Building the multilingual TUT parallel treebank

Manuela Sanguinetti
Università di Torino
manuela.sanguinetti@studenti.unito.it

Cristina Bosco
Dipartimento di Informatica,
Università di Torino
bosco@di.unito.it

Abstract

The paper introduces an ongoing project for the development of a parallel treebank for Italian, English and French annotated in the pure dependency format of the Turin University Treebank, i.e. Parallel–TUT. We hypothesize that the major features of this annotation format can be of some help in addressing the typical issues related to parallel corpora, e.g. alignment at various levels. Therefore, benefitting from the tools previously used for TUT, we applied the TUT format to a multilingual sample set of sentences from the JRC-Acquis Multilingual Parallel Corpus and the whole text of the Universal Declaration of Human Rights.

1 Introduction

Parallel corpora are currently considered among the crucial resources both for a variety of NLP tasks, e.g. machine translation and cross-lingual information extraction, and for research in the field of translation studies and contrastive linguistics with respect to terminology and syntax in particular.

Since the utility of parallel corpora is increased by forms of annotation which make explicit the linguistic knowledge involved in the raw data, parallel treebanks have proved to be valuable resources for a number of purposes (see e.g. (Ahrenberg et al., 2010; Grimes et al., 2010; Rios et al., 2009)). As far as translation studies are concerned, the FuSe project (Cyrus, 2006), for example, aims at studying translation shifts in an English-German corpus annotated with regard to the predicate-argument structure, while the LinEs parallel treebank for Swedish and English (Ahrenberg, 2007) focuses on this aspect by means of complete alignments of segment pairs. As for contributes to the improvement of machine translation quality (both rule-based and statistical), a few examples are provided by SMULTRON (Volk et al., 2010), with a constituency-based parallel treebank for English, German and Swedish; the Prague Czech-English Dependecy Treebank (Čmejrek et al., 2004); the Copenhagen Dependency Treebank[1] for Danish, English, German, Italian and Spanish; and the Swedish-Turkish Parallel Treebank (Megyesi et al., 2008).

In this paper, we introduce a new parallel treebank for Italian, English and French, henceforth Parallel–TUT. The annotation schema for this new resource is that of the Turin University Treebank (TUT), which has been applied in a dependency-based treebank used for training of parsing systems and as reference for the evaluation campaigns for Italian parsing. By featuring a rich set of grammatical relations, it shows a representation centered on the predicate-argument structure, a linguistic knowledge that is proximate to semantics and underlies syntax and morphology, essential for the efficient processing of human language. We developed our project also in order to test the hypothesis that this kind of knowledge, and thus the schema representing it, can be useful also in bridging the differences among languages, e.g. in translation.

Therefore, as far as the annotation of the Parallel–TUT corpus is concerned, our approach consists in extending and applying the same tools designed for Italian, within the TUT project, to two other languages, i.e. English and French. The result is the extension of the same format and relations for all the languages of the new parallel corpora, with the same granularity in the representation of the linguistic knowledge. On the one hand, this is motivated by the fact that, as suggested in (Paulussen and Macken, 2010), the use of

[1] http://code.google.com/p/copenhagen-dependency-treebank/

Proceedings of Recent Advances in Natural Language Processing, pages 19–28,
Hissar, Bulgaria, 12-14 September 2011.

different annotating tools and formats for each monolingual corpus may have a negative impact on the following exploitation and processing of corpora, such as alignment at various levels. On the other hand, the literature shows several examples of application to different languages of formats originally developed for a given language, by using the same features of the native format to address new linguistic phenomena encountered in the other languages. For instance, the format of the Prague Dependency Treebank (PDT), developed for Czech, has been afterwards applied to Arabic (Hajič and Zemánek, 2003), or the Penn Treebank format, which has been applied e.g. to Chinese[2] and Arabic[3]. An especially relevant side effect of the application of such kind of methodology consists in increasing the portability across languages of NLP tools and in making available data useful for the comparison and study of models and strategies underlying NLP tools when applied to different languages.

