The Parallel-TUT: a multilingual and multiformat treebank

The Parallel-TUT: a multilingual and multiformat treebank

Cristina Bosco, Manuela Sanguinetti, Leonardo Lesmo

Dipartimento di Informatica, Università di Torino
Corso Svizzera, 195, 10149, Torino (Italy)
bosco,msanguin,lesmo@di.unito.it

Abstract

The paper introduces an ongoing project for the development of a parallel treebank for Italian, English and French, i.e. Parallel–TUT, or simply ParTUT. For the development of this resource, both the dependency and constituency-based formats of the Italian Turin University Treebank (TUT) have been applied to a preliminary dataset, which includes the whole text of the Universal Declaration of Human Rights, sentences from the JRC-Acquis Multilingual Parallel Corpus and the Creative Commons licence. The focus of the project is mainly on the quality of the annotation and the investigation of some issues related to the alignment of data that can be allowed by the TUT formats, also taking into account the availability of conversion tools for display data in standard ways, such as Tiger–XML and CoNLL formats. It is, in fact, our belief that increasing the portability of our treebank could give us the opportunity to access resources and tools provided by other research groups, especially at this stage of the project, where no particular tool – compatible with the TUT format – is available in order to tackle the alignment problems.

Keywords: parallel treebanks, annotation formats, alignment

1. Introduction

Parallel multilingual corpora can be considered as crucial resources in several tasks, e.g. Machine Translation (MT) and Computer–Assisted Translation (CAT), language learning and terminology extraction, but also projection of annotation on new (less-resourced) languages (Buch-Kromann, 2007; Ahrenberg et al., 2010). Their usefulness, as in the case of single language resources, increases when they are annotated and their annotations allow forms of alignment at various levels of linguistic knowledge, see e.g. (Ahrenberg et al., 2010; Grimes et al., 2010; Rios et al., 2009).

In particular, research in data-driven methods for MT has greatly benefitted from the increasing availability of parallel aligned treebanks for the training of statistical systems. But the development of such kind of resources raises several unresolved applicative and theoretical issues. First, as usual in the case of mono-lingual resources, parallel treebanks are usually semi-automatically developed by applying a very time-consuming and error prone process. Second, several levels of alignment of data, e.g. sentence, words or other syntactic components, can be in principle of some interest for the extraction of information relevant for translation and other tasks, but the development of tools for the alignment is currently limited to particular linguistic knowledge levels and annotation formats. Because of this, on the one hand, only a few of statistical MT models have only recently begun to really take advantage of higher level linguistic structures as annotated in treebanks; on the other hand, only a few parallel treebanks aligned at some level exist, while none of them is of sufficient use in any statistical MT application, see e.g. (Ahrenberg, 2007), (Volk et al., 2010), (Čmejrek et al., 2004) and (Megyesi et al., 2008).

This paper introduces the ongoing project of a new parallel treebank for Italian, English and French, henceforth Parallel–TUT (or, more simply, ParTUT) featured by both a pure dependency format (as described in (Sanguinetti and Bosco, 2011)) and a constituency-based annotation like that of the Penn Treebank (PTB), i.e. TUT–Penn. Even if the project concerns a resource missing for Italian, the development of a new treebank large enough for training of statistical systems is currently beyond our interest. The focus of the paper is therefore mainly on the features and quality of the annotation, and the investigation of some issues related to the alignment of data allowed by the formats applied in ParTUT. In fact, it will be described both the dependency-based annotation, called *native TUT*, and the conversion from this format to others useful in the cross-paradigm perspective and in order to increase the portability of data (e.g. Penn), or simply to make the data in native TUT compliant with different standards for displaying and analysis (e.g. TigerXML or CoNLL).

