Exploiting *catenae* in a parallel treebank alignment

Manuela Sanguinetti[1], Cristina Bosco[1], Loredana Cupi[2]
[1]Dipartimento di Informatica, [2] Dipartimento di Lingue e Letterature Straniere e Culture Moderne
Università di Torino (Italy)
{manuela.sanguinetti;cristina.bosco;loredana.cupi}@unito.it

Abstract

This paper aims to introduce the issues related to the syntactic alignment of a dependency-based multilingual parallel treebank, ParTUT. Our approach to the task starts from a lexical mapping and then attempts to expand it using dependency relations. In developing the system, however, we realized that the only dependency relations between the individual nodes were not sufficient to overcome some translation divergences, or shifts, especially in the absence of a direct lexical mapping and a different syntactic realization. For this purpose, we explored the use of a novel syntactic notion introduced in dependency theoretical framework, i.e. that of *catena* (Latin for "chain"), which is intended as a group of words that are continuous with respect to dominance. In relation to the task of aligning parallel dependency structures, catenae can be used to explain and identify those cases of one-to-many or many-to-many correspondences, typical of several translation shifts, that cannot be detected by means of direct word-based mappings or bare syntactic relations. The paper presented here describes the overall structure of the alignment system as it has been currently designed, how catenae are extracted from the parallel resource, and their potential relevance to the completion of tree alignment in ParTUT sentences.

Keywords: parallel dependency treebanks, syntactic catenae, alignment

1. Introduction

Several parallel treebanks have been developed in particular in the last few years. In order to make them useful for any further purpose, the data of these linguistic resources have to be properly aligned. The more challenging aspects in the alignment task is the treatment of all those translational divergences that in some contexts are also referred to with the term *shifts* (Catford, 1965; Cyrus, 2009). The well-known IBM translation models, for example, take into account complex aspects such as the word order and the probability that a source word aligns to more than one target word (Brown et al., 1993). There are, however, other aspects that word-based IBM models could not explain and for which syntactic information is needed.

Our work mainly focuses on the issues raised by the development of a newly created resource, namely a parallel dependency treebank for Italian, English and French – i.e. ParTUT – and on the potentiality of the structural information encoded in the resource for its alignment.

The choice to use dependency paradigm and its features for the automatic alignment is mainly dictated by the acknowledged fact that dependencies can better represent linguistic phenomena typical of morphologically rich and free-word order languages (see (Covington, 1990; Goldberg et al., 2013)); furthermore, dependencies show a higher degree of cohesion, compared to constituencies (Fox, 2002), and are closer to the representation of the predicate-argument structure, which is the linguistic level that we hypothesize that remains stable while shifting from one language to another. In addition, dependency formalism has also proved suitable for structural representation formats efficiently oriented to relation and information extraction tasks, such as the increasingly widespread Stanford typed dependencies (de Marneffe and Manning, 2008).

Neverthless, dependency formalism is typically based on relations between single nodes in the tree structure; when it comes to dealing with the alignment of such parallel structures, however, it is necessary to identify complete translational correspondences, which may involve whole phrasal units, instead of single nodes.

The above observations led us to assume that:

a) syntactic dependencies may play a crucial role in the alignment of divergent structures between parse tree pairs, especially when dealing with reordering issues;

b) to capture translational divergences that derive, for example, from conflation (or expansion) of lexical items, it is necessary to consider a syntactic (and alignment) unit that goes beyond the single node.

Previous works attempted to tackle this issues in particular in several ways:

- resorting to a full constituency-based representation, e.g. (Hearne et al., 2007)
- introducing a "phrasal" component when learning dependency structure mappings (Ding and Palmer, 2004)
- exploiting a further abstraction layer, such as a logical form (Menezes and Richardson, 2001) or the tectogrammatical layer of the Prague Dependency Treebank (Mareček et al., 2008)

For that purpose, we introduced in our approach the novel notion of *catena*, a syntactic unit that defines possible groups of nodes in a dependency tree that are linked together according to specific criteria (see Section 2. for a description).

