Translational Divergences and Their Alignment in a Parallel Multilingual Treebank

Translational divergences and their alignment in a parallel multilingual treebank*

Manuela Sanguinetti, Cristina Bosco

Università di Torino
Department of Computer Science
E-mail: {msanguin;bosco}@di.unito.it

Abstract

The usefulness of parallel corpora in translation studies and machine translation is strictly related to the availability of aligned data. In this paper we discuss the issues related to the design of a tool for the alignment of data from a parallel treebank, which takes into account morphological, syntactic and semantic knowledge as annotated in this kind of resource. A preliminary analysis is presented which is based on a case study, a parallel treebank for Italian, English and French, i.e. ParTUT. The paper will focus, in particular, on the study of translational divergences and their implications for the development of an alignment tool of parallel parse trees that, benefitting from the linguistic information provided in ParTUT, could properly deal with such divergences.

1 Introduction

Parallel corpora are currently considered as crucial resources for a variety of NLP tasks (most notably machine translation), and for research in the field of translation studies and contrastive linguistics. In order to be of any help for such purposes, parallel corpora have to be correctly aligned, that is to say that translational correspondences between the pairs of languages involved should be properly detected and exploited. Several approaches have been used in order to automatically align these multilingual resources, based both on deterministic rules or specific heuristics (see, e.g. [11]), and statistical techniques (e.g. [9], and [16]). The latter in particular have been highly successful in recent years. The reasons for such success are manyfold: among them we can find their capability to process even less-studied or resource-poor languages, or the large amount of time required to create robust and accurate rule-based systems. It is our belief, however, that linguistic insights can

*This work has been partially funded by the PARLI Project (Portale per l'Accesso alle Risorse Linguistiche per l'Italiano - MIUR - PRIN 2008).

be of great help and that the presence of a set of rules for the detection of translational correspondences can, if not replace, significantly complement the work done by statistical systems. Linguistic knowledge can be of help in identifying not only the exact matches, but also (we would rather say in particular) all those cases in which there are partial or fuzzy correspondences due, for example, to the individual translator choices or to differences - which often occur in a systematic way - between language pairs.

In this paper we focus our attention especially on these differences (which we will designate with the term "shift"), on their classification, and finally on a proposal for their automatic processing. We start, in Section 2, from the discussion of some works in the field of Translation Studies where reference is made to the notion of translation shift, and we then present an existing resource, i.e. ParTUT, which has been used as the basis for our analysis. The remainder of the paper is structured as follows: in Section 3 we provide a brief description of the corpus content and of the annotation format of the treebank, while Section 4 is devoted to the cases of translation shifts encountered so far in the corpus and to the identification of the main points our research should be addressed to.

2 Related Work

Due to the peculiarities of each language system, translation process may be quite complex and correspondences could be drawn up on various linguistic levels: lexical, structural or semantic; it therefore entails the need to implement strategies (or "shifts") so that the meaning is properly transposed and preserved in such process.

John Catford [1] was the first one to use the term *shift*, which he defined as the departure from formal correspondence in the process of going from the source language to the target language. His definition relied upon the notions of *formal correspondence* and *textual equivalence*. A formal correspondence is a relationship between two linguistic items that play the same role in their respective systems, while textual equivalence is defined, in a broader sense, as any kind of translational equivalence in a pair of texts or sentences. A shift then occurs in translation whenever there is not a formal correspondence relationship between two elements in source and target texts. In his work, Catford discussed two types of shift: level and category shift. The former deals with shifts between different linguistic levels (usually grammar and lexis), while category shift is further subdivided in: class shift (e.g. in the word class used), unit or rank shift (e.g. when a single word is translated by means of a phrase), structural shift (e.g. when word order is modified) – this is considered the most frequent among the category shifts – and intra-system shift (i.e. when, despite the presence of a formal correspondence between source and target elements, a non-corresponding form is chosen while translating).