The work presented here aims at going beyond the creation of a parallel treebank where Italian language is included. It aims, in fact, at extending and applying a single treebank schema to other languages, and study how the schema can be meaningfully used to address issues typically related to parallel corpora, e.g. alignment at various levels. The focus of this work is therefore the format of the treebank and the consequence of the application of this format on a parallel corpus.

The remainder of the paper is structured as follows. The next section describes the TUT annotation schema while Section 3 shows the content and size of the corpus on which the schema has been applied. Section 4 describes the annotation process for the three monolingual corpora, while Section 5 shows the alignment issues related to the effects of applying the TUT format to English and French. Finally, we discuss the current state of the project, analyze the future developments of Parallel–TUT and briefly summarize the project.

2 The Turin University Treebank: the resource and its annotation schema

TUT is a resource developed in the last ten years by the Natural Language Processing group of the University of Turin (`http://www.di.unito.it/~tutreeb/`). It currently consists in more than 102,000 annotated tokens (around 3,500 sentences).

The treebank annotation is automatically performed by the Turin University Linguistic Environment (henceforth TULE[4]) (Lesmo et al., 2002; Lesmo, 2007; Lesmo, 2009) and then semi-automatically checked in order to recover errors in the morphological and syntactic annotation. TULE is a rule-based dependency parsing system which includes also the modules needed for tokenization, PoS tagging and morphological analysis, as well as parsing. The parsing module produces a projective dependency tree for each given sentence in input. In the last evaluation campaign for Italian parsing, held in 2009 (Bosco et al., 2009b), TULE achieved the best scores currently at the state of the art (Labelled Attachment Score 88.73), which are very close to the scores known for English parsing.

The core of the treebank is a dependency-based annotation scheme (on which we will focus in this paper), but the resource has been also enriched by the converted versions of all the annotated data in a Penn-like format (Bosco, 2007), in a Combinatory Categorial Grammar format (Bos et al., 2009)[5] and in other constituency-based annotations. This results both in an increased quality of the annotated material and portability of the resource. Beyond allowing the training of parsing systems, TUT has been used as a testbed for evaluation campaigns (Bosco et al., 2007; Bosco et al., 2009a; Bosco et al., 2009b) and analyses of parsing models' performance with respect to variation in tag sets, paradigms and annotation schemes (Bosco and Lavelli, 2010).

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 applies the major 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. For instance, in figure 1 the tree for the sentence NEWS-355 from TUT, i.e. *"L'accordo si è spezzato per tre motivi principali"* (The agreement has been broken for three main reasons)[6], shows the features of the an-

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

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

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

[5] `http://www.di.unito.it/~tutreeb/CCG-TUT/`

[6] English translations of the Italian examples are literal

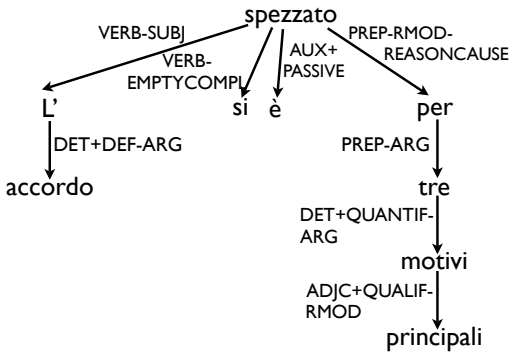

Figure 1: Sentence NEWS–355 of TUT.

notation schema. In particular, we see the role of complementizer played by Determiners (i.e. the article *"L'"* (The) and the numeral *"tre"* (three)) and Prepositions (i.e. *"per* (for)).

