy is evaluated on all semantic elements of the target language (both *targets* and *frame elements* together) and only on FEs. The transfer of target annotations was considered correct if the alignment was correct, even if the frame labels were different in the two languages. As for FEs, two kinds of match were computed: Perfect Matching (the projected segments in the target language exactly match with the gold standard ones) and Partial Matching (the intersection between the target projected segments and the ones in the gold standard is not empty). Moreover, in order to measure the gap between perfect and partial matching, evaluation included also token precision, recall and f-measure computed over all transferred labels (micro-average). In Table 4 we report the evaluation of our annotation transfer with algorithm 2, which conveys better performance than algorithm 1, run on the *Europarl* corpus following the above mentioned criteria. We show the results of perfect and partial match applied to all semantic elements (targets + FEs), while the values for FEs only are reported between parenthesis.

Europarl	**PerfMatch** (FEs only)		**PartialMatch** (FEs only)
	0.77 (0.66)		0.90 (0.89)
Token	**Precision**	**Recall**	**F1**
	0.83 (0.82)	0.75 (0.78)	0.79 (0.80)

Table 4: *Evaluation 2 of Alg. 2 on Europarl*

The best model reported in [1] on the same testset scored 0.73 PerfMatch and 0.90 PartialMatch on LUs+FEs, and 0.42 and 0.78 respectively as PerfMatch and PartialMatch on FEs only. This means that both approaches reach high accuracy on target words, whereas our model performs significantly better on FEs only. In general, the two results reflect the different goals of the two approaches: [1] are interested in investigating and adopting unsupervised techniques

with poor semantic and syntactic information to automatically annotate a large scale (but noisy) training set and exploit it for semantic role labelling. On the contrary, we are interested in developing annotated resources with nearly manual quality, so we consider particularly important FE transfer precision.

We report in Table 5 the evaluation of algorithm 2 on the *MultiBerkeley* corpus following the same criteria mentioned above.

MultiBerkeley	**PerfMatch** (FEs only)	**PartialMatch** (FEs only)
	0.84 (0.75)	0.92 (0.88)
Token	**Precision**	**Recall**	**F1**
	0.88 (0.85)	0.84 (0.86)	0.86 (0.85)

Table 5: *Evaluation 2 of Alg. 2 on M.Berkeley*

As expected, the algorithm behaves differently on the two corpora, and all values obtained on *MultiBerkeley* outperform those on *Europarl*, except for PartialMatch on FEs only (0.88 vs. 0.89). This may depend on the fact that the constituents in the *MultiBerkeley* corpus are generally quite short, so the annotation transfer tend to be either a perfect match or to fail. On the contrary, the constituents in the *Europarl* sentences tend to be more complex, thus it is likely that they have at least one aligned token with the English source FE that matches with the gold standard, but exact match is less probable.

5.3 Evaluation 3: a proposal

A common feature of the two evaluation frameworks presented in Section 5.1 and 5.2 is that they exclude from evaluation cases of missing parallelism between source and target sentences. We propose a third approach based on 3 main ideas: 1) we think that it is preferable to evaluate separately targets and frame elements, because of the different nature of the two tasks: target transfer is more influenced by word alignment quality and is generally more straightforward than FE projection. On the other hand, the latter requires a different strategy because it involves selection procedures at chunk or constituent level. While target projection is mainly based on single-word alignment, FE projection requires both role identification and boundary detection. 2) Since we are interested in the (semi) automatic creation of FrameNet for new languages, we want to evaluate the quality of the resulting corpus as a whole, so we consider all transferred annotation regardless of parallelism between the two gold standards. 3) As for the evaluation of FE transfer, we propose two different criteria for assessing the match between automatic annotation and gold standard that are looser than the exact match condition. In both cases, the automatically annotated FE matches the gold standard FE if they share at least the same semantic head. However, type 1 is more strict in that it requires that also the annotation of the corresponding targets match. Type 2, instead, considers correct all matching frame elements between automatic and manually annotated sentences regardless of whether the target has been annotated with the right frame.