An important contribution which largely relied upon the linguistic theories described above is that of Cyrus [3], who explicitly annotated translation shifts in a parallel treebank. Contrarily to most of the works on parallel treebanks pro-

posed so far, Cyrus' work did not aim at creating a machine translation system, but at building a resource for English and German (FuSe) in which translation shifts were explicitly annotated on the basis of the predicate-argument structure. The annotation system was based on the distinction between two main classes, i.e. grammatical and semantic shifts. In the first one, all those cases of passivisation–depassivisation, pronominalisation–depronominalisation, as well as category and number changes were included. Semantic shifts comprised the cases when meaning is somewhat involved and were classified as follows: semantic modification, explicitation– generalisation, addition–deletion and mutation.

These works have been an important theoretical reference for our research. The analysis of translational correspondences and divergences made on our small set of sentences, besides taking into account the observed cases, as well as some peculiarities of the corpus (such as the annotation format), has been largely inspired by such works in its theoretical formalization and systematization.

3 Corpus description

3.1 Data

ParTUT[1] currently comprises 42,347 tokens distibuted as shown in Table 1. The corpus consists of an average amount of 465 sentences per language. They were retrieved from the JRC-Acquis multilingual parallel corpus[2] [15] and the entire text (about 100 sentences) of the Universal Declaration of Human Rights[3]. More recently this preliminary set has been enlarged with an additional corpus extracted from the open licence "Creative Commons"[4] composed by around 100 sentences, and from publicly available pages from Facebook website.

The corpus is aligned on the sentence level with the Microsoft Bilingual Sentence Aligner ([12]) and the LF Aligner[5], an automatic tool based on Gale and Church algorithm which enables the storage of sentence pairs as translation units in TMX files and the review of the output in formatted xls spreadsheets.

3.2 The annotation format

The parallel treebank comprises a collection of sentences represented in the form of dependency structures. The dependency relations are described in compliance with the same criteria adopted for the creation of the Italian monolingual treebank TUT (Turin University Treebank)[6].

[1] http://www.di.unito.it/~tutreeb/partut.html
[2] http://langtech.jrc.it/JRC-Acquis.html, http://optima.jrc.it/Acquis/
[3] http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx
[4] http://creativecommons.org/licenses/by-nc-sa/2.0
[5] http://sourceforge.net/projects/aligner/
[6] http://www.di.unito.it/~tutreeb/

Corpus	sentences	tokens
JRCAcquis–It	181	6,304
JRCAcquis–En	179	4,705
JRCAcquis–Fr	179	8,667
UDHR–It	76	2,387
UDHR–En	77	2,293
UDHR–Fr	77	2,537
CC–It	96	3,141
CC–En	88	2,507
CC–Fr	102	3,624
FB–It	115	1,947
FB–En	114	1,723
FB–Fr	112	2,512
total	1,377	42,347

Table 1: Corpus overview.

As far as the native annotation schema is concerned, a typical TUT tree shows a pure dependency format centered upon the notion of argument structure and is based on the principles of the *Word Grammar* theoretical framework [7]. This is mirrored, for instance, in the annotation of Determiners and Prepositions which are represented in TUT trees as complementizers of Nouns or Verbs. By contrast, the native TUT scheme exploits also some representational tools, i.e. null elements (which are non–standard in dependency-based annotations), in order to deal with particular structures, such as pro–drops, long distance dependencies and elliptical structures.

As for the dependency relations that label the tree edges, TUT exploits a rich set of grammatical items designed to represent a variety of linguistic information according to three different perspectives, i.e. morphology, functional syntax and semantics. The main idea is that a single layer, the one describing the relations between words, can represent linguistic knowledge that is proximate to semantics and underlies syntax and morphology, i.e. the predicate-argument structure of events and states. To this end, a distinction is drawn between modifiers and subcategorized arguments on the one hand and between surface and deep realization of any admitted argument; these annotation criteria have proven particularly useful while detecting cases of nominalisations and passivisation, which are common cases of divergence in translation.

In a cross-linguistic perspective, the choice to use a single and coherent representation format, and the TUT format in particular, proved to be suitable in the development of a parallel resource in that the rich morphological tag set allowed an adequate representation for both morphologically rich languages (Italian and French) and simpler ones (English), and the richness and flexibility of relations

allowed an appropriate coverage of new linguistic phenomena encountered so far.

