Dependency and Constituency in Translation Shift Analysis

Dependency and constituency in translation shift analysis

Manuela Sanguinetti and **Cristina Bosco** and **Leonardo Lesmo**
Universitá di Torino
Dipartimento di Informatica
Italy
{manuela.sanguinetti;cristina.bosco;leonardo.lesmo}@unito.it

Abstract

Exploiting data from a parallel treebank recently developed for Italian, English and French, the paper discusses issues related to the development of a dependency-based alignment system. We focus on the alignment of linguistic expressions and constructions which are structurally different in the languages that have to be aligned, and on how to deal with them using dependency rather than constituency. In order to analyze in particular the shifts related to syntactic structure, we present a selection of cases where a dependency-based and a constituency-based representation has been applied and compared.

1 Introduction

In the last few years several resources have been developed for improving Machine Translation tools, applying corpus-based approaches. Among them, there are parallel multilingual treebanks, which are also valuable for the extraction of linguistic knowledge and for translation studies. Their usefulness can be strongly improved by data alignment in particular on the syntactic level, but this task is very time-consuming if manually performed and especially challenging for automatic systems.

The main challenge for such kind of systems is the alignment of linguistic constructions which are expressed by different structures in different languages. Based on past work on translational divergences, or *shifts* – according to Catford's terminology (Catford, 1965) – we thus present in this paper a corpus-based analysis and a comparison, with respect to translation shifts and their possible alignment, of parse tree pairs represented both in a dependency and constituency-based format. The aim of our research is to create a syntax-driven alignment system for parallel parse trees. Our intuition is that, as it has been shown for other tasks, the use of syntactic information on dependency relations and on the predicative structure provided by annotated corpora can be useful while tackling the alignment task, and, as a result, for translation purposes. We therefore developed an alignment system based on dependency information. While our alignment system is now at a prototyping stage, what we intend to define in this paper is a feasibility study on the information that could be exploited by such system. Moreover, in order to examine whether and to what extent the dependencies are able to capture parallelisms, we compared them to a constituency representation. The observations emerged from this study, as well as being the main focus of this paper, constitute the theoretical framework upon which our alignment system can be based. For the preliminary nature of our research, the approach is strongly rule-based, and this allows us to have more control over what information is actually relevant, and which is not.

The paper is organized as follows: after a presentation of the main contributions presented in the last decade concerning parse tree alignment, we describe the linguistic resource we used for our study, focusing on both size and annotation formats applied to the treebank. In the last sections we provide some detailed analyses of the data, and we present a selection of shifts where dependency and constituency-based representations have been compared, with final remarks on the observations emerged from the comparison.

2 Parse tree alignment and related work

When it comes to parse tree alignment, the structures involved are mostly represented in the form of syntactic constituents. Alignment of constituency trees typically includes a sub-sentential level: first, a lexical mapping is performed to terminal nodes (i.e. words), then the non-terminal nodes (i.e. phrases) are aligned so that ances-

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 282–291,
Prague, August 27–30, 2013.

tor/descendant in the source tree are only aligned to an ancestor/descendant of its counterpart in the target tree (Tiedemann, 2011; Tinsley et al., 2007; Wu, 1997). Constituency paradigm is still the most common and widespread in the field of parsing and treebank development, and phrase alignments are considered useful for Syntax-based Machine Translation (which is, in fact, the main use of aligned parallel resources) (Chiang, 2007; Tiedemann and Kotzé, 2009), or for annotating correspondences of idiomatic expressions (Volk et al., 2011). Furthermore, they were also used to make explicit the syntactic divergences between sentence pairs, as in Hearne et al. (2007). In this work in particular the major benefit from aligning phrase structures is claimed to be the opportunity to infer translational correspondences between two substrings in the source and target side by allowing links at higher levels in the tree pair.

Our hypothesis is based on the fact that certain equivalence relations, despite divergences in translations, can be detected using dependency trees. This hypothesis is supported in literature by some previous work on the alignment of deep syntactic structures. For example Ding et al. (2003) developed an algorithm that uses parallel dependency structures to iteratively add constraints to possible alignments; an extension of such work is that of Ding and Palmer (2004), who used a statistical approach to learn dependency structure mappings from parallel corpora, assuming at first a free word mapping, then gradually adding constraints to word level alignments by breaking down the parallel dependency structures into smaller pieces. Mareček et al. (2008) proposed an alignment system of the tectogrammatical layer of texts from the Prague Czech-English Dependency Treebank[1] with a greedy feature-based algorithm that exploits some measurable properties of Czech and English nodes in the corresponding tectogrammatical layers. Among these works, three in particular presented a common approach consisting in the creation of an initial set of word alignment which is then propagated to the other nodes in the source and target dependency trees using syntactic knowledge, formalized in a set of alignment rules (Menezes and Richardson, 2001; Ozdowska, 2005) or extracted by means of unsupervised machine learning techniques (Ma et al., 2008).