By contrast, the native TUT scheme exploits also some representational tools which are non–standard in dependency-based annotations, i.e. null elements, in order to deal with particular structures. In particular, null elements are used for pro–drops and missing subject (e.g. equi), long distance dependencies and elliptical structures. These phenomena are quite common in Italian, a morphologically rich language where verbal inflection leads to a widespread diffusion of the pro–drop phenomenon and to a relatively free order of words and constituents. For instance, the subject deletion is very common with tensed verbs in declarative clauses, as confirmed by the data in TUT corpora, where this phenomenon occurs an average of 0.28 times per sentence[7].

On the one hand, an advantage in using null elements in the annotation is that they permit dependency trees to be without crossing edges and projective structures also for non–projective sentences. On the other hand, by using null elements it is possible to give an explicit representation also of parts of the argument structure that can be missing, but crucial for some task. For instance, in machine translation, if the source language allows argument deletion and the target language does not, in order to make possible for the system to handle the translation, it is crucial that in the source language the dropped argument is explicitly marked. An alike situation can happen in a translation from Italian to English or French, where, on the contrary, the subject is always lexically realized in tensed clauses.

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 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) 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). 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.

Observing related works, we think that the TUT schema can be a good candidate for the development of a parallel treebank for various reasons. First of all, it is oriented to the representation of the predicate-argument structure, a kind of information that can be useful as a pivot for

and s, they thus may appear awkward in English.

[7] But the frequency of pro–drop varies from 0.17 to 0.64 times per sentence according to the text genre included in the treebank.

the alignment in translation, but is also crucial in tasks such as Information Extraction. 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. Second, this schema gives the means for the development of annotations at various degrees of specificity of grammatical relations, thus extending the comparability and compatibility with other existing resources. Finally, another aspect to be taken into account is the availability of automatic tools for the conversion of the native TUT format in other constituency-based representations, among which the most known and used format in the world (i.e. that of the Penn Treebank), and in a Combinatory Categorial Grammar format too, which is a semantic-oriented representation.
In the next sections we describe the parallel corpus on which we have applied the TUT format for the development of the Parallel-TUT.

3 The data in the Parallel–TUT

The Parallel–TUT currently comprises a small set of sample texts, which have been annotated in order to assess our methodology and test our hypothesis. They are organized in two sub-corpora, as outlined in Table 1.

The first sub-corpus consists of about 50 sentences extracted from the JRC-Acquis multilingual parallel corpus[8] (Steinberger et al., 2006) for each of the three languages involved in the Parallel–TUT. In particular, the sentences for Italian are shared by TUT and the corpus used within the French parsing evaluation campaign Passage[9], respectively in Italian version annotated in the TUT format, and in French version annotated in the EASy format. The English counterpart of the corpus was retrieved from English section of the JRC-Acquis corpus. We will refer to these data as JRCAcquis–ITA, JRCAcquis–FR and JRCAcquis–EN, respectively for Italian, French and English.
The second sub-corpus, which will be referred as UDHR–ITA, UDHR–FR and UDHR–EN, includes the entire text of the Universal Declaration of Human Rights, as available in the official Web page of the UN Office of the High Commissioner of Human Rights[10], and consists of about 76 sentences for each language.

Corpus	sentences	tokens
JRCAcquis–ITA	50	2,205
JRCAcquis–FR	52	2,297
JRCAcquis–EN	50	1,895
UDHR–ITA	76	2,387
UDHR–FR	77	2,537
UDHR–EN	77	2,293
total	382	13,614

Table 1: Corpus overview.

For what concerns the texts of the JRCAcquis corpus in particular, they were selected because of their availability in two different annotation formats developed by two independent research groups, as mentioned above. Moreover, choosing texts from legal documents, we benefitted from the expertise in the field of legal language processing acquired within the TUT project[11]. Last but not least, the data included in our corpus are representative of the development of raw text parallel corpora developed in the last decades, e.g. from the European Community. Nevertheless, we know that analyses based on such kind of unbalanced material may lead to misleading results if applied in general context, as the syntax in this corpus is typical of a quite particular kind of documents. This will be taken into account in the further development of our corpus.
In general, our selection of texts includes raw materials which are in translation relation to each other, and free of Intellectual Property Rights problems, which allows us to release treebank data under an open license.