Such format proved also to be useful in comparative analyses, by means of which some typical phenomena of the three languages involved could be easily queried and quantified. Among these we can find the higher frequency of nominal modifiers (4.5%) in English texts (expressed by the NOUN-RMOD label), with respect to Italian (0.9%) and French (0.6%), or, on one hand, the presence (marked, as explained above, by null elements) of pro-drops in Italian sentences (56 occurrences against one single entry in English and the absence in French) and, on the other, the use of expletive subjects in English and French (respectively 19 and 21 occurences). Furthermore, the annotation of Determiners as complementizers of Nouns (as also pointed out above) and the lower frequency of determiners in English (11.6%, compared to 13.4% in French and 16.1% in Italian) led to a different representation of English trees with respect to Ialian.

These preliminary considerations, together with the absence of an appropriate tool that allows the alignment of sentences represented according to the TUT format, have led to the decision of working on the development of a new system. As, in fact, the choice of the aligment tool is strongly related to the final task, we cannot abstract away from the input format, especially if it provides linguistically relevant information which can be useful for any further corpus processing and exploitation.

4 Alignment issues

In this section we describe the main steps of the investigation which has lead us to the definition of our approach for the alignment of the trees in ParTUT.

As a first step, we selected a small subset of sentences from the collection and attempted to align them manually, in order to specifically study the various issues that could arise, and in particular translation shifts. We selected for that purpose the first two subcorpora of the treebank, i.e. the JRC-Acquis and the UDHR, which were already aligned on the sentence level. In order to avoid further drawbacks deriving from multiple or null correspondences – which were caused in some cases by the different segmentation criteria adopted by the parser and the sentence aligner – we only worked on 1:1 sentence pairs (constituting almost the 90% of the overall amounts of sentence pairs). In particular, the experimental corpus was composed by 50 sentences per language divided into three pairs (i.e. Italian - English, English - French, French - Italian). While comparing the tree pairs, we observed the various cases of divergences, or shifts, and attempted to classify them.

Similarly to what proposed in [3], we identified two main classes of shifts, each one involving respectively morpho-syntactic and structural level on one hand, and semantic level on the other. In the first class we include:

Category shift[7]: when different parts of speech are used in source and target text.

[7] Although Catford used this term for describing a main class of shifts that included other sub-

EN: ***Improving*** *the efficiency* [...]
FR: ***L'amélioration de*** *l'efficacité* [...]
(*The improvement of the efficiency*)[8]

Structural shift: this subclass comprises all those cases when syntactic level is directly involved and affected from translator's choices or word order constraints. We then include, for example, the cases of discontinuous correspondences:

EN: *Nor shall* ***a heavier penalty*** *be imposed* ***than*** *the one that was applicable* [...]
IT: *Non potrà del pari essere inflitta* ***alcuna pena superiore a*** *quella applicabile* [...]
(Cannot be likewise imposed any penalty heavier than the one applicable)

passivisation–depassivisation[9]:

EN:*This Directive* ***seeks to*** *achieve* [...]
IT:*La presente Direttiva* ***è intesa a*** *conseguire* [...]
(The present Directive is sought to achieve)

different syntactic realizations (e.g. light verb constructions or paraphrases):

EN: [...] ***to achieve the promotion*** *of respect* [...]
IT: [...] ***promuovere*** *il rispetto* [...]
(to promote the respect)

As already observed in [1], this is the most frequent type of translation shift and, we would rather say, the most evident, when comparing a tree pair.
The second main class often involves some of the structural shifts as well; despite this, we preferred to classify them separately. They may include:

addition–deletion:

EN: [...] *the respect for* ***and observance of*** *human rights and fundamental freedoms* [...]
FR: [...] *le respect* ***universel et effectif*** *des droits de l'homme et des libertées fon-*

classes, similarly to [3], with this expression we prefer to indicate only morpho-syntactic categories.

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

[9]Since in ParTUT translation direction is unknown, we consider the two transformation strategies as counterparts one of each other and put them in the same subclass. We applied the same principle even for the cases of additiondeletion, cited below.

damentales [...][10]
(the universal and effective respect of human rights and of fundamental freedoms)

mutation: whenever there is a textual equivalence (according to Catford's terminology), but the correspondence is characterised by a high degree of fuzziness.