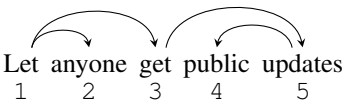
To summarize, assuming that the syntactic knowledge is crucial for the alignment of divergent structures, the current stage of our work consists in the creation of a syntactically-motivated alignment system that exploits information on dependencies, both in terms of single relations and of *catenae*, provided by ParTUT.

Therefore, the aim of this paper is manyfold: to provide a description of the current state of ParTUT,and to describe the method we are tuning up for its alignment bearing in

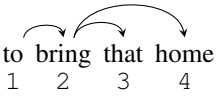
mind these premises and attempting to emphasize the potential of catenae and their exploitation.
In the next section, we then provide a brief theoretical background on the notion of catena, while Section 3. and Section 4. offer an overview of the resource and of the alignment process respectively. Section 4.3., in particular, provides a description of a core reference corpus consisting of a sample of manually aligned sentence pairs; this corpus has been created in order to have a better understanding of the implications of the use of catenae in the alignment process, and, at the same time, to create a gold standard for a preliminary evaluation of the automatic system, whose figures are finally reported in Section 4.4..

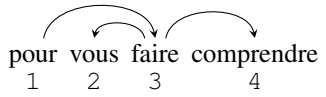
2. What is a *catena*?

Dependency is typically recognized for considering as syntactic unit the single words, which are in one-to-one relation with the node in the syntactic tree. This is its basic difference with respect to constituency, where, in contrast, the syntactic unit is the phrase, or constituent.
Recent years, however, have seen the recognition of a new unit type in dependency framework: that of *catena*. This notion is based on past work by O'Grady (1998), who used the term *chain* to designate a unit type for explaining the syntax of idioms; lately this notion has been extended to represent other linguistic phenomena, such as elliptical constructions (Osborne, 2005); in order to distinguish it from the previous notation, it then has been designated with the latin term catena (pl. *catenae*), (Osborne and Putnam, 2012).
A catena is defined as "*a word, or a combination of words which is continuous with respect to dominance*" (Osborne et al., 2012). This basic notion distinguishes catena from other units, such as strings or constituents. The figure below shows an example of a sentence represented in an unlabelled dependency graph where each word is assigned an identifier (`1, 2, 3, 4, 5`). In the sentence, 9 distinct catenae can be identified, besides single nodes[1] : `[1 2]`, `[1 3]`, `[4 5]`, `[1 2 3]`, `[1 3 5]`, `[3 4 5]`, `[1 2 3 5]`, `[1 3 4 5]`, and `[1 2 3 4 5]` (i.e. the whole dependency graph).



A catena may thus include both contiguous and non-contiguous sequences of words. Because of the dominance constraint, however, it cannot be compared neither to a string nor to a constituent. In the sentence we used as example, we can trace a continuous path from "*Let*" through "*get*" to "*updates*". Consider as a counter example the word combination "*Let anyone get public*", which is a string. We start with "*Let*" but we have no direct path to "*public*". In order to reach the former word, we would have to pass through "*update*", but this is not included in the word combination taken as example, therefore it cannot be considered a catena.
Catenae are also claimed to be more inclusive than constituents, as they do not require the unit to include all the nodes that are dominated.
As catenae can capture combinations of words consisting of a head and multiple dependents, they can behave as a "bridge" notion while dealing with divergences in translation; they can be used, in fact, to explain and properly identify those cases of one-to-many or many-to-many correspondences, typical of several translation shifts, such as the English idiomatic expression "*to bring that home*", reported in the example below together with its French counterpart "*pour vous faire comprendre* (*to let you understand*)[2].





This is the reason why we attempted to use them in the pipeline designed for the alignment of trees in ParTUT.