Our approach to the alignment task has been largely inspired by such works. What we seek to verify is how such an approach can be a valid alternative to classical phrase-based ones, especially when encountering translational shifts and linguistic differences of various nature.

3 Annotations and data

In this section we describe the data exploited in our study, focusing on the dependency and constituency formats applied to the parallel treebank, together with a brief overview of its size and content.

3.1 Annotation formats

The resource exploited in this study, i.e. ParTUT[2], is a parallel dependency treebank annotated according to the principles and using the same tags for Part of Speech (PoS) and syntactic labels of the Italian monolingual treebank TUT (Turin University Treebank[3]), whose format has been the reference for parsing evaluation campaigns[4], on which is currently defined the state-of-the-art for Italian. TUT trees can be partially compared to surface-syntactic structures (*SSyntS*) as proposed in the Meaning-Text Theory (Melčuk, 1988) and to the analytical layer in the Prague Dependency Treebank style (Böhmová et al., 2003).

As far as the native TUT dependency format is concerned, it uses projective structures whose nodes are labeled with words, and whose arcs are labeled with the names of syntactic relations. Figure 1 shows an example of a typical TUT tree. The arc labels include two components: the second one specifies if the dependent is an argument (ARG) or a modifier (in this case there are only *restrictive* modifiers: RMOD). The first component is the category of the governing item, in case the relation is ARG, or of the dependent, in case of RMOD. In some cases, the subcategory (type) is also included (after the plus sign). So PREP-RMOD should be read as *prepositional restrictive modifier* and DET+DEF-ARG as *argument of a definite determiner*. Note that, in TUT, the root of noun groups is the Determiner (if any), while the root of a prepositional group is the Preposition, as prescribed in the *Word Grammar* (Hudson, 1984) theoretical framework. In the actual TUT

[1] http://ufal.mff.cuni.cz/pcedt2.0/

[2] http://www.di.unito.it/~tutreeb/partut.html

[3] http://www.di.unito.it/~tutreeb

[4] http://www.evalita.it/

there is a third component (omitted here) concerning the semantic role of the dependent with respect to its governor. An important feature is that the format is oriented to an explicit representation of the predicate-argument structure, which is applied to Verb, but also to Nouns and Adjectives; to this end, a distinction is drawn between modifiers and subcategorized arguments and between surface and deep realization of any admitted argument. TUT format is also enhanced by a trace filler mechanism to deal with discontinuous structures, pro-drops and elliptical constructions. Furthermore, compound nouns and contracted forms are split into their components, with an associated node in the parse tree for each of them. The same happens for multi-word espressions, where each of their components is associated with a different node, although in this case they share the same lexical (i.e lemma) and morpho-syntactic information. This means, for example, that the Italian preposition in the example Figure "*dei*", resulting from the contraction between the preposition "*di*" (*of*) and the masculine plural article "*i*" (*the*), is split in two distinct nodes for each of their components.

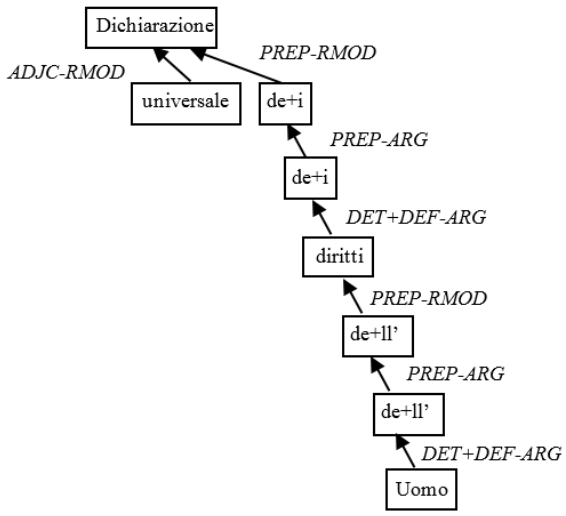

Figure 1: Example of the Italian sentence "*Dichiarazione Universale dei Diritti dell'Uomo*" (*Universal Declaration of Human Rights*) annotated in the TUT format.

The resource has been made available by conversion also in other formats, among them TUT-Penn, a format compliant (except for a few aspects described below) with the English Penn Treebank (PTB) standard. TUT-Penn has a richer morpho-syntactic tag set than Penn format, but it implements almost the same syntactic structure. With respect to the syntactic annotation, it differs from PTB only for some particular constructions and phenomena. It features, for example, a special representation for post-verbal subjects: though a quite common phenomenon in Italian, this is typically challenging for phrase structures (since the subject is considered as external argument of the VP). The standard PTB inventory of null elements is also adopted in TUT–Penn, but while for English null elements are mainly traces denoting constituent movements, in TUT–Penn they can play different roles: zero Pronouns, reduction of relative clauses, elliptical Verbs and also, as said before, the duplication of Subjects which are positioned after Verbs.