EN: *the right to* ***recognition as a person before the law***
FR: *le droit à la* ***reconnaissance de sa personalité juridique***
(*the right to the recognition of his legal personality*)

Although to a different extent, both classes of shift (i.e. syntactic and semantic) were equally relevant, being the most prominent the structural shifts (with an overall frequency rate of 47.6%), the addition-deletion semantic sub-class (21%) and the category shifts (16.8%)[11]. The description of such shifts in quantitative terms, besides their classification, may provide an insight on the extent to which it is necessary to take into account such differences while designing an automatic alignment system, as well as on the impact that they may have on its overall performance.

4.1 Alignment proposals

Our proposal is based on some fundamental criteria which depend on the observation of the translation shifts. First of all, since in the shifts both morphology and syntax can be involved, we decided that we have to take into account at the same time a morphological and syntactic perspective, as allowed by a format encoding like TUT, with linguistic knowledge for the lexical morphology, dependency structure and argument structure. Observing the several guidelines (see [9] and [12]) and related works (e.g. [7], or [16]) consulted before we started the manual annotation of translational correspondences, we found only one study [15] which dealt with alignment of tree pairs in terms of phrases, rather than single words, and no one where the alignment was based on dependency relations. Useful information on the alignment of argument structure in parallel treebank were instead found in [3] and [6], but in all these guidelines there were significant discrepancies and divergences in the choice of what should actually be put into correspondence.

Our proposal includes two distinct steps, respectively referring to the lexical level and to syntactic dependencies. As for the alignment on the word level, we first used the WordAligner, a web-based interface which allows for manual editing and browsing of alignments and represents each pair of sentences as a grid of squares. For the syntactic level, we worked on an alignment procedure that could then be formalized and implemented benefitting from the syntactic information

[10] In this example, in particular, we observe both additions and deletions while comparing the English sentence to the French version.

[11] And among them, respectively, the shift between name and adjective, (36.3%), and between noun and verb (21.2%).

provided by the annotation. This procedure, as currently conceived, consists of two distinct steps.

Step 1: in a first stage, lexical correspondences are identified and stored by means of a probabilistic dictionary. We obtained such resource by running the same tool used for the sentence alignment, i.e. the Microsoft Bilingual Sentence Aligner (see Section 3.1), which contains an implementation of the IBM Model One. Since data gathered in ParTUT could not be sufficient to obtain reliable results, we ran the tool bidirectionally with texts retrieved from different parallel resource, i.e. Europarl and the Web Inventory of Transcribed and Translated Talks (WIT3)[12][2] (see Table 2 for details on the Italian-English pair).

Source	**sentences**	**tokens**
Europarl En	56,057	1,437,469
Europarl Ita	56,057	1,515,431
WIT3 En	9,266	176,345
WIT3 Ita	9,315	175,496

Table 2: Some preliminary results on the creation of a lexical resource for Italian and English with the IBM Model 1 implementation of the Microsoft Bilingual Sentence Aligner.

Step 2: starting from the lexical pairs obtained in the first step, correspondences between neighbouring nodes are verified by means of the information provided by the annotation (i.e. morpho-syntactic category and dependency relations). Such information is useful to predict alignment links between nodes that were not detected as translational equivalents by means of the dictionary, then completing the tentative alignment performed in the previous step.

The procedure, in its general approach (i.e. without considering more specific cases), is described below (see Algorithm 1). For each lexical pair between source and target sentences retrieved in the first step, referred to as *lex_pair(s,t)* in the pseudo-code, the algorithm iteratively searches for head and dependents of the source node *s* in the lexical pair and verifies, at first attempt, whether they belong to other lexical pairs; otherwise, it looks for their linguistic features, first Part of Speech (*PoS*), then syntactic relations (*dep_rel*), and compares them with the corresponding features of head and dependents of *t*.

The choice to use the relational and categorical information proved useful in the identification and resolution of some cases of translation shifts.