For the development of ParTUT, we applied to English and French the same tools designed for Italian and applied within the TUT project[1]. In particular, we used the parser TULE and the TUTtoPENNconverter[2], respectively for the application to the raw texts of the dependency-based annotation and the conversion of the resulting data, annotated in TUT, to the TUT–Penn format. On the one hand, the application of existing formats to other languages has been often reported in literature, see e.g. the application of the Prague Dependency Treebank (PDT) format to Arabic (Hajič and Zemánek, 2003), or the PTB format to Chinese[3] and Arabic[4]. This allowed in fact the improvement and extension in multi-lingual perspective of approaches originally developed for single languages, also increasing the portability of NLP tools and the availability of data useful for their com-

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

[2] http://www.di.unito.it/~tutreeb/-TUTtoPENNconverter/

[3] See http://www.cis.upenn.edu/~chinese/

[4] See http://www.ircs.upenn.edu/arabic/

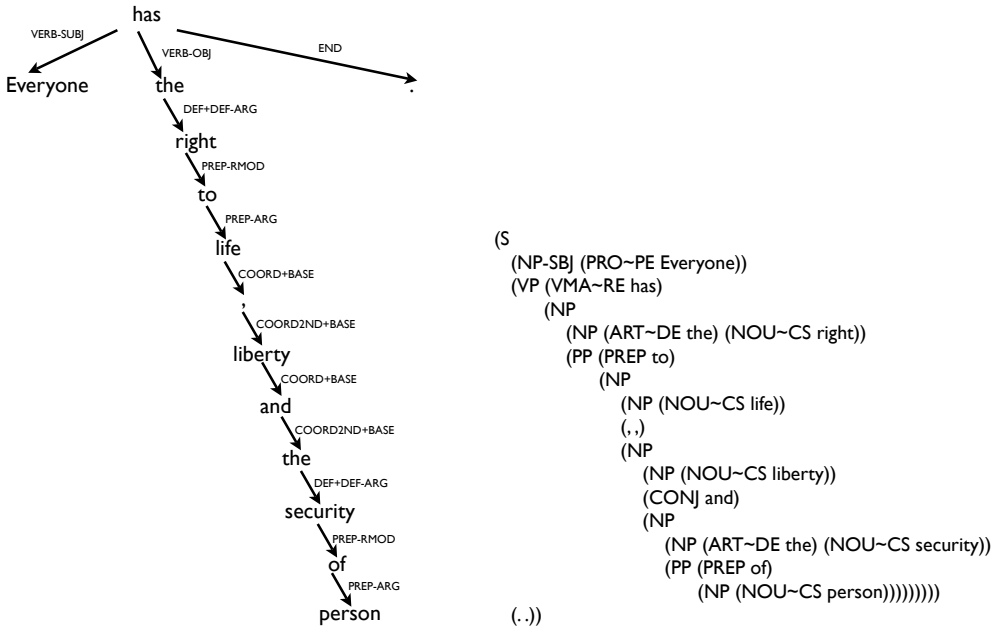

Figure 1: The English sentence HUMAN-RIGHTS-21, as annotated in TUT (a) and TUT–Penn (b).

parison and study. As suggested in (Paulussen and Macken, 2010), the use of the same annotating tools and formats for each monolingual corpus may also have a positive impact on the following exploitation and processing of the resulting parallel corpora. On the other hand, the availability of multi-format annotations for parallel treebanks, like that described in (Francom and Hulden, 2008), can be of some help in the analysis of the adequateness of specific format for particular languages and phenomena.

The next section describes the formats of the parallel treebank. The following section is instead devoted to the description of the data collected in order to build the corpus. The final section discusses issues related to the alignment of the parallel annotation of Italian, English and French allowed by TUT formats.

2. TUT formats

TUT is a resource developed by the Natural Language Processing group of the University of Turin (`http://www.di.unito.it/~tutreeb/`) which currently consists of more than 102,000 annotated tokens (around 3,500 sentences) extracted from texts varying from newspapers, to legal, to Wikipedia. The development of TUT includes two steps: the first one, which is devoted to the dependency-based native annotation of data, is the application of an annotation system to raw texts; the second, which outputs the data in a constituency-based format, consists in a conversion applied to the data in the dependency format produced by the first step. In the current phase of development, both steps require check and a limited amount of corrections which are applied in a semi-automatic way by exploiting tools intended for this purpose.

The core of the first step is the Turin University Linguistic Environment (henceforth TULE[5]) (Lesmo, 2007; Lesmo, 2009), which includes a rule-based parser developed in parallel with TUT and the modules needed for tokenization, PoS tagging and morphological analysis. The second step, which is the annotation in the Penn–TUT format, i.e. the constituency-based Penn-like format designed for TUT, includes the application of conversion tools (Bosco, 2007)[6] to the data in TUT native format.