3. ParTUT

The resource used to experiment this approach is ParTUT, a parallel treebank for Italian, English and French. ParTUT has been designed as a multilingual development of an Italian existing treebank, i.e. the Turin University Treebank, or TUT, i.e. the reference treebank for the parsing evaluation campaigns for Italian Evalita[3]. The whole treebank currently comprises around 89,000 tokens, with an average of 1,060 sentences per language, and it includes different text genres, from debates of the European Parliament, instructions on how to create a Facebook account and multilingual transcriptions of talks from the TED Conferences to legal texts[4]. At this stage of the treebank development, we are also working on its extension with the addition of new parallel texts retrieved from EurLex[5], whose release is expected by the end of this year.
The treebank is annotated in a dependency-based formalism, partially inspired by the Word Grammar, in particular for what concerns the choice to represent determiners and prepositions as heads respectively of nominal and prepositional groups, a feature that is not shared by other dependency theories. Other typical features of TUT and ParTUT

[1] In accordance with the convention used in (Osborne et al., 2012), the words that form a catena are listed in a left-to-right order, following their linear order in the sentence.

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

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

[4] For further details on the composition of the collection and to download the annotated texts, see `http://www.di.unito.it/~tutreeb/partut.html`

[5] `http://eur-lex.europa.eu/en/index.htm`

trees are the use of null elements and the explicit representation of the predicate-argument structure not only for verbs but also for nouns and adjectives. This means, for example, that the arguments of a deverbal noun are annotated as arguments of its corresponding verb. The latter in particular is a notable feature that could prove useful whenever we have to compare parallel structures, as the explicit annotation of argumental roles in the source and target nodes could make the identification of translational correspondences easier; this is one of the main reasons why we have chosen this format for our experiments on alignment.

The treebank is automatically annotated with the broad-coverage parser included in the Turin University Linguistic Environment (TULE) (Lesmo, 2009), and then manually corrected.

Raw texts are also aligned at the sentence level with Microsoft Bilingual Sentence Aligner (Moore, 2002).

Despite its still reduced size, its content offers an overview of different text varieties and, at the same time, it mirrors some of the well-known distinctive features of the three languages involved[6], which constitutes a challenging factor in the alignment task as well.

The resource is available in several formats, among them the CoNLL, a widely used format for a variety of NLP tasks. For its interoperability and the opportunity to exploit it as a "pivot" format for the conversion to other ones[7], this is the format we have chosen for the development and testing of the alignment system.

4. Alignment method

To perform a tree-to-tree alignment means to find mappings between linguistically motivated analyses across languages. The ultimate goal of our work therefore consists in the development of an automatic system that fully exploits the syntactic information provided by these analyses to identify translational correspondences despite the presence of shifts. Our approach to the task starts from a lexical mapping and then attempts to exploit dependency relations, in a similar fashion to Menezes and Richardson (2001), Ozdowska (2005) and Ma et al. (2008). The whole alignment process is illustrated in Figure 1.

4.1. Extraction of *catenae* from the treebank

As shown in Figure 1, one of the preprocessing operations carried out before the alignment phase consists in the extraction of the possible catenae from parse trees of ParTUT. If we stick to the "standard" definition of catena shown in Section 2., a catena can also be formed, in principle, by a single word (then a node of the tree) or by the entire tree. Since we intend to make use of this concept for the identification of non-isomorphic sub-structures (and therefore neither a single word, nor the structure as a whole), we excluded these two cases in the extraction process. The script therefore proceeds recursively for each node by extracting all the possible catenae that may include it.

[6]For a partial description of the linguistic analysis performed on ParTUT texts, see (Sanguinetti et al., 2013)

[7]More recently, a tool has been made available for the conversion in the Stanford Dependencies, see (Bosco et al., 2013)

The output then returned is a list, for each node, of the possible catenae that can be formed with that node. Using the dependency graph of Section 2. as example, the set of possible catenae where the node "*get*", in position `3`, is involved is displayed as follows:

```
[1  3]          "let  get"
[1  3  5]       "let  get  updates"
[1  3  4  5]    "let  get  public  updates"
```

The variety of text genres in ParTUT is also mirrored by the different degree of structural complexity detected in different sub-corpora. One of the measures used to detect such complexity is the average length of catenae, shown in Table 1.