4 Treebank Development

Except for the Italian part of the first sub-corpus of the Parallel–TUT, i.e. JRCAcquis–ITA (which was already available in the annotated version[12] as described above), for the English and French counterparts, as well as for the entire second

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

[9] `http://atoll.inria.fr/passage/index.en.html`.

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

[11] Around the 30% of TUT data are extracted from legal texts, i.e. the *Codice Civile* and the *Costituzione Italiana.*

[12] Available from the TUT Web page at `http://www.di.unito.it/~tutreeb/` (EUDIR Section)

sub-corpus (UDHR), we processed the texts following the same strategies applied in the TUT project and using the same tools both for parsing and checking.
Being the original materials in XML format (eg. texts collected in the JRC multilingual corpus) or directly exctracted from a Web page, the first step was to clean up files from noisy data (eg. markups) and to convert them to plain text files with UTF-8 encoding. In this way, texts can be exploited for our further linguistic analyses.
Despite other parallel treebanks, where monolingual corpora were processed independently with different tools (cf. (Megyesi et al., 2008)), or created from already existing monolingual treebanks (cf.(Klyueva and Mareček, 2010)), the texts of our collection were analyzed from scratch with the same tool, i.e. TULE. Although TULE supports in principle linguistic analysis in several languages (English in particular, but also French, Spanish, Catalan and Hindi), its output quality achieves satisfactory results mostly for Italian, since it has been extensively tested in the development of the Italian treebank TUT. Since TULE is a rule-based parser, the annotation phase for English and French therefore entailed alternating steps of rules insertion in TULE and automatic analysis, until an acceptable output was produced. Rule-insertion steps included mainly the enrichment of lexical knowledge, e.g. insertion of new lexical entries (including proper nouns, named entities, compounds and locutions), modifications in the suffix tables, and new disambiguation rules for linguistic phenomena previously unseen in Italian. A typical example of such phenomena is the English genitive for regular plural nouns (*-s'*). Since in Italian (and French too) the apostrophe is normally considered a graphic sign indicating an elision, during the automatic analysis, tokenization in particular, it is kept attached to the previous token. The English possessive case, however, is normally isolated and treated as a single token. Its recognition in this form by the TULE tokenizer has therefore requested the integration of a new condition in the set of disambiguation rules. Other types of intervention focused on the syntactic representation of those phenomena that distinguish the two languages from Italian. For example, the French superlatives formed by the definite article and *plus/moins* follow a word order which is quite different from that of the Italian superlatives: it was therefore necessary to modify the representation scheme already present in the TUT annotation guidelines for Italian. The treatment of the expletive subject (ie. a purely syntactic subject, not semantically realized), which is a common occurrence both in English and French, but not in Italian (where, as we said, the subject can be omitted) also required the inclusion of additional labels in the annotation schema.
The whole procedure above described had a twofold goal: to improve the output quality of TULE for English and French, and, as a result, to reduce to a feasible extent manual intervention of human annotators in future annotation work.
Because of the current small size of the corpus and the consequently limited training on English and French of our tools, we expect that a considerable amount of manual intervention (eg. enriching the knowledge base of the parsing system) will be necessary also in the next step of the development of our parallel treebank. In fact, the variety of new syntactic structures encountered so far in English and French data is quite small, and the probability that the treebank could miss some syntactic phenomena is high.

The relatively lower quality of the output of TULE for English and French with respect to Italian (as reported in Section 6) made the final stage of manual correction crucial to verify that linguistic phenomena were annotated appropriately and consistently. In this stage, the same tools used in the development of TUT were exploited. For instance, for displaying the dependency trees, the viewerTULETUT Java graphical interface was used, thus allowing the observation of the structures in a more readable graphic form.