2.1. Dependency: TUT native format

As far as the native annotation schema is concerned, a typical TUT tree (see Figure 1 (a)) shows a pure dependency format centered upon the notion of argument structure and is based on the principles of the *Word Grammar* theoretical framework (Hudson, 1984). This is mirrored, for instance, in the annotation of Determiners and Prepositions which are represented in TUT trees as complementizers of Nouns or Verbs. See, for instance, in Figure 1(a) the Determiner "the" which is the head for the Noun "security" and the Preposition "of" which is the head of the Noun "person".

For what concerns 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, which has proven essential for efficient processing of human language. Therefore, each relation label

[5] `http://www.tule.di.unito.it/`

[6] The conversion tools can be freely downloaded from the TUT web site.

can in principle include three components, i.e. morpho-syntactic, functional-syntactic and syntactic-semantic, but can be made more or less specialized, including from only one (i.e. the functional-syntactic) to three of them (see e.g. (Bosco and Lavelli, 2010; Alicante et al., 2012) for more details). For instance, the relation used for the annotation of the Prepositional modifiers in figure 1, i.e. PREP-RMOD-REASONCAUSE (which includes all the three components), can be reduced to PREP-RMOD (which includes only the first two components) or to RMOD (which includes only the functional-syntactic component). In figure 1 several relations involving two components are showed: e.g. VERB-SUBJ for the subject of a Verb, PREP-RMOD for the restrictive modifier introduced by a Preposition and PREP-ARG for the argument of a Preposition. This variable degree of specificity is a useful means for the human annotator in that it meets his/her different degree of confidence about a given relation. Moreover, it can also be applied in particular tasks in order to increase the comparability of TUT with other existing resources, by exploiting the amount of linguistic information more adequate for the comparison, e.g. in terms of number of relations.

Last but not least, as Italian requires, the TUT format provides an extended morphological tag set including all the categories and features needed to describe morphologically rich languages. This tag set allowed therefore for an accurate description both for French, whose morphological richness resembles that of Italian, and English, which is morphologically poorer.

Moreover, contrary to most of dependency-based annotations, in order to deal with pro–drop and equi, long distance dependencies and elliptical structures, the native TUT exploits also null elements. In most of cases, null elements are co–indexed with some word of the sentence (e.g. for gapping or equi phenomenon). Non co–indexed null elements are instead used e.g. for the representation of elliptical constructions, pro–drop subjects or other dropped complements playing some role in argument structure of Verbs. Exploiting null elements permits dependency trees to avoid crossing edges and to be projective. In practice, null elements are useful in giving an explicit representation also of those parts of the argument structure that could be missing, but sometimes crucial for some task. For instance, the exploitation of null elements can make the alignment easier, in all cases where the source language, e.g. Italian, allows the dropped subject and the target language does not, as English or French. Finally, as described in (Chung and Gildea, 2010), adding some empty elements can help building machine translation systems, which benefit from training on corpora with annotated empty elements, even when empty element prediction is slightly far from what would be conventionally considered robust.

2.2. Constituency: TUT–Penn format

As far as the constituency-based annotation is concerned, the annotation in TUT–Penn (see Figure 1 (b)) is structurally the same as in Penn Treebank, but it varies from this model because of a richer morphological tag set and an extended inventory of functional relations.

In fact, for what concerns morphology, the size of the Pos tag set of the TUT–Penn, if compared with that exploited in English PTB, clearly reflects the fact that Italian is morphologically richer than English, in particular with respect to the inflection of Verbs. Beyond the information that the PTB tag set makes explicit, TUT–Penn takes into account a richer variety of features for Verbs, Adjectives and Pronouns, apart from a few cases of English morphological features which do not exist (e.g. possessive ending) or do not correspond with Italian forms (e.g. comparative Adjective and Adverb).

Instead, for what concerns functional relations, in order to deal with phenomena related to the flexibility of Italian word order, some label has been added to the small PTB inventory. For instance, the label EXTPSBJ, which is used for the annotation of subjects in post-verbal position. 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 the duplication of Subjects which are positioned after Verbs.

3. Data and development of ParTUT

The parallel treebank currently comprises a preliminary set of sample texts, which have been annotated in order to assess our methodology. The corpus consists of 50 sentences extracted from the JRC-Acquis multilingual parallel corpus[7] (Steinberger et al., 2006) and the entire text (about 100 sentences) of the Universal Declaration of Human Rights[8]. More recently this preliminary set has been enlarged with an additional corpus extracted from the open licence "Creative Commons"[9] composed by around 100 sentences. All the data gathered in ParTUT up to the present (included raw texts) can be consulted and downloaded from the ParTUT web page[10]. These texts are represented in the ParTUT corpus in Italian, English and French and the exact amount both in terms of sentences and tokens can be seen in table 1[11].