These two types of representation, i.e. TUT and TUT-Penn, are those used in our study (in Figure 2 the two formats are shown in parallel)[5]; the observations emerged during their comparison with respect to the alignment issue are described in Section 5.1.

3.2 Data set size and content

ParTUT includes 3,184 sentences corresponding to 85,821 tokens: 28,772 for Italian, 30,118 for French and 26,931 for English, organized in different sub-corpora and text genres, as outlined in Table 1. The content of each corpus varies from legal texts, namely legislative texts of European Community (JRCAquis)[6], to texts extracted from the proceedings of the European Parliament (Europarl)[7] and the Creative Commons license (CC)[8], from the Universal Declaration of Human Rights (UDHR)[9] to instructions on how to create a new Facebook account (FB) and multilingual transcriptions of TED talks[10] (WIT3)[11].

Although the limited size of the treebank, which is still far from being a representative resource of the languages involved, the variety of genres included in the collection also allows to detect some

[5]While for the implementation of the alignment tool we use data annotated with TUT labels but formatted in CoNNL tabs.

[6]http://langtech.jrc.it/JRC-Acquis.html

[7]http://www.statmt.org/europarl/

[8]http://creativecommons.org/licenses/by-nc-sa/2.0

[9]http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx

[10]http://www.ted.com/talks

[11]https://wit3.fbk.eu/

relevant linguistic phenomena and their regularity.

Corpus	sentences	tokens
JRCAcquis_It	181	5,984
JRCAcquis_En	179	4,705
JRCAcquis_Fr	179	6,580
UDHR_It	76	2,072
UDHR_En	77	2,293
UDHR_Fr	77	2,329
CC_It	96	3,252
CC_En	88	2,507
CC_Fr	102	3,097
FB_It	115	1,893
FB_En	114	1,723
FB_Fr	112	1,964
Europarl_It	505	14,051
Europarl_En	515	14,204
Europarl_Fr	480	14,480
WIT3_It	97	1,520
WIT3_En	92	1,499
WIT3_Fr	99	1,668
total	3,184	85,821

Table 1: Corpora and size of ParTUT

4 Data analysis

We applied several different analyses to the data using a set of tools which take, as input, data in the native TUT format. The assumptions of our analysis are based on preliminary studies (Sanguinetti and Bosco, 2012) on the presence, and their classification, of translation shifts in the dataset. The results of those studies had shown that, as expected, the highest number of shifts occurred essentially on the morpho-syntactic and, especially, structural level (see Section 5 for their description). In order to both support and integrate those preliminary studies, in the current analysis we focused our attention basically on the degree of structural complexity of the texts in the different languages, described in terms of word order and dependency distance (Hudson, 1995). We also selected these two metrics as they are good indicators of potential cross-linguistic differences and translational divergences, as well as discrepancies in the structural representation using different formalisms.
As a side effect of the application of these tools, we also obtained a validation and an improved quality of the data set.

4.1 Word order

As for the word order (whose statistics are summarized in Table 2), although the high number of contributions in literature on the matter, it is difficult to find quantitative and cross-language results about the behavior of languages with respect to the movement of major constituents within the sentence structure. A reliable and wide study about word order should be based on a carefully balanced very large dataset, and this goal is beyond the scope of this work. The limits of our analysis are those imposed by the limited size and content of the dataset currently available, the results obtained, however are in line with common knowledge on typical behaviours of English, Italian and French with respect to this issue.
In our analysis, we focused mainly on four elements, i.e. Verb, Subject, Object and Complement[12] and on their relative positions within the sentence. We excluded in advance from the analysis data such as marked structures, interrogative and relative clauses, or infinitival structures, in order to concentrate our attention on unmarked declarative clauses only. For the same motivations we did not consider expletive, progressive and passive structures. The remaining verbal structures consist of 782 clauses for English, 886 for French and 597 for Italian distributed within the three monolingual treebanks. The far smaller amount of verbal structures taken into account for Italian is motivated by the exclusion of structures affected by pro-drop, i.e. the absence of subject in finite clauses, which occurs 32.6% of unmarked declarative verbal structures.
The most frequent word order for all the three languages is the classical SVO, as assumed in literature (Dryer, 1998); however, if we focus our attention on the relative position of Subject and Verb, a typical issue that can be problematic for constituency-based formats, we can see that this phenomenon is quite rare in French (4.7%) and English (7.3%), but far more frequent in Italian (17.1%) verbal structures. Such figures, as far as Italian language is concerned, are in line with the results obtained in previous studies on the influence of the constituent order on data-driven parsing (Alicante et al., 2012), where the Subj/Verb order is attested at 79.10% and its inverted order

[12] We encompassed on the label Compl the Indirect Object, the Agent complement, predicatives and other indirect complements that act as arguments of the verb encountered.

at 20.90%.