It is known that the conversion of dependency trees into phrase structures is in itself a comparative test of the adequateness of the involved representation formats with reference to the features of the language and the quality and consistency of annotation (Musillo and Sima'an, 2002). Therefore, some preliminary experiment was also performed by applying to the English and French data the procedures for the conversion in the Penn Treebank format developed for Italian. The results are promising in particular for English, as we expected, since this is the reference language for the

Penn format. For French the conversion should be further refined by including in the Penn format the representation of particular phenomena.
As far as the annotation phase of the Parallel–TUT is concerned, it can be currently considered as concluded and the corpus will be soon released and made available for research purpose. In the next section, we describe the alignment phase which is the less advanced part of the project, currently under development.

5 Aligning the Parallel–TUT

Several techniques have been developed and made available for aligning texts at various granularities. They vary from document-structure to sentence, word, phrases or dependency subtrees (see e.g. (Wu, 2010; Li et al., 2010)).
Each level implies several and different issues that are currently in part unresolved also because does not exist an objective and universally shared notion of correspondence between sentence units. For instance, it is difficult to decide which words in a given target string correspond to which words in its source string (especially where idiomatic expressions are involved) and often, an alignment includes effects such as reorderings, omissions, insertions (Och and Ney, 2003).
Moreover, tools implementing aligment techniques are often designed with reference to some particular kind of annotation and schema, and cannot be applied to different formats, such as TUT. This is currently the major limit of the project that should be addressed in the next future. In fact, even if in our project we are interested in the alignment at various levels, we applied until now only some preliminary form of alignment, and the most of the time devoted to this part of the Parallel–TUT project has been spent in the analysis and report of the issues raised by our data.
First of all, the Parallel-TUT has been developed taking into account the issues related to the alignment at sentence and word level. Therefore, after the linguistic annotation, a further step has been the detection of lexical and structural correspondences between language pairs. As for the sentence level, the alignment was performed with Omega Aligner[13], a simple Python script used for the alignment of translation units within Computer Aided Translation (CAT) systems. The files produced conform to the Translation Memory eXchange (TMX) standard, an XML-compliant formatting standard normally used for storing and exchanging translation memories among CAT systems. Since the script expects the same number of segments in the source and target texts, some pre-processing was required, in order to avoid mismatches, in particular for punctuation marks.
As for the word alignment, considering the current absence of a tool which was compatible with the TUT format, the process was performed only preliminarily, using empirical methods, mainly in order to develop alignment guidelines that can drive the development of a tool suitable for such a task in the future. We observed that the alignment is made easier by the fact that languages are annotated using the same format, and because of TUT format strategy for the annotation of idiomatic expressions or compound words, which consists in splitting them in one line for each lexical word. In order to keep alignments as fine-grained as possible, two link types were designed to capture linguistic correspondences: exact and fuzzy. 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). However, untranslated words, incorrect or deeply divergent translations are left unaligned.
At the same time, we chose to link correspondences at a structural level too, so that parallelisms between pairs of syntactic trees (or subtrees) could be easily detected and studied. In recent years, in fact, a number of syntactically motivated approaches to statistical machine translation have been proposed which focused on the fact that syntactic constituents tend to move as units with systematic differences in the word order of the languages involved (Zhang and Gildea, 2004). In the case of Parallel–TUT, a syntactically motivated alignment may be driven by the argument structure as annotated according to the TUT format. In particular, we planned to implement forms of alignment based on (a selection of the major) grammatical relations that are involved in the predicate argument structure, as figure 2 shows. We hypothesize that the features of the annotation schema of TUT can be of some help for the alignment at this granularity. Nevertheless, these features and the richness of the annotation schema of TUT are currently the major limits in the application of a standard tool for the alignment of the Parallel-TUT.