The full corpus consists currently in less than 23,000 annotated tokens and represents only very specific text genres. The further development of ParTUT, planned for the future, includes the annotation of a larger set of data that will be collected by taking into account the issues related to the text genre too. It is in fact crucial to enlarge the corpus, in order to both address a larger and more meaningful set of linguistic phenomena, and more reliable analyses not affected by sparseness, like e.g. in (Ahrenberg, 2010).

Nevertheless, as deeply unbalanced the treebank might be at the moment, the choice of the texts of this collection was not fortuitous, and several criteria were considered before

[7] See `http://langtech.jrc.it/JRC-Acquis.html`, `http://optima.jrc.it/Acquis/`

[8] See `http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx`

[9] See `http://creativecommons.org/licenses/by-nc-sa/2.0`

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

[11] JRCAcquis indicates the JRC-Acquis multilingual subcorpus of ParTUT, UDHR indicates the Universal Declaration of Human Rights and CC the Creative Commons licence texts.

Corpus	sentences	tokens
JRCAcquis–It	50	2,205
JRCAcquis–Fr	52	2,297
JRCAcquis–En	50	1,895
UDHR–It	76	2,387
UDHR–Fr	77	2,537
UDHR–En	77	2,293
CC–It	96	3,141
CC–Fr	102	3,624
CC–En	88	2,507
total	688	22,886

Table 1: Corpus overview.

their selection: above all, practical reasons of easy availability from the web and the absence of Intellectual Property Rights problems, which allow us to process the data freely and release them under an open licence. Moreover, choosing texts from legal documents, we benefitted from the expertise in the field of legal language processing acquired within the TUT project by the group of the University of Turin. The data included in our corpus are representative of the development of unannotated parallel corpora developed in the last decades, in particular by the European Community. Finally, these texts includes raw materials which are in translation relation to each other and this should be relevant in the perspective of studies about human and machine translation.

The output produced by our annotation tool, however, was somehow affected by this bias, by virtue of the high number of long sentences[12], subordinate clauses, parentheticals and coordinated structures (constituting, by themselves, a well-known problem within automatic tools), which are all typical features of normative texts. Therefore, as stated above, and for the reasons we have just explained, we plan in the immediate future to extend the treebank so as to make our resource less biased and more complete for any further application and research.

For what concerns in particular the application of the annotation, although the TULE parser supports in principle linguistic analysis in several languages (English in particular, but also French, Spanish, Catalan and Hindi), its output quality currently achieves satisfactory results mostly for Italian, since it has been extensively tested in the development of the Italian TUT. That is to say, since TULE is a rule-based parser, it needs in the current phase of development of ParTUT rule-insertion and enrichment of the lexical knowledge for English and French, e.g. insertion of new lexical entries including, in particular, proper nouns, named entities, compounds and locutions, and new disambiguation rules for previously unseen linguistic phenomena.

Also the application of tools developed for the conversion of native TUT into TUT–Penn format has required some limited update of tables containing the linguistic knowledge exploited by tools for English and French. In general, we observed that applying to the ParTUT in native TUT format the tools for the conversion in TUT–Penn, Tiger–XML[13] (see an example in figure 2) and CoNLL has been a very useful practice for error detection and consequently quality improvement of the annotated data.

It is our belief that the availability of several formats, in particular those compliant to known standards, increases the portability of our treebank and could give us the opportunity to access resources and tools provided by other research groups, especially at this stage of the project, where no particular tool compatible with the TUT format is available in order to face some typical problems in parallel treebanking. This is particularly true for the alignment phase, which is currently one of the aspects to which we are focusing our attention and whose problems we attempt to describe in the next section.

4. Aligning ParTUT

Because of the correspondences between the information encoded in the same sentences in different languages, processing the same text in two languages yields useful information on how words and structures are translated from a source to a target language.