Corpus	avg. length
CC	7.43
EURO	7.02
FB	5.46
JRC	8.17
UDHR	5.98
WIT3	5.04

Table 1: Average length of syntactic *catenae* in ParTUT

The overall average length of catenae reported in the table has also been used to set a maximum length of catenae in the final alignment step, as described in the next section.

4.2. Alignment Steps

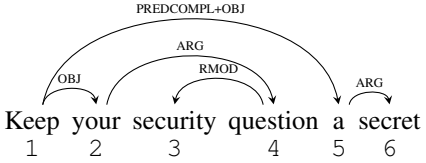
The whole alignment process entails multiple steps that can be summarized as follows:

Step 1: Lexical mapping the alignment starts from a basic lexical mapping that provides the initial anchor set of word pairs in the source and target trees.

The following subsections, that describe the multiple steps of the process, are also accompanied by alignment matrices that show the mapping output for the following English-Italian bisentence:

FB_En#48[8]: Keep your security question a secret
FB_It #45: Non rivelare a nessuno la domanda di sicurezza
(*Do not reveal to anyone the security question*)

whose dependency graph is shown below.



[8]The label refers to the position of the sentence in the given sub-corpus (in this case, we report sentence n. 48 of the English part of the Facebook section.)

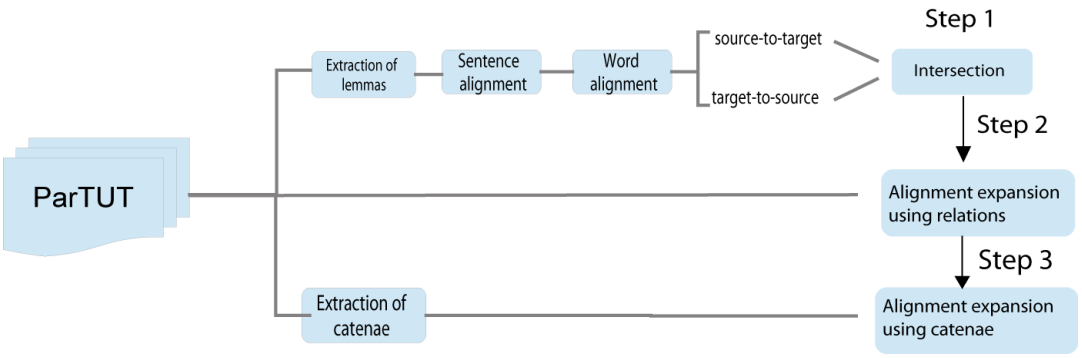
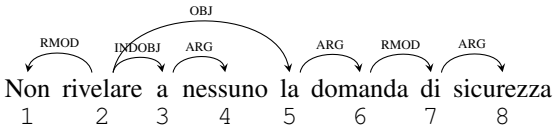
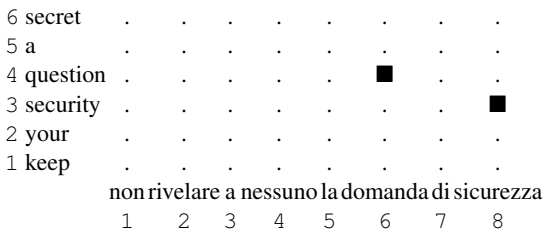


Figure 1: Alignment system pipeline.



Step 1: Intersection The anchor set of source-target word pairs in our case is obtained through several preprocessing stages. A first alignment is performed at the sentence level with the Microsoft Bilingual Sentence Aligner. Aligned sentences are then run bidirectionally (from source to target and from target to source) with the state-of-the-art tool for word alignment GIZA++[9] (Och and Ney, 2003). In order to obtain a high-precision set, the two alignments were then symmetrized with the Lingua-Alignment Set[10] and only the word pairs in the intersection set were retained. Considering the limited amount of data submitted to GIZA++, and the resulting risk to get too sparse data, we ran the tool on lemmatized rather than raw texts. Although not to a completely satisfying extent, this significantly increased the Precision score from 34.37 (raw data) to 63.95 . For generating GIZA++ alignment, we used the default parameters, therefore the alignments were bootstrapped from IBM Model 1 (iterations), HMM model (5 iterations), IBM3 (5 iterations) and IBM4 (5 iterations), as increasing the number of iterations did not lead to significant results in the final measures.

