10.12871/CLICIT2014161

Converting the parallel treebank ParTUT in Universal Stanford Dependencies

Manuela Sanguinetti
Dipartimento di Informatica
Università di Torino (Italy)
Corso Svizzera, 185, 10149 Torino
manuela.sanguinetti@unito.it

Cristina Bosco
Dipartimento di Informatica
Università di Torino (Italy)
Corso Svizzera, 185, 10149 Torino
cristina.bosco@unito.it

Abstract

English. Assuming the increased need of language resources encoded with shared representation formats, the paper describes a project for the conversion of the multilingual parallel treebank ParTUT in the *de facto* standard of the Stanford Dependencies (SD) representation. More specifically, it reports the conversion process, currently implemented as a prototype, into the Universal SD format, more oriented to a cross-linguistic perspective and, therefore, more suitable for the purpose of our resource.

Italiano. *Considerando la crescente necessità di risorse linguistiche codificate in formati ampiamente condivisi, l'articolo presenta un progetto per la conversione di una risorsa multilingue annotata a livello sintattico nel formato, considerato uno standard de facto, delle Stanford Dependencies (SD). Più precisamente l'articolo descrive il processo di conversione, di cui è attualmente sviluppato un prototipo, nelle Universal Stanford Dependencies, una versione delle SD maggiormente orientata a una prospettiva inter-linguistica e, per questo, particolarmente adatta agli scopi della nostra risorsa.*

1 Introduction

The increasing need to use language resources for the development and training of automatic systems goes hand in hand with the opportunity to make such resources available and accessible. This opportunity, however, is often precluded by the use of different formats for encoding linguistic content. Such differences may be dictated by several factors that, in the specific case of syntactically annotated corpora, or treebanks, may include the choice of constituency *v*s dependency-based paradigm, the specific morphological and syntactical features of the language at issue, or the end use the resource has been designed for. This variety of formats makes it more difficult the reuse of these resources in different contexts.

In the case of parsing, and of treebanks, a few steps towards the spread of formats that could be easily shared by the community has led, also thanks to the efforts devoted to the organization of evaluation campaigns, to the use of what have then become *de facto* standards. This is the case, for example, of the Penn Treebank format for constituency paradigms (Mitchell et al., 1993).
Within the framework of dependency-based representations, a new format has recently gained increasing success, i.e. that of the Stanford Typed Dependencies. The emergence of this format is attested by several projects on the conversion and harmonization of treebanks into this representation format (Bosco et al., 2013; Haverinen et al., 2013; McDonald et al., 2013; Tsarfaty, 2013; Rosa et al., 2014).

The project described in this paper is part of these ones and concerns in particular the conversion into the Stanford Dependencies of a multilingual parallel treebank for Italian, English and French called ParTUT. The next section will provide a brief description of ParTUT and its native format, along with that of the Universal Stanford Dependencies, while Section 3 will be devoted to the description of the conversion process, with some observations on its implications in the future development of ParTUT.

2 Data set

In this section, we provide an overview of ParTUT and of the two annotation formats at issue, focusing on their design principles and peculiarities.

2.1 The ParTUT parallel treebank

ParTUT[1] is a parallel treebank for Italian, English and French, designed as a multilingual development of the Italian Turin University Treebank (TUT)[2] (Bosco, 2001), which is also the reference treebank for the past parsing tracks of Evalita, the evaluation campaign for Italian NLP tools[3].

The whole treebank currently comprises an overall amount of 148,000 tokens, with approximately 2,200 sentences in the Italian and English sections, and 1,050 sentences for French[4].
ParTUT has been developed by applying the same strategy, i.e. automatic annotation followed by manual correction, and tool exploited in the Italian TUT project, i.e. the Turin University Linguistic Environment (TULE) (Lesmo, 2007; Lesmo, 2009), first developed for Italian and then extended for the other languages of ParTUT (Bosco et al., 2012). Moreover, one of the main developments of the treebank is also the creation of a system for the automatic alignment of parallel sentences taking explicitly into account the syntactic annotation that is included in these sentences (Sanguinetti and Bosco, 2012; Sanguinetti et al., 2013; Sanguinetti et al., 2014).

2.2 The TUT representation format

The treebank is annotated in a dependency-based formalism, partially inspired by the Word Grammar (Hudson, 1990), in particular for what concerns the head selection criteria for determiners and prepositions (that are considered as governors of the nominal and prepositional groups respectively). Other typical features of TUT and ParTUT 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.