The ParTUT project is oriented to the development of a data set on which such hypothesis can be tested. ParTUT, assuming the annotation typical of TUT, features a rich annotation and it is oriented to the representation of the predicate-argument structure, a kind of information that we hypothesize that can be useful as a pivot for alignment in translation. As observed above, both the dependency core and the inventory of null elements introduced in the annotation schema of TUT contribute to a more accurate representation under this respect. Moreover, it makes available a set of data in different annotation formats, both dependency and constituency-based, that can allow for the comparison of alignment based on these two paradigms. This kind of comparison has been developed e.g. in (Gildea, 2004) showing that, for the Chinese-English case, constituent-based alignment significantly outperforms the dependency-based. Since ParTUT features formats belonging to two different paradigms and linguistic theories, it should be exploited as a testbed for similar comparison with reference to Italian, English and French.

Up to the present, the issues related to the alignment at sentence, word and syntactic level have been taken into account in ParTUT, but mainly by applying tools not specifically implemented for our formats and using empirical methods in order to develop guidelines that can drive the development of suitable tools. For instance, as for the sentence level, the alignment was performed with Omega Aligner[14], a simple Python script which produces files conforming to the Translation Memory eXchange (TMX) standard.

For the word alignment as well, we tested a number of freely available resources and took into account the useful suggestions proposed in several guidelines (as those in

[12] A high percentage of sentences reaches a length of 70 to 100 tokens.

[13] Texts and their relative encoding in the Tiger–XML format preserve the original dependency representation, in a similar fashion to what recommended by the Nordic Treebank Network (Hall and Nilsson, 2005).