We report in Table 6 the evaluation of target transfer on the two corpora. We don't distinguish between algorithm 1 and algorithm 2 on the *Europarl* corpus because the alignment step for targets is the same and relies on word alignment.

	Precision	**Recall**	**F1**
Europarl	0.71	0.50	0.59
MultiBerkeley	0.93	0.81	0.86

Table 6: *Target transfer evaluation*

Europarl	**Precision**	**Recall**	**F1**
Algorithm1			
Type 1	0.46	0.30	0.37
Type 2	0.64	0.41	0.49
Algorithm2			
Type 1	0.55	0.28	0.37
Type 2	0.64	0.32	0.43

Table 7: *FE transfer evaluation 3 on Europarl*

In Table 7 we report the evaluation of FE transfer on the *Europarl* corpus according to the two criteria we have proposed, using both algorithm 1 and algorithm 2. The results reflect different features of the two algorithms that had not been highlighted in the previous evaluations. In particular, algorithm 2 achieves a better performance on precision for evaluation Type 1, but the overall recall value are worse for both types. Since FE transfer in algorithm 2 depends on a correct target transfer, it is clear that missing target alignments influence in turn also the FE transfer performance. The evaluation shows that it is probably better to make the two transfer steps independent, like in algorithm1, so that one can try and align FEs even if no target has been transferred. In Table 8 we report the evaluation of FE transfer on the *MultiBerkeley* corpus according to the two criteria we have proposed and applying algorithm 2.

MultiBerkeley	**Precision**	**Recall**	**F1**
Type 1	0.68	0.54	0.60
Type 2	0.69	0.55	0.61

Table 8: *FE transfer evaluation 3 on MBerk.*

All results on *MultiBerkeley* generally achieve an improvement w.r.t. *Europarl*, particularly on recall. This can be explained by the nature of the corpus, that maximizes word alignment, so that less constituents are left out in the alignment step. Moreover, we noticed in the *Europarl* corpus a greater difference between type 1 and type 2 than in *MultiBerkeley*. In fact, in the former there are a lot of frames that are semantically related and share the same frame elements (for example *Cognizer* is a core FE of several frames in the corpus such as AWARENESS, CERTAINTY, COMING_TO_BELIEVE, JUDGMENT, OPINION, etc.). For this reason, the set of all matching frame elements between automatic and manually annotated sentences regardless of the frame identity (type 2) is bigger than that

considering also the corresponding target match (type 1). In *MultiBerkeley*, instead, the two sets almost coincide because the frame variability is much higher, thus it is less likely that two frame elements of different sentences are the same even if the frame is different.
Error analysis shows that transfer quality of the *Europarl* corpus is crucially affected by syntactic complexity and free translation of the target corpus, which in turn impact on alignment quality. See the example reported at (3):

(3) EN: 85% of Mexico's exports go north.
ITA: L'85 percento delle esportazioni messicane è destinato all'America del nord.
(Literal transl.: 85 percent of Mexican exports are destined to North America)

In order to determine the parallelism between the two sentences, we need to make the inference that North America is north of Mexico, which is out of the current capability of any word-alignment tool. Furthermore, *"go"* and *"essere destinato (to be destined)"* do not exactly express the same predicate and it is likely that they won't be aligned. Other problems involve both corpora and arise from different interpretations given by the annotators to the aligned sentences, which may depend also on inherent ambiguity of FrameNet definitions. For example, in the STATEMENT frame, English annotators tend to prefer to label as *Topic* the content of the communication, whereas in Italian it is mostly annotated as *Message*. Probably the difference between the two frame elements is not clear enough, especially if not applied to English. Other minor problems depend on the recognition and alignment of multiwords in Italian. In general, both algorithms fail to find the correct constituent for frame element transfer in case of complex interpolated tree nodes, where different terminals and nodes dominated by the same parent bear different FE labels.