For what concerns the dependency labels, they were conceived as composed of two components[5] according to the following pattern:

morphoSyntactic–functionalSyntactic.

The main (and mandatory) feature is the second one, specifying the syntactic function of the node in relation to its governor, i.e. whether the node is an argument (ARG), a modifier (MOD) or a more specialized kind of argument (e.g. OBJ or SUBJ) or modifier (e.g. RMOD for restrictive modifiers and APPOSITION for the others) or something else (e.g. COORD or SEPARATOR). This component can be preceded by another one that specifies the morphological category *a*) of the governing item, in case of arguments (e.g. PREP-ARG for the argument of a Preposition), *b*) of the dependent, in case of modifiers (e.g. PREP-RMOD for a prepositional restrictive modifier). In some cases, the subcategory type of this additional component is also included (after a '+' sign), as in DET+DEF-ARG, which should be read as argument of a definite Determiner.

TUT aims at being as linguistically accurate as possible, providing a large number of labels for each of these two components, which can be easily combined together to express the specificity of a large variety of syntactic relations. It thus results in a high flexibility of the format that allowed its application to languages different from the original one (that is Italian).

2.3 The Stanford Typed Dependencies

The Stanford Dependencies (SD) representation (de Marneffe et al., 2006; de Marneffe and Manning, 2008; de Marneffe and Manning, 2008; de Marneffe et al., 2013) was originally developed for English syntax to provide a scheme that could be easy to use in practical NLP tasks, like Information Extraction. This led to the choice of a format that was theory-neutral as regards the specific grammar, and of a set of widely recognized grammatical relations. Indeed, one of the key features of SD representation, throughout the different versions proposed, is namely the trade-off between linguistic fidelity and readability, which is probably the main factor that determined its usability, and, finally, its success.

Recently, a new version of the SD scheme has been proposed, i.e. the Universal Stanford Dependencies (USD)[6], a revised set of relations more oriented to provide a uniform and consistent structural representation across languages of different linguistic typologies (de Marneffe et al., 2014).

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

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

[3] `http://www.evalita.it/`

[4] The resource is under constant development, and the French part of the newest texts recently added to the collection is yet to be analyzed and included.

[5] In the Italian TUT there is also a third one (omitted here and in the current ParTUT annotation) concerning the *semantic role* of the dependent with respect to its governor.

[6] `http://universaldependencies.github.io/docs/`

By virtue of this claim, more emphasis is put on the *lexicalist hypothesis*, that ultimately favors the relations between content words, with the aim of properly dealing with compounding and rich morphology. This affects, among the other things, the treatment of prepositions, which – rather than mediate between the modified word and the modifier – are now attached as dependents of the latter.
Furthermore, in order to allow the proper recognition of language-specific phenomena, USD representation also opens to possible extensions by adding new grammatical relations as subtypes of the existing ones. This flexibility in the labeling scheme is a valuable feature that USD has in common with the TUT format.

In light of these observations, in this conversion project we opted for the USD representation scheme as the target format.

3 Converting ParTUT

In this section, we describe the current, preliminary, stage of this project. This phase mainly consists in a qualitative comparison of the two formats at hand, drafting a basic mapping scheme between the two relation taxonomies and highlighting the main factors that could impact – both positively and negatively – the conversion process, currently implemented as a prototype.

Mapping scheme As expected, we encountered only 13 cases of 1:1 correspondences between the items of the two relation sets, although, conversely, in relatively few cases (9) a counterpart could not be found either in the source or the target format. The remaining ones entailed a multiple correspondence either 1:*n* or *m*:1. A small selection of such cases, based on the 15 most commonly used relations in ParTUT, is proposed in Table 1.

Preliminary observations The conversion from TUT to USD seems to be especially feasible because of the high flexibility of the involved schemes and their openness to cross-linguistic applications. Furthermore, we can benefit from the fact that we are moving from a source format with a high level of detail to a target format that is more underspecified[7].

[7]TUT scheme comprises an overall amount of 11 *morphoSyntactic* and 27 *functionalSyntactic* features (not to mention their subtypes) that can be combined together, while USD taxonomy includes 42 grammatical relations (which is a further reduction in number, with respect to the previous SD versions).