Language	order	frequency
Italian (597)	Subj/Verb(/Obj)	74.5%
	Subject after Verb	17.1%
	Compl between	9.9%
French (886)	Subj/Verb(/Obj)	82.4%
	Subject after Verb	4.7%
	Compl between	9.4%
English (782)	Subj/Verb(/Obj)	88.5%
	Subject after Verb	7.3%
	Compl between	1.02%

Table 2: Word order in the ParTUT languages.

The well-known assumption that English is featured by a fixed word order, with respect to French or Italian, is clearly attested by our results also observing that in the former it is very rare that a Complement or an Object is positioned between Subject and Verb. In Italian and French various kinds of Complements can be positioned between Subject and Verb[13], thus making the structure more complex.

4.2 Dependency distance

Concerning the results obtained in the analysis of dependency distance, which is measured here as the distance between words and their parents in terms of intervening words (Hudson, 1995), we considered also its correlated measure, that of dependency direction, i.e the contrast between governor-initial (which means that the position number of the governor is lower than that of the dependent) and governor-final dependencies (see Table 3). In view of a comparison with a constituency representation, this measure is a good indicator of how the relationship between a dependent and its head, within a dependency framework, is still preserved despite their distance, and the direction of this distance. This seems even more important when we have to find correspondences between parallel parse trees in different languages.

The distance is reported in terms of percentage of dependencies, while the direction is expressed by the labels POS (POSITIVE, i.e. governor-initial cases) and NEG (NEGATIVE, i.e. governor-final cases). With respect to this matter, we observed that English has a higher number of dependency relations with governor-final cases (25.19%), although their distance is lower if compared to Italian and French. This could be easily explained by the higher frequency of English premodifiers, with respect to Italian and French.

Despite the small amount of data available for our experiments, from a comparison of the data for the Italian in ParTUT and those extracted from the TUT monolingual treebank[14] (a more extended dataset, with a different text composition from the multilingual treebank) there is a substantial similarity with respect to dependency distance and its direction (see the rightmost column in Table 3). In light of this, we expect similar results for English and French as well, once we can rely on a larger dataset.

Distance	En.	Fr.	It.	TUT
POS	74.81	81.91	81.01	76.65
POS <=10	98.12	97.89	97.72	97.88
10 > POS < 20	1.43	1.64	1.72	1.62
POS >= 20	0.45	0.47	0.56	0.49
NEG	25.19	18.09	18.99	18.59
NEG <=10	95.70	92.81	93.62	92.24
10 > NEG < 20	2.99	4.91	4.54	5.17
NEG >= 20	1.31	2.28	1.84	2.58

Table 3: The table shows statistics on dependency distance and direction distributed per language, with a comparison of Italian data of ParTUT with the overall figures extracted from the monolingual treebank TUT.

5 Translation shifts and their alignment

The search for matches between pairs of non-isomorphic trees requires an extended knowledge (whether formalized by a set of rules or learned automatically) on the divergences, or shifts, that may occur during the translation process. While designing our alignment system, we attempted in a first step to determine what types of shifts may be encountered in ParTUT. The classification was made on a sample of the treebank sentences extracted from each of the sub-corpora that compose the collection.

This comparison led to a first basic classification[15] which includes essentially three levels:

[13] Such complements are mainly in the form of clitics expressing a direct or indirect object ("***me l'****ont demandé*" – *I was asked to*), or predicative complements ("*non* ***lo*** *sono mai*" – *they are never like that*)

[14] The treebank currently consists of 3,542 sentences and 102,150 tokens.

[15] It was difficult to establish a clear-cut disticntion for each

morpho-syntactic (*Category Shifts*) and syntactic level (*Strucutral Shifts*) on one hand, and that of meaning (*Semantic Shifts*) on the other[16].

Category Shifts may involve a change in the Part of Speech;

Structral Shifts are the most complex and include a number of different situations that can be determined both by linguistic constraints imposed by the respective languages, or, more simply, by individual translator's choice. Structural shifts may thus comprise cases of:

– different word order and discontinuous correspondences;

– passivization/depassivization;

– function word introduction/elimination;

– conflation (i.e. tre translation of two words using a single word equivalent in meaning);

– paraphrases; – idioms.

Semantic Shifts mainly concern the level of meaning; they include cases of:

– addition/deletion (i.e. the introduction or elimination of pieces of information);

– mutation (whenever the correspondence is characterised by a high degree of fuzziness, or the content substantially differs).

In order to handle properly with such divergences, we therefore designed an alignment system that starting from a lexical mapping of the nodes in the tree pair, it moves outwards to the unaligned nodes using the information available on syntactic structure, with a focus in particular on the argument structure (which, in ParTUT is applied to Nouns and Adjectives as well).

The algorithm, which is currently in a prototype implementation stage, includes two distinct steps, respectively referring to the lexical level and to syntactic dependencies.

Step 1: lexical correspondences are identified and stored in lexical pairs; the mapping of source and target nodes is carried out by means of a probabilistic dictionary created using the IMB Model 1 implementation in the Bilingual Sentence Aligner (Moore, 2002).