6 Conclusions and future work

In this work, we presented two algorithms for the crosslingual projection of frame semantic information and tested it on an English-Italian parallel corpus extracted from *Europarl*. Since the comparative evaluation of the two algorithms highlighted advantages and disadvantages for each approach, we think that a combination of the two algorithms should be implemented, trying to preserve the good precision performance of algorithm 2 and to improve on recall via the head-based approach of algorithm 1. Another main concern of our investigation was to understand to what extent different types of corpora can influence the transfer process. For this reason, we tested and evaluated algorithm 2 also on the *MultiBerkeley* corpus, which was produced by manually translating a selection of English sentences from the Berkeley FrameNet database. While evaluation results on the *Europarl* subcorpus were still unsatisfactory because they did not allow for a completely automatic development of FrameNet-like resources, we noticed that *MultiBerkeley* allowed to optimize algorithm performance and minimize alignment errors. Evaluation results show that the translation effort to produce the corpus is repaid by the remarkable reduction of correction work. On the other hand, we are aware that transferring only one sentence per frame and controlling translation allows to cover only one of the possible valence patterns of the frame. For this reason, we believe that the methodology should be considered only a starting point for the creation of FrameNet-like resources for languages different from English. For example, it would be interesting to investigate procedures to automatically acquire new example sentences starting from *MultiBerkeley*, also exploiting existing lexical resources like MultiWordNet. In the future, we plan to improve the projection algorithm exploiting all annotation layers present in the FrameNet corpus. In particular, information about the grammatical function of frame elements could help improving constituent alignment and candidates selection.

References

[1] R. Basili, D. D. Cao, D. Croce, B. Coppola, and A. Moschitti. Cross-language frame semantics transfer in bilingual corpora. In *Proceedings of CICLing*. Springer-Verlag, 2009.

[2] L. Bentivogli and E. Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(03):247–261, 2005.

[3] A. Burchardt and A. Frank. Approximating Textual Entailment with LFG and FrameNet Frames. In *Proceedings of the 2nd PASCAL RTE Workshop*, Venice, Italy, 2006.

[4] A. Burchardt, A. Frank, S. Padó, and M. Pinkal. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974, 2006.

[5] B. Coppola, A. Moschitti, S. Tonelli, and G. Riccardi. Automatic FrameNet-Based Annotation of Conversational Speech. In *Proceedings of the 2nd IEEE Workshop on Spoken Language Technology*, Goa, India, 2008.

[6] A. Corazza, A. Lavelli, and G. Satta. Analisi Sintattica-Statistica basata su Costituenti. *Intelligenza Artificiale*, (2):38–39, 2007.

[7] L. Cyrus. Building a resource for studying translational shifts. In *Proc. of 5th LREC*, Genoa Italy, 2006.

[8] C. Fillmore, C. Johnson, and M. R. L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16:235–250, September 2003.

[9] R. Johansson and P. Nugues. A FrameNet-based Semantic Role Labeler for Swedish. In *Proc. of Coling/ACL 2006*, 2006.

[10] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit*, 2005.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source tool for statistical machine translation. In *Proceedings of ACL 2007*, Prague, CZ, 2007.

[12] S. Padó. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Universität des Saarlandes, 2007.

[13] S. Padó and M. Lapata. Cross-linguistic Projection of Role-Semantic Information. In *Proceedings of Human Language Technology Conference and EMNLP*, pages 859–866, Vancouver, Canada, 2005.

[14] S. Padó and G. Pitel. Annotation précise du français en sémantique de roles par projection cross-linguistique. In *Proceedings of TALN-07*, Toulouse, France, 2007.

[15] E. Pianta and L. Bentivogli. KNOwledge Intensive Word Alignment with KNOWA. In *Proceedings of Coling 2004*, pages 1086 – 1092, 2004.

[16] D. Shen and M. Lapata. Using Semantic Roles to Improve Question Answering. In *Proceedings of EMNLP and CONLL*, pages 12–21, Prague, CZ, 2007.

[17] C. Subirats-Rüggeberg. *FrameNet Español: un análisis cognitivo del léxico del español*. Peter Lang, Frankfurt am Main, 2009.

[18] S. Tonelli and E. Pianta. Frame Information Transfer from English to Italian. In E. L. R. Association, editor, *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.