TUT	USD	H.m.
VERB-SUBJ	*nsubj, csubj*	Y
VERB-OBJ	*dobj, xcomp*	N
VERB-SUBJ/ VERB-INDCOMPL-AGENT	—	—
VERB-OBJ/VERB-SUBJ	*nsubjpass*	N
PREP-ARG	*case*	Y
DET+DEF-ARG	*det, poss*	Y
DEF+INDEF-ARG	*det*	Y
CONJ-ARG	*mark, xcomp*	Y
PREP-RMOD	*case*	Y
ADJC+QUALIF-RMOD	*amod*	N
COORD2ND+BASE	*conj, cc*	Y
COORD+BASE	*cc*	N
END	*punct*	N
SEPARATOR	*punct*	N
TOP-VERB	*root*	N

Table 1: A mapping scheme between the 15 most used syntactic relations in ParTUT and their counterparts in USD. The third column reports whether there is a (either direct or complex) head movement (H.m.) when transforming TUT representation into USD.

English-particular relations, for example, can be easily mapped onto the ones used in ParTUT, and, except for one specific case (that of verb particles), can also be applied to Italian as well as French constructions. Such cases are, respectively, *a*) temporal modifiers expressed with a NP; *b*) predeterminers; *c*) words preceding a conjunction; *d*) possessives.

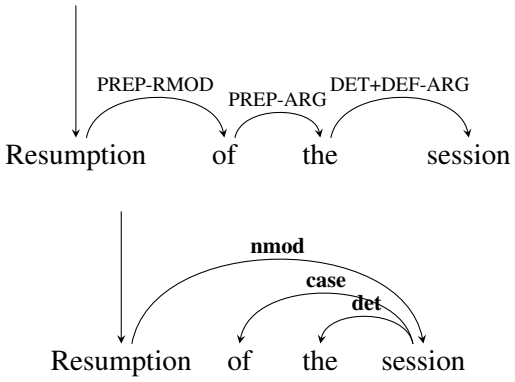
TUT	USD	H.m.
PARTICLE* (*English only)	*prt*	N
NOUN-RMOD-TIME	*tmod*	Y
PDET-RMOD	*predet*	N
COORDANTEC	*preconj*	N
DET+DEF-ARG	*poss*	Y

Table 2: English-particular relations in USD that can be mapped onto the ones used in ParTUT. Unless stated otherwise, all the relations reported in the table can also be applied to Italian and French.

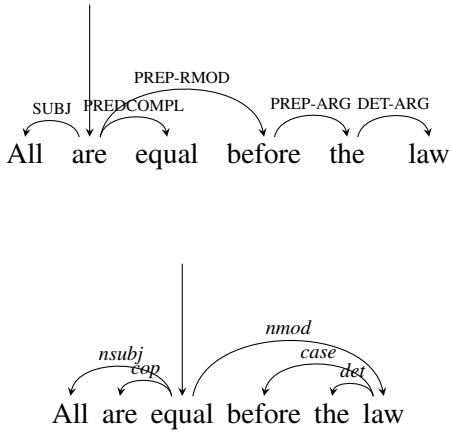
However, as briefly introduced in Section 2.3, the choice to establish meaningful syntactic links between content words not only characterizes this version of SD with respect to the previous ones,

but it also marks a clear boundary with the TUT representation. This aspect entails two basic types of conversion procedures in case of non-direct correspondences, that mainly concern the head selection criteria, and that can be summarized as follows:

- a direct head swapping, where conversion is carried out by a simple inversion of head and dependent roles, as in the case of determiners and prepositions (see below two parallel examples of TUT, in the upper part, and USD, in the lower one):



- a complex transformation that may involve the whole subtree. This is the case, for example, of copulative verbs, that are annotated as heads in ParTUT, and as dependents – together with the subject itself – of the predicative complement in USD (see below).



On the other hand, a more semantically-oriented representation has its benefits as well, especially when dealing with parallel texts in different languages annotated according to the same scheme[8]. This proves useful for translation purposes, which is one of the main goal ParTUT has been conceived for, since it could make it easier the identification of translational correspondences, both manually and automatically, and it constitutes therefore a meaningful step for the further development of the resource as a whole.