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

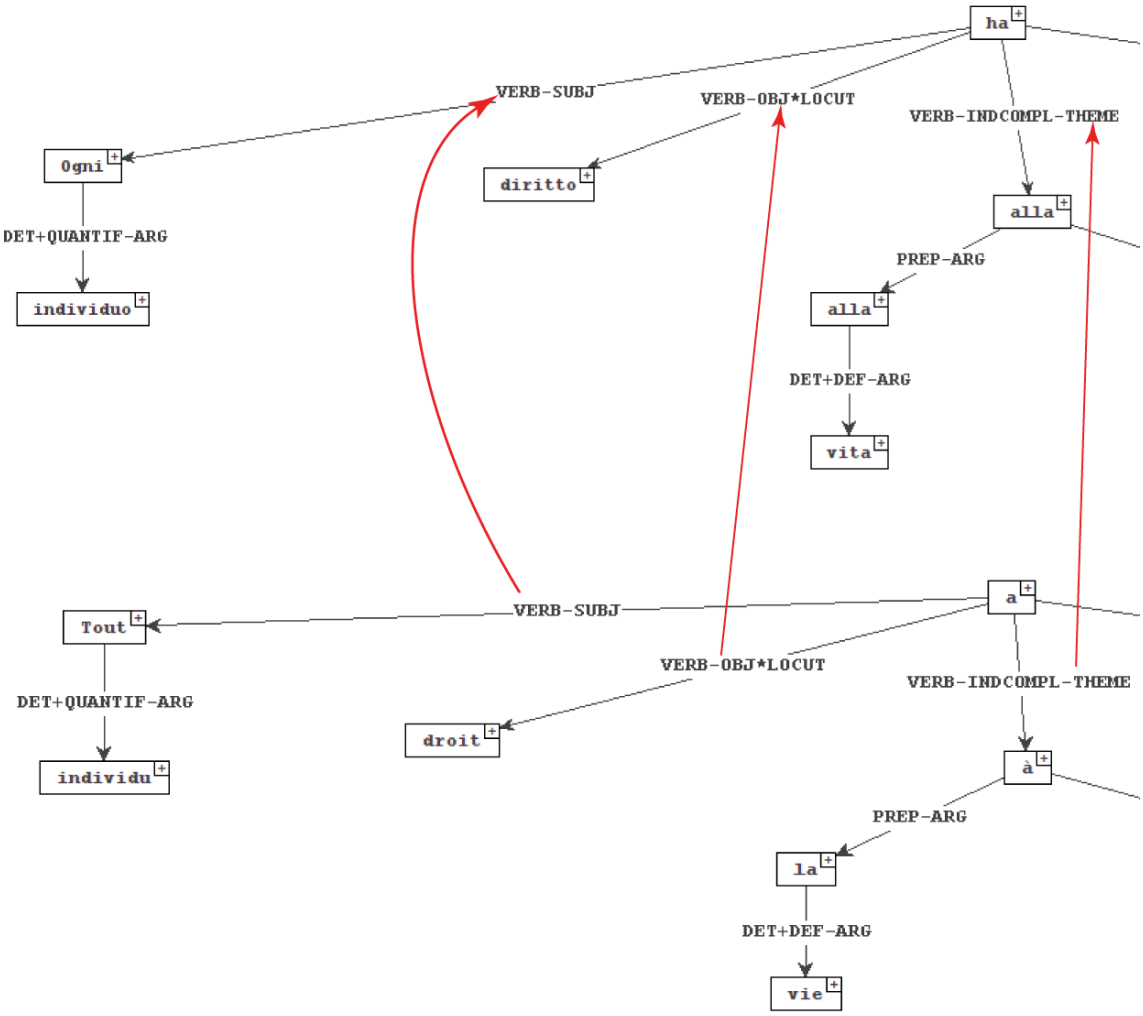

Figure 2: A sample of Italian-French alignment at dependency relation level in Parallel–TUT, for a fragment of the sentence ”Ogni individuo ha diritto alla vita” (UDHR–ITA–20) – ”Tout individu a droit à la vie” (UDHR–FR–19), corresponding to the UDHR–EN–21 for English: ”Everyone has the right to life”.