Step 2: starting from the lexical pairs obtained in the first step, correspondences between neighbouring nodes are verified comparing in parallel the respective relational structure, such that:

$d_s > d_t$ if:
$(w_s;w_t)$
$rel\ (w_s;d_s) = rel(w_t;d_t)$

where d_s and d_t are a source and a target node of a tree pair whose governors are the word w_s and its counterpart w_t; d_s and d_t can be aligned ($d_s > d_t$) whenever their governors are selected as anchor pair (w_s;w_t) during the lexical mapping step, and the syntactic relation *rel* $(w_s;d_s)$ between the source anchor word w_s and its dependent d_s is the same as $rel(w_t;d_t)$, i.e. that between the target anchor word w_t and its dependent d_t. This means that, for example, in the expressions "*no one*" – "*nessun individuo*", given the anchor pair (*no; nessun*), and the syntactic relations ARG(*no; one*) and ARG(*nessun; individuo*), then the alignment can be expanded to the dependents (*one;individuo*).

Our hypothesis is that tree alignment of dependency structures could work because, besides lexicon, it is based on predicative structure, which (provided that this is shared by the two parse trees) will remain stable in different languages despite variations in the realization of the constituents; as a result, whenever the algorithm attempts to search for correspondences between a source and a target dependency tree, it may be able to find, within a reasonable distance from the head of a predicative structure, the relations that make up that structure. This reasonable distance can be approximated by taking into account the elements we reported in the analysis on word order and dependency distance.

5.1 Constituency and dependency: cross-linguistic comparison

In the previous section, we described the overall framework of our alignment system; in this section, we attempt to describe its strengths and weaknesses while comparing trees in ParTUT as represented in the dependency-based TUT format and in the constituency-based converted format TUT-Penn. The comparison mainly deals with the types of shift introduced in Section 5. What emerged from this investigation is that the choice to compare sentence pairs considering their deep structure and relations, rather than grouping them together into constituents, can help

kind of shifts, especially when multiple divergences cooccurred. Their classification was made based on the predominant aspects that characterize each shift.

[16]This classification is similar in spirit to the work by Cyrus (2006), Dorr (1994) and Melčuk and Wanner (2006), and partially adopts their terminology and definitions. In particular, like in Cyrus (2006), we opted for maintaining the notion of *shift* as, in our view, particularly conveying the idea of the transfer that takes place during the process of transposition of meaning from one language to another.

to overcome some of the limitations imposed by such non-isomorphism. This proved true in the case of category shift. With respect to the classic case of nominalization, for example, while a hierarchical costituency representation gives rise to two different phrases, dependents identification of corresponding heads is facilitated by the fact that, as mentioned in Section 3.1, even Nouns are assigned a predicative structure. Dependents are therefore labeled as arguments of a same predicative structure, as in the example below[17]:

(1a) TUT:

1-improving[TOP] *2-the*[1;OBJ] *3-efficiency*[2;ARG]

1-l'[TOP] *1-amélioration*[1;ARG] *3-de*[2;OBJ] *4-l'*[3;ARG] *5-efficacité*[4;ARG]

(*The improvement of the efficiency*)[18]

(1b) TUT-Penn:

(VP (V *Improving*) (NP (ART *the*) (N *efficiency*)))

(NP (ART *L'*) (NP (N *amélioration*) (PP (PREP *de*) (NP (ART *l'*) (N *efficacité*)))))

As they include linguistic aspects of various nature, structural shifts require broader and more articulated considerations. On the one hand the dependency structure, and in particular the predicative structure as encoded in TUT, once again may be useful in overcoming translational divergences and reducing them to a common structure. This is the case, for example, for long-distance dependencies - which are difficult to represent as such in a phrase structure - but also for word order. Below we report an English-Italian bisentence that may exemplify this issue:

(2a) TUT:

***1-the*[18;SUBJ]** *2-exchange*[1;ARG] *3-of*[2;RMOD]
4-information[3;ARG] *5-on*[4;RMOD]
6-environmental[9;RMOD] *7-life*[8;RMOD]
8-cycle[9;RMOD] *9-performance*[5;ARG]
10-and[5;COORD] *11-on*[10;COORD2ND] *12-the*[11;ARG]
13-achievements[12;ARG] *14-of*[13;RMOD]
15-design[16;RMOD] *16-solutions*[14;ARG] *17-is*[18;AUX]
18-facilitated[TOP]

1-è[2;AUX] ***2-agevolato*[TOP]** ***3-uno*[2;SUBJ]**
4-scambio[3;ARG] *5-di*[4;RMOD] *6-informazioni*[5;ARG]
7-su[6;RMOD] *8-l'*[7;ARG] *9-analisi*[8;ARG] *10-di*[8;RMOD]
11-la[10;ARG] *12-prestazione*[11;ARG]
13-ambientale[12;RMOD] *14-di*[12;RMOD] *15-il*[14;ARG]
16-ciclo[15;ARG] *17-di*[16;RMOD] *18-vita*[17;RMOD]
19-e[7;COORD] *20-su*[19;COORD2ND] *21-le*[20;ARG]
22-realizzazioni[21;ARG] *23-di*[22;RMOD]
24-soluzioni[23;ARG] *25-di*[24;RMOD]
26-progettazione[25;ARG]

(*is facilitated an exchange of information on the analysis of the environmental life cycle perfomance and on the achievements of design solutions*.)