The alignment matrix below shows the output obtained in this step.



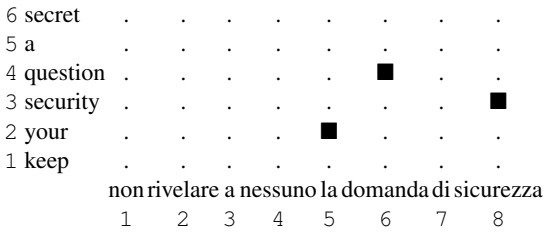
[9]As only 1:1 correspondences were retained, the number of sentence pairs processed by GIZA++ was further reduced to 855 (English-French), 822 (Italian-English) and 774 (French-Italian).

[10]http://gps-tsc.upc.es/veu/personal/lambert/software/AlignmentSet.html

Step 2: Alignment expansion using relations Starting from the lexical pairs obtained in the first step, correspondences between neighbouring nodes are verified comparing in parallel the respective relational structure.

The algorithm iteratively searches for head and dependents of the source node in the lexical pair and verifies, at first attempt, whether they belong to other lexical pairs; otherwise, it looks for their syntactic labels, and compares them with the corresponding labels of head and dependents of the target node.

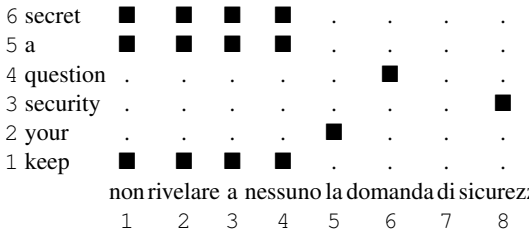
In the alignment example below, the anchor set is bootstrapped with the alignment of the English and Italian lexical items, respectively in position 2 ("*your*") and 5 ("*la*"), and playing the syntactic role of direct object (OBJ) of their corresponding root verbs.



This step mainly relies on the assumptions for which a dependency-based approach has been preferred in this work, i.e. that syntactic dependencies, together with the information on predicative structure, can be helpful in identifying some translational divergences, such as nominalizations (i.e. shifts into a name of linguistic elements belonging to another category) and passivizations (the shift from a passive to an active verb form). As remarked in Section 1., in fact, dependencies tend to hold between the languages, and provided that a common predicative structure is shared by the two parse trees, it will remain stable in the different languages despite variations in the realization of its arguments or their distance from the predicate.

There is, however, a number of shifts that cannot be captured and properly aligned with this system. These shifts include all those cases of non-compositional expressions, such as idioms, light-verb constructions, Multi-Word Expressions or paraphrases. For this purpose, we decided to explore a further step involving the use of catenae.

Step 3: Alignment expansion using catenae The group of nodes that are left unaligned are then compared in terms of catenae: by comparing the catenae extracted in both source and target files, such group of nodes are checked, and if they belong to the set of extracted catenae on both source and target sides, they are aligned. In this example, the catenae that are put in correspondence are those formed by the node sequences `[1 5 6]` ("*keep a secret*") and `[1 2 3 4]` ("*non rivelare a nessuno*").



Catenae are thus aligned as a whole in a typical multiple word-to-word fashion (each node of the source catena are aligned to each node of the target catena).
All nodes aligned in this step of the process are connected to their counterparts with a Possible link by default.

4.3. Reference corpus

Evaluation is the weakest part of the projects, as the system still needs major improvements. Even so, we defined a gold standard based evaluation methodology and we applied it to a small data set as a test. The use of a core "gold standard" in this study has a two-fold aim; firstly, it provides a reference material for the comparison of an alignment set, produced by a manual annotator, to the automatic tool that extracts catenae from ParTUT: such comparison may serve the purpose of showing whether and to what extent multiple and phrasal alignments can be considered as syntactic catenae, thus justifying the use of the latter in the automatic alignment as well. Secondly, it would help us in planning the development of the larger data set, that will be done in the next few months. In this section, we will therefore describe the main corpus charachteristics, and its degree of reliability, defined in terms of inter-annotator agreement.
We performed our experiments with the Italian–English subsection of ParTUT.
The sampled selection comprises 60 sentences from the different subcorpora of the entire treebank. The sample was manually aligned by two independent annotators, A_1 and A_2.
Being the most challenging aspects to be tackled by the alignment system, we selected the sentences so as to include cases of translation shifts that fall into at least one of the following categories[11] :

- nominalizations and other shifts in the morpho-syntactic category (i.e. from verb to noun, from noun to adjective, from adjective to adverb, etc.)
- passivizations or depassivizations (the shift from active to passive form or vice versa)
- different word order and discontinuous correspondences
- conflations (i.e. the translation of two words using a single word equivalent in meaning)
- paraphrases
- idioms
- additions/deletions (i.e. the introduction or elimination of pieces of information)
- mutation (whenever the correspondence is characterized by a high degree of fuzziness, or the content substantially differs between source and target version)

The inclusion of these specific cases has the aim of providing both a reference set of translation shift manual alignment for the further development of the gold standard, and a preliminary starting point for the evaluation of the automatic system and its performance in relation to these aspects.
For the manual annotation of alignment, we discussed several guidelines, both for word and treebank alignment (Graça et al., 2008; Kruijff-Korbayova et al., 2006; Lambert et al., 2005; Melamed, 1998; Samuelsson et al., 2010; Simov et al., 2011), in order to verifiy whether and how the linguistic phenomena mentioned above were treated in other similar tasks; a document was finally compiled explicitly dealing with such cases.
Furthermore, in order to evaluate the alignment annotation system, an inter-annotator agreement was calculated over the sample pairs using the Cohen's Kappa statistical measure (Carletta, 1996). For that purpose, and similarly to Macken (2010), each alignment link was classified based on the type of connection between the source and target word. Therefore, the scheme includes the following categories:

- ***Direct_S***: a one-to-one alignment that is linked as Sure
- ***Direct_P***: a one-to-one alignment that is linked as Possible
- ***Indirect_S***: a link that entails a one-to-many or many-to-many alignment and that can be considered as Sure
- ***Indirect_P***: a link that entails a one-to-many or many-to-many alignment and that can be considered as Possible

Contrarily to other (word) alignment tasks, we do not take into account NULL alignments (which is the link type used to label and classify unaligned words), neither in the manual nor in the automatic process; as a result, no label has been conceived for such cases in the alignment type classification. The inter-annotator agreement was then computed over the above mentioned categories, resulting in an overall $\kappa = 0.78$, which can be considered a good result in terms of reliability and consistency of the annotation system, however not to a completely satisfying extent. The

[11] For a description of the translation shift classification devised in our study, see (Sanguinetti and Bosco, 2012)

measure is in fact lowered by the relatively high disagreement on what should be discarded as translation correspondent, i.e. on those nodes that are left unlinked by each annotator. As it could impact the final evaluation process of the alignment system, this aspect should be thoroughly discussed in the further development of the gold standard. The overall fairly positive result, however, allow us to account for this resource for other considerations related to the relevance of multiple alignments within the corpus, as they typically express a shift in translation, and their correlation with catenae. In fact, a low discrepancy can be observed, both when comparing the annotations provided by A_1 and A_2, and when comparing them to the final gold standard, in relation to some data in particular. Such data are indicators of some aspects that we consider relevant to our study, and they include:

Mean Fertility: this measure is used to express the amount of words that have fertility bigger than one. As it represents a known difficulty for word alignment models, mean fertility serves as a good indicator of the difficulty of the corpus (Graça et al., 2008), especially when it comes to deal with multiple links.

Percentage of Multiple Alignments as Possible links: this indicator is designed to justify the choice, in the automatic system, to annotate as Possible any alignment link obtained with the addition of catenae.

Percentage of Multiple Alignments as Catenae: it shows the percentage of multiple alignments (both one-to-many and many-to-many alignments) that can be considered as syntactic catenae.
The data described here are reported in Table 2.

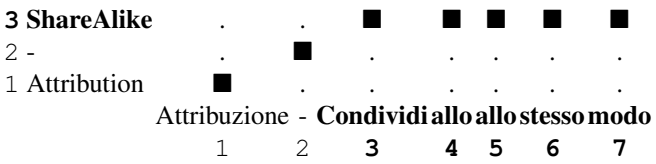
	A_1	A_2	Gold
Mean Fertility	1.3	1.3	1.2
% Multiple Alignments as P-links	66.1	61.6	61.2
% Multiple Alignments as Catenae	40.9	41.4	52.5

Table 2: Table that summarizes the results obtained for the three measures (Mean Fertility, Percentage of Multiple Alignments as Possible links, Percentage of Multiple Alignments as Catenae) in the annotator files (A_1 and A_2) and in the final gold corpus.

The latter measure in particular is a useful indicator of the degree of coverage, with the use of catenae, of the potential alignments that cannot be detected in Steps 1 and 2 described in Section 4..
The alignment matrix shown below reports an example of a manually aligned pair in the gold corpus, namely the bisentence:

CC_En#1: Attribution-ShareAlike 2.0
CC_It#1: Attribuzione-Condividi allo stesso modo 2.0 (ITALIA)

where the set of nodes that are also recognized as catena is highlighted in bold[12].



4.4. Preliminary Results

In order to assess the impact of the different alignment steps, a preliminary evaluation has been carried out, whose results are shown in Table 3. We compared the results obtained at each phase of the process to the alignment in the gold sample described in the previous section, attempting to assess its intrinsic quality by using Precision, Recall and F-measure (distinguished for Sure and Possible links). The figures in the table show that still major improvements have to done on the system; however, they also show interesting results that open to further discussions.

Step	**Ps**	**Rs**	**Fs**	**Pp**	**Rp**	**Fp**
Intersection	57.3	69.3	62.4	76.3	39.1	51.2
with relations	58.2	76.3	62.2	71.1	43.8	53.5
with catenae	63.9	76.3	67.9	62.2	57	56.6

Table 3: Results of each alignment step: Precision, Recall and F measure for Sure (Ps, Rs, Fs) and Possible (Pp, Rp, Fp) alignments.

Contrarily to the commonest expectations, and to past experiments, for example, Precision scores in Step 1 are lower than the scores obtained for the same measure in the next steps. The most interesting data observed, however, is that, the use of catenae contribute, although still not to a satisfying extent, to an improvement of the alignment system.

5. Conclusions and Future Work

In this paper, we described the ParTUT parallel treebank and the ongoing work on its alignment, in particular by introducing in the task the novel syntactic notion of catena, which has not been explored yet in other NLP applications. The aim of exploiting such notion is to find the set of links which most precisely detect translational divergences, also known as shifts, between the parallel trees. With this paper, we then attempted at showing the validity of our approach, for which full evaluation figures on aligned data will be provided in the near future. For the reliable evaluation of the aligner in particular, we are currently working on the improvement of performaces of the system and, in parallel, on the extension of the dataset, which can include a larger

[12]The raw sentences used for manual alignment have been tokenized using the same criteria adopted by the TULE parser for the automatic analysis of texts in ParTUT; such criteria provide for the splitting of contracted forms such as the one in the example, i.e. "allo", which is the result of a contraction between the preposition "a" (*to*) and the definite article "lo" (*the*); that is the reason why this item occurs twice in the example text.

variety of phenomena related to the alignment and possibly not occurring in ParTUT. The development of the new dataset is currently planned following the same steps of the development of the original ParTUT. Such extension will include new texts from legal domain.

One of the aims of ParTUT is also to allow its use and its comparison to other formats and paradigms of structural representation. The next planned direction for the development of ParTUT and its alignment system is therefore that of testing the aligner on a version of the treebank annotated according to the Stanford Dependencies, whose annotation format can be available for ParTUT as well by applying the conversion system used in Bosco et al. (2013).

6. References

Bosco, C., Sanguinetti, M., and Lesmo, L. (2012). The parallel-TUT: a multilingual and multiformat parallel treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 1932–1938.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting Italian treebanks: Towards an Italian Stanford Dependency treebank. In for Computational Linguistics, A., editor, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, jun.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254, June.

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

Covington, M. A. (1990). Parsing discontinuous constituents in dependency grammar. *Comput. Linguist.*, 16(4):234–236, December.

Cyrus, L. (2009). Old concepts, new ideas: approaches to translation shifts. *MonTI. Monografías de Traducción e Interpretación*, 1.

de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford Typed Dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ding, D. and Palmer, M. (2004). Automatic learning of parallel dependency treelet pairs. In *Proceedings of the 1rst International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 233–243.

Fox, H. J. (2002). Phrasal cohesion and statistical machine translation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 304–3111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Goldberg, Y., Marton, Y., Rehbein, I., and Versley, Y., editors. (2013). *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. Association for Computational Linguistics, Seattle, Washington, USA, October.

Graça, J., Pardal, J. P., Coheur, L., and Caseiro, D. (2008). Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Language Resources and Evaluation Conference (LREC'08)*, Marrakech.

Hearne, M., Tinsley, J., Zhechev, V., and Way, A. (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)*.

Kruijff-Korbayova, I., Chvatalova, K., and Postolache, O. (2006). Annotation guidelines for Czech-English word alignment. pages 1256–1261.

Lambert, P., de Gispert, A., Banchs, R. E., and Mario, J. B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4).

Lesmo, L. (2009). The Turin University Parser at Evalita 2009. In *Proceedings of Evalita '09, Reggio Emilia, Italy*.

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

Macken, L. (2010). An annotation scheme and gold standard for dutch-english word alignment. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Mareček, D., Žabokrtský, Z., and Novák, V. (2008). Automatic alignment of Czech and English deep syntactic dependency trees. In *Proceedings of EAMT08*, Hamburg, Germany, September.

Melamed, D. (1998). Manual annotation of translational equivalence: The blinker project. Technical report, University of Pennsylvania.

Menezes, A. and Richardson, S. 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 at ACL-2001*, pages 39–46.

Moore, R. 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 (AMTA-02)*, pages 135–144.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 1(29):19–51.

O'Grady, W. (1998). The syntax of idioms. *Natural Language and Linguistic Theory*, 16:279–312.

Osborne, T. and Putnam, M. (2012). Constructions are catenae: Construction grammar meets dependency grammar. *Cognitive Linguistics*, 23:165–215.

Osborne, T., Putnam, M., and Gross, T. (2012). Catenae:

Introducing a novel unit of syntactic analysis. *Syntax*, 15:354–396.

Osborne, T. (2005). Beyond the constituent: A dependency grammar analysis of chains. *Folia Linguistica*, 39:251–297.

Ozdowska, S. (2005). Using bilingual dependencies to align words in english/french parallel corpora. In *Proceedings of the ACL Student Research Workshop*, pages 127–132.

Samuelsson, Y., Volk, M., and Sofia Gustafson Capková, E. J. S. (2010). Alignment guidelines for SMULTRON. version 2.1, august 9, 2010. Technical report, Stockholm University, Department of Linguistics; University of Zürich, Institute of Computational Linguistics.

Sanguinetti, M. and Bosco, C. (2012). Translational divergences and their alignment in a parallel treebank. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT-11)*, pages 169–180.

Sanguinetti, M., Bosco, C., and Lesmo, L. (2013). Dependency and constituency in translation shift analysis. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 282–291, Prague, Czech Republic, August. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

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.