6 Discussion and future work

In this section we discuss the implications of applying the TUT format to English and French for the development of Parallel–TUT.

The first aspect we focused on, while evaluating our methodology and its effects, was the parser output, the type of errors produced and their investigation.

After the work phase described in Section 4, TULE, when evaluated according to its precision in building and labelling dependency trees, reached an error rate of around 9% for Italian, but 15,6% for English and 17,8% for French.

Errors detected during manual correction mainly dealt with tokenization and, to a larger extent, morphological analysis and Part of Speech tagging. This is maybe due to an incorrect application of disambiguation rules by the parser or to a lack of information about the lexical items in the TULE dictionary. As a result, these errors deeply affected the parser performance, and, despite rule-insertion operations, its output quality for English and French languages is still lower if compared to Italian. This suggests that further improvements in the system are required.

In addition to these errors, two other types have been identified. For their special character, we could define them as as “ language-dependent” and “genre-dependent” errors. In the first case, errors have to do with the distinctive feature of each language. The most frequent phenomenon (among those encountered in our corpus) included

in the former is that of the pre-modification in English, ie. all those cases of noun phrases where one or more units preceding the head of the phrase are syntactic modifiers of the head itself[14] structured in a hierarchic order. Since Italian language prefers post-modification, a parser trained for such linguistic patterns, in most cases, is unable to recognize the appropriate syntactic order between the units of the pre-modification.

As for the second type of errors, defined here as "genre-dependent", we include all those cases of errors directly attributable to the genre of the texts collected and analyzed in our small corpus. As we said, the collection comprises legal documents, where the recurrence of complex and ambiguous syntactic constructions (a feature shared by the three languages considered) is quite common. The high number of embedded prepositional phrases, subordinate clauses and parentheticals contributes to the lowering of the output quality.

As for the application of the TUT format and schema to the other two languages, distinctive features of these linguistic systems result in a lack of an appropriate structural representation, for which new relational labels were introduced, as described in Section 4. We tackled this problem with the two-fold goal of providing a coherent framework of annotation (like for Italian[15]), and taking into account the linguistic peculiarities of each language. This was made possible by a number of factors. First, the choice of a dependency (rather than constituent) structure better suits for both morphologically rich languages (such as Italian and French) and morphologically simpler ones (English). Moreover, the richness of relations provided in the TUT scheme, in addition to the use of null elements, which is another feature of the TUT format, allows a flexible annotation and the coverage of those linguistic phenomena which distinguish French and English from Italian (to name a few, the relative superlative in French, or the possessive case in English, as already mentioned in Section 4).

[14]This can be noticed, in the annotated texts, by the higher frequency of nominal modifiers (expressed by the NOUN-RMOD label) in English texts, rather than in the French and Italian sub-parts of the corpus; the occurences of such relation are 103 in English texts, 25 in the French and 17 in the Italian ones, covering respectively 2.5%, 0.5% and 0.4% of the total amount of relational labels.

[15]See the *linguistic notes* of TUT at `http://www.di.unito.it/ tutreeb/documents/noteling-engl-15-11-08.pdf`

As said at the beginning, the Parallel–TUT is currently an ongoing project, and the aim of the present work is mainly at raising and investigating issues related to its development. Nevertheless, in this phase of our project we observed that using the same format, and the TUT format in particular, has proved useful in the detection of similarities during the alignment phase at all the levels currently taken into account. The decision to adopt the same annotation scheme and grammatical description for the three languages can also contribute to the comparison of grammatical patterns.

As for future development of this work, a number of issues must be further pursued.

First of all, by taking into account the directions collected in the alignment guidelines developed during this first phase of the Parallel–TUT project, we will address the development and the integration of suitable tools, in particular for the alignment at the predicative structure level and for displaying such kind of information.