(2b) TUT-Penn:

((S (NP (NP (ART *The*) (N *exchange*)) (PP (PREP *of*)(NP (NP (N *information*))(PP (PP (PREP *on*)(NP (NP (NP (N life)) (N *cycle*)) (NP (ADJ *environmental*) (N *performance*))))(CONJ *and*)(PP (PREP *on*)(NP (NP (ART *the*) (N *achievements*))(PP (PREP *of*)(NP (NP (N *design*)) (N *solutions*)))))))))(VP (V *is*)(VP (V *facilitated*)))))

((S (VP (V *é*) (VP (V *agevolato*)(NP (ART *uno*)(N *scambio*))(PP (PREP *di*)(NP (NP (N *informazioni*))(PP (PREP *su*)(NP (NP (ART *l'*)(N *analisi*))(PP (PREP *di*)(NP (ART *la*)(NP (N *prestazione*))(ADJP (ADJ *ambientale*)(PP (PREP *di*) (NP (NP (ART *il*) (N *ciclo*)) (PP (PREP *di*)(NP (NP (N *vita*))(CONJ *e*)(PP (PREP *su*)(NP (NP (ART *le*) (N *realizzazioni*))(PP (PREP *di*)(NP(NP (N *soluzioni*))(PP (PREP *di*)(NP (N *progettazione*)))))))))))))))))))))

The English sentence presents a standard Subj-Verb order, although their dependency distance (as measured with the tools used for analysis described in Section 4) equals to 17; on the contrary, its Italian counterpart shows a Verb-Subj order with a positive dependency distance of 1. While such figures affected the phrase structure representation, mainly because of the post-positioned Subject in the Italian version, this was not the case in dependency analysis, where the respective arguments of the corresponding verbs were appro-

[17]The examples are here represented in a compact form where only the major annotated information is shown: for each dependency node we provide information on *position-word[governorposition;relation]*, while for constituency only some phrase label is abbreviated. Each example reports a sentence pair where the source language is always English and the target language is Italian or French. Bold characters are used to highlight the dependency distance between a head and its dependent in the linear order of the sentence (see example 2a and 2b).

[18]The glosses for non-English examples are intended as literal and do not necessarily correspond to the correct English expression.

priately assigned, despite the high distance of the Subject from the main verb in English, thus preserving the parallelism between the two structures, and as a result, their alignment.

The same can be said for passivization, which can be easily detected and aligned by means of the explicit representation of deep relations. Considering the bisentence below, for example, a common predicative structure can be observed for the main verbs in the respective languages, although in the passive form surface syntactic roles are also expressed, so as to specify that the verb has undertaken a transformation: the surface Subject is thus linked to its predicate with the relation [OBJ/SUBJ], meaning that it corresponds to a deep Object. While in the phrase structure the arguments of the predicate are moved during transformation, resulting in a different realization.

(3a) TUT:

1-we[2;SUBJ] *2-allow*[TOP] *3-accounts*[2;OBJ]

1-gli[4;OBJ/SUBJ] *2-account*[1;ARG] *3-sono*[4;AUX] *4-consentiti*[TOP]

(*accounts are allowed*)

(3b) TUT-Penn:

((S (NP (PRO *we*)) (VP (V *allow*) (NP (N *accounts*)))))

((S (NP (ART *gli*) (N *account*)) (V *sono*) (VP (V *consentiti*))))

A more tricky cases are those of paraphrases, idioms and the conflation of two lexical items into a single item semantically equivalent. In Figure 2 we represented in a graphic form an example of paraphrase, where a Verb in English is expressed with a Verb followed by the nominalized form of the English Verb in Italian, and of an idiom in English and its translation in French. In the sub-class of idioms we also included multiword expressions: although their overall presence in the treebank is not so relevant (1,15% in Italian, 0,86% in French and 0,05% in English), it is a phenomenon that we should take into account, as they share with idioms the features of non-compositionality and an idiosyncatic use, which make them a very complex linguguistic phenomenon for several NLP tasks, not only in the alignment issue. It should also be pointed out that, despite the problematic identification of a multi-word unit, in the TUT format a number of these linguistic items are already recognized as such. This means that the aligner also can take advantage of this information, as it is provided in the annotation.