Implemented conversion The implementation of the converter is driven by the mapping scheme and observations mentioned above. Each single relation is classified according to different perspectives, including e.g. granularity and mapping cardinality. Adequate procedures are developed to deal with the transformations necessary to the conversion for each relation class. Some procedures, e.g. those implementing a complex restructuring rather than a simple relation renaming, exploit not only the syntactic knowledge but also PoS tagging associated to dependency nodes.
The output of the conversion is made available in different notations known in literature: besides the typical bracketed notation of SD, the converted version will be also released in CoNLL-U[9] and using the Universal PoS tagset proposed by Petrov et al. (2012)

4 Conclusion

In this paper, we briefly described the ongoing project of conversion of a multilingual parallel treebank from its native representation format, i.e. TUT, into the Universal Stanford Dependencies. The main advantages of such attempt lie in the opportunity to release the parallel resource in a widely recognized annotation format that opens its usability to a number of NLP tasks, and in a resulting representation of parallel syntactic structures that are more uniform and, therefore, easier to put in correspondence. Conversion, however, is not a straightforward process, and a number of issues are yet to be tackled in order to obtain a converted version that is fully compliant with the target format. The next steps of this work will focus in particular on such issues.

[8] Although recent works (Schwartz et al., 2012) seem to point to the fact that while content word-based schemes are more readable and "interlingually" comparable, they are harder to learn by machines; this is, in fact, an aspect we intend to verify in the validation phase of the converted resource, by using it as training set for a statistical parser, as also described in Simi et al. (2014).

[9] http://universaldependencies.github.io/docs/format.html

References

Cristina Bosco. 2001. Grammatical relation's system in treebank annotation. In *Proceedings of Student Research Workshop of Joint ACL/EACL Meeting*, pp. 1–6.

Cristina Bosco and Alessandro Mazzei. 2012. The EVALITA Dependency Parsing Task: From 2007 to 2011. In B. Magnini, F. Cutugno, M. Falcone and E. Pianta (Eds.), *Evaluation of Natural Language and Speech Tools for Italian*, pp. 1–12.

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

Cristina Bosco, Simonetta Montemagni and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford Dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 61–69.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), pp. 449–454.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford Typed Dependencies representation. In *Coling2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, CrossParser 08*, pp. 1–8 .

Marie-Catherine de Marneffe and Christopher D. Manning 2008. Stanford Typed Dependencies manual (Revised for the Stanford Parser v. 3.3 in December 2013). `http://nlp.stanford.edu/software/dependencies_manual.pdf`.

Marie-Catherine de Marneffe and Miriam Connor and Natalia Silveira and Samuel R. Bowman and Timothy Dozat and Christopher D. Manning. 2013. More Constructions, More Genres: Extending Stanford Dependencies. In *Proceedings of the Second International Conference on Dependency Linguistics* (DepLing 2013), pp. 187–196.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen,Filip Ginter, Joakim Nivre and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), pp. 4585–4592.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski and Filip Ginter 2013. Building the essential resources for Finnish: the Turku Dependency Treebank. In *Language Resources and Evaluation*, 48:3, pp. 494–531.

Richard Hudson. 1990. *Word Grammar*. Basil Blackwell, Oxford and New York.

Leonardo Lesmo 2007. The rule-based parser of the NLP group of the University of Torino. In *Intelligenza artificiale*, IV:2, pp. 46–47 .

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

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of 51st annual meeting of the Association for Computational Linguistics* (ACL'13), pp. 92-97 .

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics*, 19:2.

Slav Petrov, Dipanjan Das and Ryan McDonald 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12).

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman and Zdeněk Žabokrtský. 2014. HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), pp. 2334–2341.

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

Manuela Sanguinetti, Cristina Bosco and Leonardo Lesmo 2013. Dependency and Constituency in Translation Shift Analysis. In *Proceedings of the 2nd Conference on Dependency Linguistics* (DepLing'13), pp. 282-291 .

Manuela Sanguinetti, Cristina Bosco and Loredana Cupi. 2014. Exploiting *catenae* in a parallel treebank alignment. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC'14), pp. 1824-1831.

Roy Schwartz and Omri Abend and Ari Rappoport. 2012. Learnability-Based Syntactic Annotation Design. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING'12), pp. 2405–2422.

Maria Simi, Cristina Bosco and Simonetta Montemagni 2014. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of*

the 9th International Conference on Language Resources and Evaluation, (LREC'14), pp. 83–90.

Reut Tsarfaty 2013. A unified morpho-syntactic scheme of Stanford Dependencies. In *Proceedings of 51st annual meeting of the Association for Computational Linguistics* (ACL'13), pp. 578–584.