[14] `http://www.omegat.org/en/resources.html`

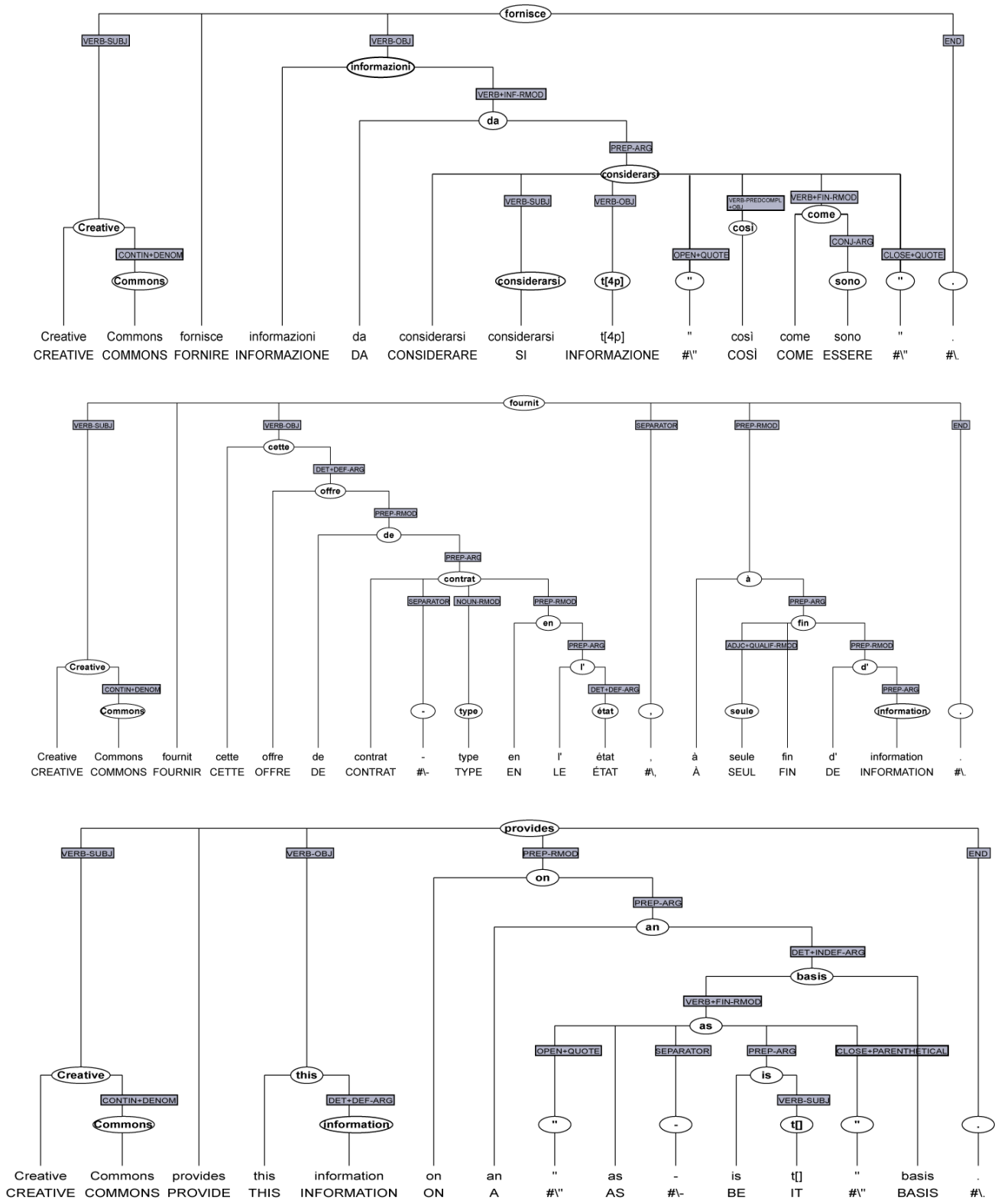

Figure 2: Three versions of the same sentence from the Creative Commons licence, represented in Tiger-XML.

(Melamed, 1998) or in (Lambert et al., 2005)), or in other similar works (see, for example, (Graça et al., 2008) or (Simov et al., 2011), just to name a few). In particular, we found a useful resource the WordAligner[15], a web-based interface which allows for manual editing and browsing of alignments. The tool represents each pair of sentences as a grid of squares (see Figure 3), which is a more useful representation device, if compared to other systems where alignments are drawn as lines, especially in cases of multiple alignment links.

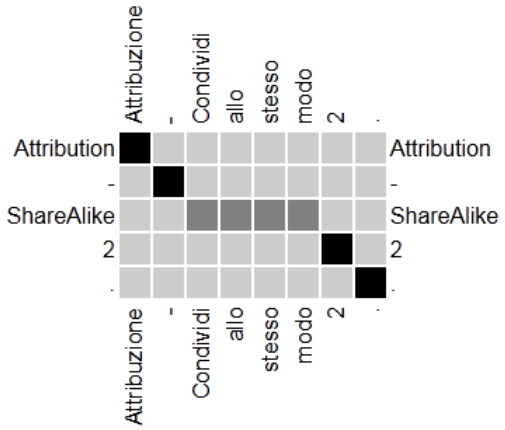

Figure 3: Example of a bi-sentence Italian-English aligned with WordAligner.

Furthermore, the WordAligner supports two types of alignment links, which are defined as sure and possible. According to our previous definition criteria, we opted for the notion of exact and fuzzy matches: the former is used to identify complete and minimal semantic translation units, and the latter to indicate valid translation pairs (including all those cases of translation shifts). Provided that the notion of sure and possible links do not differ from those we devised at the previous stage, for the sake of consistency, we decided to keep the terms of exact and fuzzy, while applying to these notions all those cases suggested by the literature respectively as sure and possible alignment links.

As pointed above, however, these two steps (sentence and word alignment) are but preliminary and totally experimental stages of a deeper level of alignment we are interested in: our goal is, in fact, to create a mapping between the tree pairs where information about the syntax-semantics interface is included. The major aim of our project for the development of ParTUT is at building a parallel treebank where alignment principles are not only lexically, but also syntactically motivated, and where the data mapped in the corpus can be of use in cross-linguistic research and applications, most notably in MT. This paper describes a first step in this direction, which consists in the creation of a golden collection of parallel parse trees where such alignment principles are investigated and tested mainly by hand.
Although we hypothesize that the features of the TUT annotation schemes can be of some help for the alignment, in particular at the syntactic level and with respect to the argument structure, these features and the richness of the annotation schema of ParTUT are currently the major limits in the application of standard alignment tools. The latter is, among the others, one of the reason why we decided to make available our resource in other exchange formats as well, such as Tiger–XML and CoNLL.