All the aspects mentioned here share some peculiarities that require particular consideration: the difficult identification of these cases, by virtue of both the absence of a direct lexical mapping and a different syntactic realization, may see the need to introduce a more extensive hierarchical notion, such as that of dependency substructure, or *treelet*, introduced in Ding and Palmer (2004)[19]. This could be useful in order to capture possible translational matches at a higher level, abstracting away from pure relations between individual nodes (supporting, though from a dependency perspective, what suggested in Hearne et al. (2007), also reported in Section 2).
Contrarily, for example, to Mel'čuk and Wanner (2006), where the level considered (i.e. the deep-syntactic structure, *DSyntS*) is abstract enough to avoid all types of lexical and syntactic divergences, the dependency format considered in this study, despite the explicit annotation of argumental roles, is more oriented to the representation of the surface dependency structure. The observations posed above, and the examples in Figure 2 suggested us the hypothesis that to overcome these limitations while attempting to map divergent (though translationally equivalent) structures, it is necessary to integrate the current alignment system with an additional layer of abstraction, such that:

$$d_{(s1,\ldots,sn)} > d_{(t1,\ldots,tn)} \text{ if:}$$
$$(w_s;w_t)$$
$$rel\ (w_s;d_{(s1,\ldots,sn)}) = rel(w_t;d_t)$$

where n is the number of nodes comprised in the substructure, and (w_s; w_t) is the lexical pair used as the closest anchor point from which the alignment can be expanded. This means that more than one node that goes down from w_s could be aligned to the subtree that goes down from d_s; i.e., for example, that in the expression given in Figure 2 "*to bring that home*" – "*pour vous faire comprendre*", given the anchor pair (*to; pour*) and the syntactic relations ARG(*to; bring*) and ARG(*pour; faire*) the descending nodes could then be aligned.

[19] As pointed out by the authors, the choice of the term *treelet* was made in order to avoid confusion with *subtree*, as treelets do not necessarily go down to every leaf.

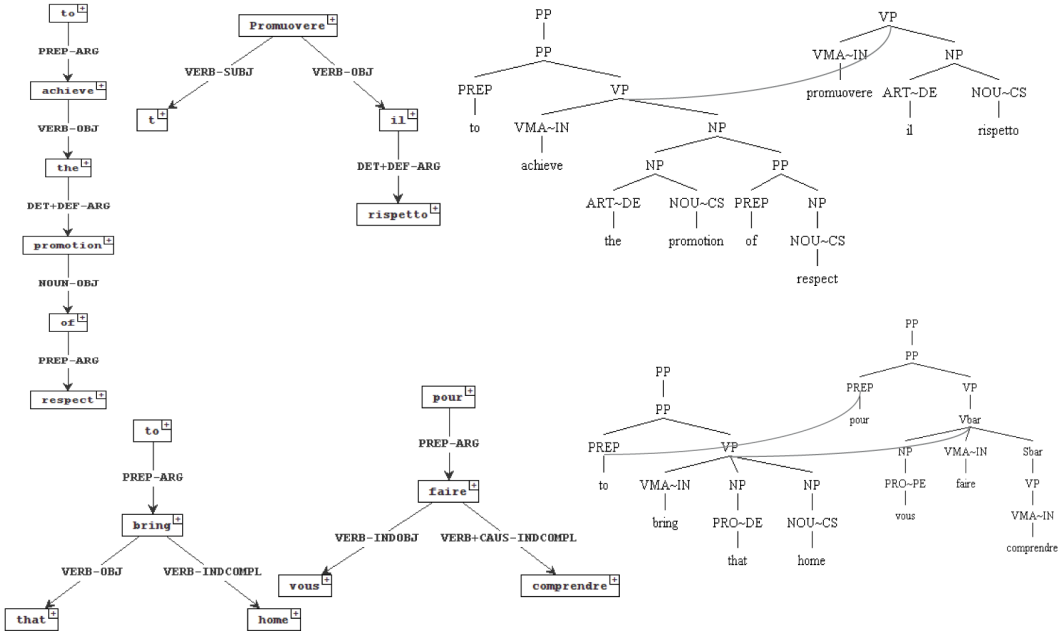

Figure 2: Graphic representations in TUT (on the left side) and TUT-Penn (on the right) of two tree pairs. The first reports a paraphrase in English, "*to achieve the promotion*", of a single Italian verb, "*promuovere*" ("*to promote*"); the second one represents an English idiom,"*to bring that home*", and its French translation, "*pour vous faire comprendre*" ("*to let you understand*"). While an alignment link can be drawn between the correspondent phrases in the constituency format, we are not able to do the same for the nodes in the dependency structures.

5.2 Discussion

Comparing the TUT dependency format to a converted version in the standard Penn Treebank, we came to the conclusion that a number of shifts could be handled with a simple approach that directly uses dependency relations expressed in the format at issue. Structural shifts when the same argumental roles are shared by the parallel trees, or wirh differences in the linear word order or distance are easily linked. However, other cases required a different treatment. Some classes of shifts, in particular those where divergences are due to differences in the idiosyncratic use between the languages or to the low compositionality of the expressions, may require the integration of a more abstract notion of substrucutre, or treelet (which can be partially assimilated to that of constituency subtree) in order to link the entire substructure to its equivalent node, that is to capture translational equivalence between these complex expressions and their counterpart in the other language. This seems to us a viable solution that could balance the limits imposed by the format with the useful linguistic information it provides.