As for categorical shifts, for example, a common case is that of nominalization: this is easily detectable by means of the third step of the algorithm, since deverbal nouns are usually annotated as such in TUT exploiting the NOUN-OBJ relation (see Figure 1).

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

for all lex_pair(s,t) **do**
  if head and dependents of s are included in other lex_pair(s,t) **then**
    align nodes
  else if their PoS is the same **then**
    align nodes
  else if their dep_rel are the same **then**
    align nodes
  end if
end for

Algorithm 1: Node alignment procedure after the detection of lexical correspondences.

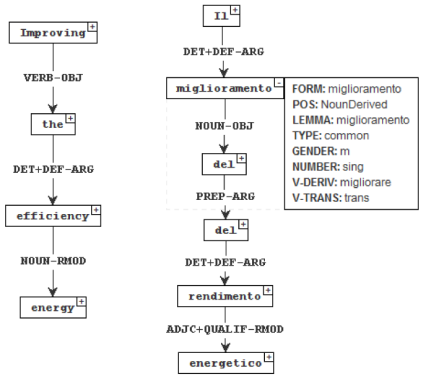

Figure 1: An example of nominalization (in the Italian version) and its annotation.

Even the other most common cases of shifts, i.e. those between adjective and name (whenever they are both modifier of the head node), are easily solved and formalized with the rule displayed above.

Also with regard to structural shifts, the systematic nature of some differences allowed the treatment of those cases with simple rules. For example, the conflation of modal, or a compound verb (especially in English), and the main verb in a single verb. The same applies to cases of change in word order, e.g. because of pre-and post-modification differences, the first being more common in English and the second in French and Italian, as briefly discussed in3.2 (see Figure 2 for an example). This is mainly due to the choice of a representation format that focuses on the dependency relation between a head and its modifier(s), rather than on constituents and their order.

For the same reason, other structural shifts involving word order were equally solvable, although they were less systematic and more subject to stylistic or individual translator's choice, as the following example shows.

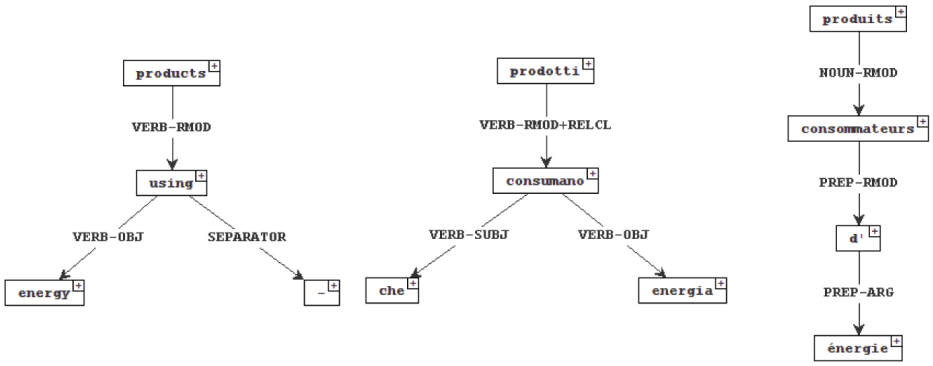

Figure 2: Representation of the expression "*energy-using products*" respectively in English, Italian and French. As observed, the heads in the sub-trees are the same, despite their positions in the three sentences.

EN: *The exchange of information on environmental life cycle performance and on the achievements of design solutions should be facilitated.*
IT: *Dovrebbero essere agevolati uno scambio di informazioni sull'analisi della prestazione ambientale del ciclo di vita e sulle realizzazioni di soluzioni di progettazione.*
(*should be facilitated an exchange of information on the analysis of the environmental life cycle perfomance and on the achievements of design solutions*.)

Even the cases of passivisation–depassivisation, due to the information encoded in the TUT format on the function of verbal arguments, may receive an alignment link: if lexical match is set in the first phase between the two verbs involved, the relational labels of their respective arguments (eg. [VERB-SUBJ] for the subject of the active form, and [VERB-OBJ/VERB-SUBJ] for the surface object of the passive form) were checked, and argument nodes are aligned.

These are all cases that show how and when linguistic knowledge and the linguistic information provided by the processing tools could manage to deal with systematic differences between two languages and with translation divergences.

5 Conclusion and future work

This paper aims to present a study for the design and development of a tool for the alignment of parallel data that exploits the linguistic information (especially on morpho-syntactic and relational level) explicitly annotated in a treebank. The purpose of this study was to investigate the usefulness of such information to overcome the limitations of alignment tools which are not linguistically motivated in

the treatment of typical translational divergences, or shifts. We described in what terms linguistic insights (such as the knowledge of some systematic differences between the language pairs involved), and the choice to use a representation format that focuses on the dependency relations and the predicative structure allows us to deal with some of these shifts by applying simple rules. Other aspects, however, are yet to be fully explored (although partially in progress for the time being), such as the creation of a reference alignment corpus, a systematic evaluation of the overall system and a comparison with the state-of-the-art tools, and the extension of the treebank to other text types in order to obtain a more balanced corpus on one hand, and to verify whether and to what extent the translation shifts classification here proposed, as well as the rules originally conceived for their automatic alignment are still valid and appropriately implemented.

References

[1] Catford, John C. (1965) *A Linguistic Theory of Translation: An Essay on Applied Linguistics*. Oxford: Oxford University Press.

[2] Cettolo, Mauro, Ghirardi, Federico and Federico Marcello (2012) WIT3: A Web Inventory of Transcribed Talks. In *Proceedings of the 16th EAMT Conference*, Trento, Italy.

[3] Cyrus, Lea (2006) Building a Resource for Studying Translation Shifts. In *Proceedings of Language Resources and Evaluation Conference* (LREC'06), Genova, Italy.

[4] Cyrus, Lea (2009) Old concepts, new ideas: approaches to translation shifts. In *MonTI. Monografías de Traducción e Interpretación.*, N. 1 (2009).

[5] Dyvik, Helge, Meurer, Paul, Rosén, Victoria and De Smedt, Koenraad (2009) Linguistically Motivated Parallel Parsebanks. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, Milan, Italy.

[6] Graça, João , Pardal, Joana Paolo, Coheur, Luísa and Caseiro, Diamantino (2008) Building a golden collection of multi-language word alignment. In *Proceedings of Language Resources and Evaluation* (LREC'08), Marrakech, Morocco.

[7] Hudson, Richard (1984) *Word Grammar*. Oxford: Blackwell.

[8] Lambert, Patrik, de Gispert, Adrià, Banchs, Rafael E. and Mariño, José B. (2005) Guidelines for word alignment evaluation and manual alignment. In *Language Resources and Evaluation*, December 2005, Volume 39, Issue 4.

[9] Lavie, Alon, Parlikar, Alok, and Ambati, Vamshi (2008) Syntax-driven Learning of Sub-sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation* (SSST-2), Columbus, OH.

[10] Melamed, Daniel (1998) *Manual annotation of translational equivalence: The BLINKER project. Technical report #98-07.* Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

[11] Menezes, Arul and Richardson, Stephen D. (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*, Toulouse, France.

[12] Moore, Robert C. (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*, Tiburon, California.

[13] Samuelsson, Yvonne, Volk, Martin, Gustafson–Capková, Sofia and Steiner, Elisabet Jönsson (2010) *Alignment Guidelines for SMULTRON*. Stockholm University, Department of Linguistics, University of Zürich, Institute of Computational Linguistics. Version 2.1, August 9, 2010.

[14] Simov, K., Osenova, P., Laskova, L., Savkov, A. and Kancheva, S. (2011) Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment. In *Proceedings of Recent Advances in Natural Language Processing* (RANLP), Hissar, Bulgaria.

[15] Steinberger, R. and Pouliquen, B. and Widiger, A. and Ignat, C. and Erjavec, T. and Tufiş, D. and Varga, D. (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of Language Resources and Evaluation Conference* (LREC'06), Genova, Italy.

[16] Zhechev, Ventislav and Way, Andy (2008) Automatic generation of parallel treebanks. In *22nd International Conference on Computational Linguistics* (COLING 2008), Manchester, UK.