Secondly, considering the opportunity of converting TUT into a Penn-like format, we can extend the conversion to our parallel treebank as well, in order to develop alignment procedures also for phrases and information expressed in constituency-based formats.

Thirdly, in order to address the languages involved beyond the limits of a toy domain, it is crucial to enlarge the corpus of the Parallel–TUT. On the one hand, applying to a larger corpus our methodology to a larger corpus will give us the opportunity for addressing a larger and more meaningful set of linguistic phenomena typical of French and English, thouh not represented in Italian. On the other hand, this will allow more detailed analyses, like e.g. in (Ahrenberg, 2010), not affected by the sparseness of data that can be currently detected using our small corpus.

Finally, we observe that currently our corpus covers a selection of texts from a specific linguistic subfield broadly corresponding to legal language; one of the main future tasks should therefore consist not only in extending the size of the annotated corpus, but also in orienting to a more balanced direction its further development, comprising different sources, e.g. technical and specialized texts, fiction, newspapers (Paulussen and Macken, 2010).

7 Conclusions

In this paper we presented preliminary results in the creation of Parallel–TUT, a multilingual parallel treebank for Italian, English and French represented in the format of the Italian resource TUT. The project mainly aims at testing the hypothesis that the annotation schema and the knowledge annotated in the TUT format can be useful also to address the issues related to parallel corpora. Therefore, the same parsing system and the tools used for the improvement of the quality of the data annotated within TUT have been extended and applied to the other two languages.
Although this attempt has produced encouraging results, the project is currently ongoing and we presented several directions for its further development, extension and improvement.

References

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

J. Bos, C. Bosco, and A. Mazzei. 2009. Converting a dependency treebank to a Categorial Grammar treebank for Italian. In *Proceedings of the 8th workshop on Treebanks and Linguistic Theories (TLT-8)*, Milan.

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, A. Mazzei, and V. Lombardo. 2007. Evalita Parsing Task: an analysis of the first parsing system contest for Italian. *Intelligenza artificiale*, 2(IV).

C. Bosco, A. Mazzei, and V. Lombardo. 2009a. Evalita'09 Parsing Task: constituency parsers and the Penn format for Italian. In *Proceedings of Evalita'09*, Reggio Emilia.

C. Bosco, S. Montemagni, A. Mazzei, V. Lombardo, F. Dell'Orletta, and A. Lenci. 2009b. Evalita'09 Parsing Task: comparing dependency parsers and treebanks. In *Proceedings of Evalita'09*, Reggio Emilia.

C. Bosco. 2007. Multiple-step treebank conversion: from dependency to Penn format. In *Proceedings of the Linguistic Annotation Workshop (LAW) at ACL'07*, Prague.

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

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 NEMLAR the NEMLAR Conference on Arabic Language Resources and Tools*, Cairo.

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

N. Klyueva and D. Mareček. 2010. Towards Parallel Czech-Russian Dependency Treebank. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, Tartu.

L. Lesmo, V. Lombardo, and C. Bosco. 2002. Treebank development: the TUT approach. In *Proceedings of ICON02*, Mumbay.

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.

X. Li, S. Strassel, S. Grimes, S. Ismael, X. Ma, N. Ge, A. Bies, N. Xue, and M. Maamouri. 2010. Parallel aligned treebank corpora at LDC: Methodology, annotation and integration. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, Tartu.

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.

G. Musillo and K. Sima'an. 2002. Towards comparing parsers from different linguistic frameworks. An information theoretic approach. In *Proceedings of Workshop Beyond PARSEVAL - Towards improved evaluation measures for parsing systems at the LREC'02*, Las Palmas.

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

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

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.

D. Wu. 2010. Alignment. In *Handbook of NLP*. Chapmand and Hale/CRC Press.

H. Zhang and D. Gildea. 2004. Syntax-based alignment: Supervised or unsupervised? In *Proceedings of COLING'04*, Geneve.