6 Conclusion and future work

In this paper we presented a comparative study between dependency and costituency representation of parallel structures with the aim of verifiyng how and to what extent dependencies are a valuable support in the alignment task. The aim of our research, in fact, is laying the ground for the development of a more linguistically motivated treebank alignment system which could properly exploit linguistic information on dependency structures in order to handle properly translational divergences, or shifts, that may occur on different levels (morpho-syntactic, syntactic or semantic). The linguistic resource we used is a parallel multilingual treebank, ParTUT, where dependency representation is more oriented to the surface order of nodes in the input sentence, rather than a deep semantic representation. Besides the extension of the treebank, in order to make it a more balanced and reliable linguistic resource, the next steps in our research will consist in improving the implementation of the alignment system so that it could consider the notion of treelet, and, in a further stage, in testing more extensively this method also

to other shifts, such as semantic shifts, which constitute an even greater challenge.

References

Anita Alicante, Cristina Bosco, Anna Corazza and Alberto Lavelli. 2012. A treebank-based study on the influence of Italian word order on parsing performance. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1985–1982.

Alena Böhmová, Jan Hajič, Eva Hajičová and Barbora Hladká. 2003. The Prague Dependency Treebank. In *Treebanks*, pp. 103-127, Springer Netherlands.

John C. Catford. 1965. *A linguistic theory of translation: An essay on applied linguistics*, University Press, Oxford.

David Chiang. 2007. Hierarchical phrase-based translation. In *Computational Linguistics*, volume 33, number 2, pp. 201–228, MIT Press, Cambridge, MA, USA.

Lea Cyrus. 2006. Building a resource for studying translation shifts. In *Proceedings of Language Resources and Evaluation Conference (LREC '06)*, Genoa, Italy.

Yuan Ding, Daniel Gildea and Martha Palmer. 2003. An algorithm for word-level alignment of parallel dependency trees. In *The 9th Machine Translation Summit of the International Association for Machine Translation*, pp. 95–101.

Yuan Ding and Martha Palmer. 2004. Automatic learning of parallel dependency treelet pairs. In *Proceedings of The First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Hainan Island, China, pp. 233–243.

Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. In *Computational Linguistics*, volume 20, number 4, pp. 597–633.

Matthew S. Dryer. 1998. Aspects of Word Order in the Languages of Europe. In *Constituent Order in the Languages of Europe*, pp. 283 - 319.

Mary Hearne, John Tinsley, Ventsislav Zhechev and Andy Way. 2007. Capturing translational divergences with a statistical tree-to-tree aligner. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*.

Richard Hudson. 1984. *Word Grammar*. Basil Blackwell, Oxford and New York.

Richard Hudson. 1995. Measuring syntactic difficulty. Unpublished paper. http://www.phon.ucl.ac.uk/home/dick/difficulty.htm

Yanjun Ma, Sylwia Ozdowska, Yanli Sun, Andy Way. 2008. Improving word alignment using syntactic dependencies. In *Proceeding of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pp. 69–77.

David Mareček, Zdeněk Žabortský and Václav Novák. 2008. Automatic alignment of Czech and English deep syntactic dependency tree. In *Proceeding of the 12th EAMT Conference*, Hamburg, Germany.

Igor Mel'čuk. 1988 *Dependency Syntax: Theory and Practice.*. The SUNY Press, Albany, N.Y.

Igor Melčuk and Leo Wanner. 2006. Syntactic mismatches in machine translation. In *Machine Translation*, volume 20, number 2, pp. 81–138 .

Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Methods in Machine Translation at ACL-2001*, pp. 39–46.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: From Research to Real Users, (AMTA '02)*, pp. 135–144.

Silvia Ozdowska. 2005. Using bilingual dependencies to align words in English/French parallel corpora. In *Proceedings of the ACL Student Research Workshop*, pp. 127–132.

Manuela Sanguinetti and Cristina Bosco. 2012. Translational divergences and their alignment in a parallel treebank. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*, pp. 169–180.

Jörg Tiedemann. 2011. *Bitext alignment*. Morgan & Claypool.

Jörg Tiedemann and Gideon Kotzé. 2009. Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT '08)*, pp. 197-208.

John Tinsley, Ventislav Zhechev, Mary Hearne and Andy Way. 2007. Robust language pair-independent sub-tree alignment. In *Proceedings of the MT Summit XI*, pp. 467-474.

Martin Volk, Torsten Marek, Yvonne Samuelsson. 2011. Building and Querying Parallel Treebanks. In *Translation: Computation, Corpora, Cognition*, vol. 1, num. 1, http://www.t-c3.org/index.php/t-c3/article/view/8

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. In *Computational Linguistics*, volume 3, number 3, pp. 377–403.