[15] http://www.bultreebank.bas.bg/aligner/-index.php

5. Conclusions

The paper describes an ongoing project for the development of a multilingual parallel aligned treebank, i.e. ParTUT, which features two annotation formats respectively based on the dependency (native TUT) and the constituency paradigm (TUT–Penn). The focus of the paper is therefore mainly on the features of these annotations and the methodology adopted for their application to the data included in ParTUT. In fact, the development of the resource is based on tools implemented for an existing treebank for Italian, namely TUT, which have been made adequate for English and French too.
Furthermore, preliminary issues related to the alignment of data allowed by the applied formats are presented, taking into account that the main goal of the ParTUT project consists in creating a mapping between the tree pairs where information about the syntax-semantics interface is included. Therefore, as for future development of this work, a number of issues must be further pursued, and in particular the development and the integration of suitable tools for alignment at syntactic level, which is currently missing.

6. References

L. Ahrenberg, J. Tiedemann, and M. Volk, editors. 2010. *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC) 2010*. Tartu.

L. Ahrenberg. 2007. LinEs: an English-Swedish Parallel Treebank. In *Proceedings of the 16th Nordic Conference on Computational Linguistics (NODALIDA '07)*, Tartu.

L. Ahrenberg. 2010. Clause restructuring in English-Swedish translation. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC) 2010*, Tartu.

A. Alicante, C. Bosco, A. Corazza, and A. Lavelli. 2012. A treebank-based study on the influence of italian word order on parsing performance. In *Proceedings of the Language Resources and Evaluation Conference (LREC'12)*, Istanbul.

C. Bosco and A. Lavelli. 2010. Annotation schema-oriented validation for dependency parsing evaluation. In *Proceedings of the 9th workshop on Treebanks and Linguistic Theories (TLT-9)*, Tartu.

C. Bosco. 2007. Multiple-step treebank conversion: from dependency to Penn format. In *Proceedings of the Linguistic Annotation Workshop (LAW) 2007*, Prague.

M. Buch-Kromann. 2007. Computing translation units and quantifying parallelism in parallel dependency treebanks. In *Proceedings of the Linguistic Annotation Workshop (LAW) 2007*, Prague.

T. Chung and D. Gildea. 2010. Effects of empty categories on machine translation. In *Proceedings of Empirical Methods in Natural Language Processing - EMNLP'10*, Boston.

J. Francom and M. Hulden. 2008. Parallel multi-theory annotation of syntactic structure. In *Proceedings of the Language Resources and Evaluation Conference (LREC'08)*, Marrakech.

D. Gildea. 2004. Dependencies vs constituents for tree-based alignment. In *Proceedings of Empirical Methods in Natural Language Processing - EMNLP'04*, Barcelona.

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

S. Grimes, X. Li, A. Bies, S. Kulick, X. Ma, and S. Strassel. 2010. Creating arabic-english parallel word-aligned treebank corpora at LDC. In *Proceedings of Language Resources and Evaluation Conference (LREC'10)*, Malta.

J. Hajič and P. Zemánek. 2003. Prague Arabic Dependency Treebank: Development in data and tools. In *Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools*, Cairo.

J. Hall and J. Nilsson. 2005. Converting Dependency Treebanks to MALT-XML. Technical report, Väjö University, School of Mathematics and Engeneering.

R. Hudson. 1984. *Word grammar*. Basil Blackwell, Oxford and New York.

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

L. Lesmo. 2007. The rule-based parser of the NLP group of the University of Torino. *Intelligenza artificiale*, 2(IV).

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

B. Megyesi, B. Dahlqvist, E. Pettersson, and J. Nivre. 2008. Swedish-Turkish Parallel Treebank. In *Proceedings of Language Resources and Evaluation Conference (LREC'08)*, Marrakech.

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

H. Paulussen and L. Macken. 2010. Annotating the Dutch Parallel Corpus. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, Tartu.

A. Rios, A. Göhring, and M. Volk. 2009. A Quechua–Spanish parallel treebank. In *Proceedings of 7th Workshop on Treebanks and Linguistic Theories (TLT-7)*, Groningen.

M. Sanguinetti and C. Bosco. 2011. Building the multilingual TUT parallel treebank. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC) 2011*, Hissar.

K. Simov, P. Osenova, L. Laskova, A. Savkov, and S. Kancheva. 2011. Bulgarian-english parallel treebank: Word and semantic level alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria.

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

M. Čmejrek, J. Hajič, and V. Kuboň. 2004. Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of EAMT 10th Annual Conference*, Budapest.

M. Volk, A. Göhring, T. Marek, and Y. Samuelsson. 2010. SMULTRON (version 3.0) - The Stockholm MULtilingual parallel TReebank. An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments.