Enhanced UD Dependencies with Neutralized Diathesis Alternation

Marie Candito, Bruno Guillaume, Guy Perrier, Djamé Seddah

Enhanced UD Dependencies with Neutralized Diathesis Alternation

Marie Candito
Univ. Paris Diderot, CNRS
Laboratoire de Linguistique Formelle
France
marie.candito@linguist.univ-paris-diderot.fr

Bruno Guillaume
Inria Nancy Grand-Est, Loria
France
bruno.guillaume@loria.fr

Guy Perrier
Univ. de Lorraine, Loria, UMR 7503
France
guy.perrier@loria.fr

Djamé Seddah
Univ. Paris-Sorbonne, Inria
France
djame.seddah@paris-sorbonne.fr

Abstract

The 2.0 release of the Universal Dependency treebanks demonstrates the effectiveness of the UD scheme to cope with very diverse languages. The next step would be to get more of syntactic analysis, and the "enhanced dependencies" sketched in the UD 2.0 guidelines is a promising attempt in that direction. In this work we propose to go further and enrich the enhanced dependency scheme along two axis: extending the cases of recovered arguments of non-finite verbs, and neutralizing syntactic alternations. Doing so leads to both richer and more uniform structures, while remaining at the syntactic level, and thus rather neutral with respect to the type of semantic representation that can be further obtained. We implemented this proposal in two UD treebanks of French, using deterministic graph-rewriting rules. Evaluation on a 200 sentence gold standard shows that deep syntactic graphs can be obtained from surface syntax annotations with a high accuracy. Among all arguments of verbs in the gold standard, 13.91% are impacted by syntactic alternation normalization, and 18.93% are additional deep edges.

1 Introduction

The Universal Dependencies initiative (UD, (Nivre et al., 2016)) is one of the major achievements of the last few years in the NLP field. Originating from the need of a better interopability in cross-language settings for downstream tasks (Petrov et al., 2011; McDonald et al., 2013), it has gathered dozens of international teams who released annotated versions of their treebanks, following the UD annotation scheme.

Although UD has raised criticisms, both on the suitability of the scheme to meet linguistic typology (Croft et al., 2017) and on the current implementation of the UD treebanks (Gerdes and Kahane, 2016), the existence of many treebanks with same syntactic scheme does however ease cross-language linguistic analysis and enables parsers to generalize across languages at training time, as demonstrated by Ammar et al. (2016).

The UD scheme favors dependencies between content words, in order to maximize parallelism between languages. Although this results in dependencies that are more semantic-oriented, the UD scheme lies at the surface syntax level and thus necessarily lacks abstraction over syntactic variation and does not fit all downstream applications' needs (Schuster and Manning, 2016).

This is partly why de Marneffe and Manning (2008) proposed a decade ago, in the Stanford Dependencies framework, several schemes with various semantic-oriented modifications of syntactic structures. Its graph-based, so-called *collapsed*, representation layer[1] has recently started to be extended and implemented as "Enhanced Dependencies" in the UD scheme family (Schuster and Manning, 2016). Current UD specifications leave open the possibility to include phenomena (cf. section 2) that make explicit additional predicate-argument dependencies. In practice, most current UD treebanks contain either very few or no enhanced dependencies at all[2].

[1] Among the various Stanford schemes, the collapsed scheme is the furthest away from the plain dependency tree.

[2] Notable exceptions in the UD 2.0 release are the SyntagRus and Finish treebanks. For English, a converter including enhanced dependencies is available within the Stanford parser (https://nlp.stanford.edu/software/stanford-dependencies.shtml).

Of course, as noted by Kuhlmann and Oepen (2016), competing proposals for deep syntactic graphs already exist and are implemented through diverse and, in some few cases, multilingual *graphbanks*. More clearly semantic schemes seem to depend on the needs of the downstream application or impose their own constraints on the syntactic layer it is either built upon or plugged in. See for example the differences between abstract meaning representations (Knight et al., 2014), designed with Machine Translation in sight, and the UDEPLAMBDA's logical structures, very recently proposed by Reddy et al. (2017) and evaluated on a question-answering over a knowledge base task.

In this paper, we build on the work of (Candito et al., 2014; Perrier et al., 2014) to propose an extension to the current *enhanced* dependency framework of Schuster and Manning (2016). First, we extend the types of argumental dependencies made explicit (taking into account participles, control nouns and adjectives, non-finite verbs and more cases of infinitive verbs). Second, we neutralize syntactic alternations, in order to make linking patterns more regular for a given verb form. We believe that making explicit and normalize the predicate-argument structures, still remaining at the syntactic level, can make downstream semantic analysis more straightforward (as shown for instance in (Michalon et al., 2016)), while remaining neutral with respect to what exact semantic representation can be further derived.

The originality of our approach is to neutralize syntactic alternations using *canonical* grammatical functions, which render linking patterns of verbs more regular but are still syntactic in nature, unlike what can be found for example in the tectogrammatical layer of the Prague Dependency bank (Hajic et al., 2006).

This proposal is currently being implemented for French, and tested on two UD treebanks (Candito et al., 2014; Nivre et al., 2016) by the means of a rule-based deterministic process. We evaluated the deep syntactic graphs automatically converted from gold UD trees and obtained a 94% F-measure on a two-hundred sentences gold standard, similar to what reported Candito et al. (2014) on a similar task. Both treebanks and building rules are made available[3] to foster further work in other languages and to gather the opinion and criticisms of the community regarding the level of abstraction we should reach when it comes to deep syntax representation.

In the following, we first briefly introduce the current Enhanced UD scheme, we detail extensions concerning arguments of non-finite verbs in section 3 and syntactic alternations for French in section 4. We present and evaluate a system to obtain enhanced graphs for French in section 5. We then discuss related work and conclude.

2 Enhanced UD representation

The current version of universal dependencies guidelines (v2.0) includes an enhanced dependencies section[4], leaving the possibility for UD treebanks to include all or only some of the following phenomena:

1. Additional subject relations for control and raising constructions
2. Propagation of conjuncts
3. Antecedent of relative pronouns in noun-modifying relative clauses
4. Modifier labels that contain the preposition or other case-marking information
5. Null nodes for elided predicates

In our implementation for French, we cope with the two first phenomena. Phenomena 3 and 4 are quite systematic and may be handled automatically and phenomenon 5 requires manual annotation. Note that while enhanced dependencies (as were Stanford dependencies) are motivated by downstream semantically-oriented applications, they remain syntactic in nature in their current stage. This results in keeping syntactic dependents that are not semantic arguments of their syntactic head, in classic cases of syntax/semantics mismatch. So for instance, subjects of raising verbs are not removed from the enhanced UD graph, although they are not a semantic argument of the raising verb, as shown in Fig. 1.

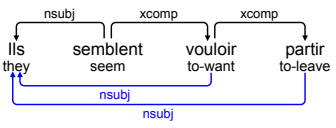

Figure 1: *Raising verb*

Following the work of Candito et al. (2014) and Perrier et al. (2014), we propose two extensions,

[3] http://github.com/bguil/Depling2017

[4] http://universaldependencies.org/u/overview/enhanced-syntax.html

that we detail in the next two sections: the first one is to extend the cases for which arguments are added to infinitive verbs and more generally to non-finite verbs. The second one concerns the neutralisation of syntactic alternations.

3 Recovering arguments of non-finite verbs

The aim of enhancing UD dependencies is to facilitate the computation of predicate-argument relations at the semantic level. In this perspective, we propose to go beyond the explicitation of control and raising verbs subjects. We detail below other cases of obligatory syntactic control, and cases which are not as systematic but which prove feasible with rather high accuracy using heuristics.

3.1 Cases fully determined by syntax

"Control nouns" In French, some nouns take a nominal and an infinitive argument, that can be both realized within the NP or as a predicative complement (Fig. 2). In both cases, the subject of the infinitive is the nominal argument.

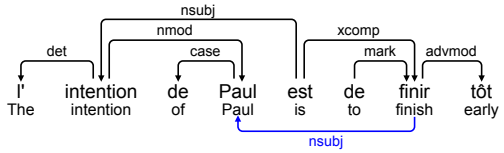

Figure 2: *Paul's intention is to finish early*

The preposition introducing the infinitival clause is determined by the control noun. It is generally *de*, more rarely *à*, as in example (1).

(1) votre capacité à **conduire** un véhicule
your capacity to drive a vehicle

"Control adjectives" Control adjectives take an infinitive complement, whose understood subject is the noun to which the adjective applies, as shown in Fig. 3.

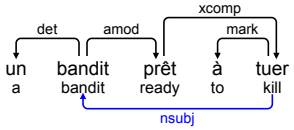

Figure 3: Control adjective

Tough movement Tough movement describes constructions in which an adjective has an infinitive as complement and the noun to which the adjective applies is the direct object of the infinitive. The adjective can be attributive or used as a predicative adjective (Fig. 4)[5]. These cases are easy to detect using available lists of tough adjectives[6].

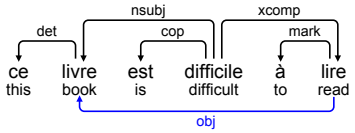

Figure 4: Tough movement

Noun-modifying participles When a past or present participle modifies a noun, the noun is the understood subject of the participle (Fig. 5).

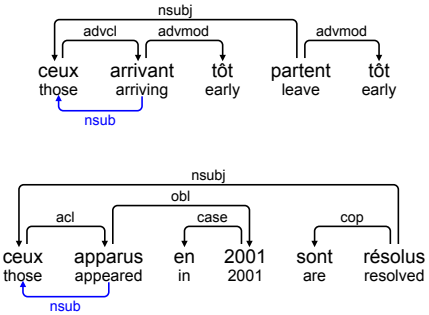

Figure 5: Noun-modifying participles

Infinitives behaving as noun modifiers In French, a transitive infinitival clause introduced with the preposition *à* can be the argument of the noun (as in example (1) in the "control nouns" section above, the noun *capacité (ability)* takes two arguments, the entity having the ability, and an infinitival clause describing what it is able of). But for any noun, an infinitival clause introduced by *à* can function as an adjunct modifying the noun, which is understood as either the object (Fig. 6) or the subject (examples (2) and (3)), depending on the transitivity of the infinitive.

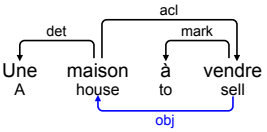

Figure 6: Infinitive modifying a noun, understood as the object of the infinitive

(2) C'est une machine à mesurer la pression
It's a machine to measure the pressure

[5]Note in this case, the modified noun is not a semantic argument of the adjective, the dependency between *difficile (difficult)* and *livre (book)* should be dropped in a semantic representation.

[6]A few "tough nouns" exist too, as in *ce livre est un plaisir à lire (this book is a pleasure to read)*.

"It's a pressure measuring machine"

(3) Elle est la première femme à y entrer
She is the first woman to in-it enter
"she is the first woman who ever entered it"

3.2 Cases requiring semantic or world knowledge

The cases we just saw correspond to situations of *obligatory control*, in which the argument to add to the non-finite verb can be deterministically identified, given the syntactic construction, and given the specific control or raising verb, control noun or adjective. Other constructions involving an non-finite verb are ambiguous with respect to which non-local argument is understood as the argument of the verb. In some of these cases though, among all the potential positions for the non-local argument to retreive is particularly more frequent, although not strictly obligatory. For the cases detailed in this section, we performed a systematic study of the occurrences in the Sequoia corpus, and concluded that simple heuristics could be used for retreiving the non-local argument of a non-finite verb with sufficient accuracy.

Dislocated participle clauses: A participle clause modifying a noun can appear "dislocated" at the beginning or end of the sentence. In that case, its subject is most often the subject of the participle, although exceptions can be built[7].

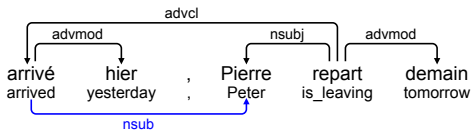

Figure 7: Dislocation

Verb-modifying infinitival and participial clauses For certain prepositions introducing infinitival clauses, the subject of the infinitive is most often the subject of the main clause, but exceptions as illustrated in ex. (4) (the subject of *terminer* is not provided in the sentence.).

(4) Cela exige beaucoup de travail pour **terminer**
it takes a lot of work to finish
à temps
on time

[7]We did not find any such exception in the Sequoia corpus. The following built up example shows one: *Exténués, on les a envoyés dormir. (Exhausted, we them have sent to-sleep) "Exhausted, they were sent to bed").*

We performed an in-depth study of these cases, using the deep Sequoia corpus (Candito et al., 2014), in which all subjects of infinitive verbs present in the sentence are marked. Breaking down the 143 infinitive heads of adverbial clauses according to the voice of the main verb, we obtain the following results:

- *main verb in the active voice:* there are 114 cases and among them, the subject of the infinitive is the subject of the main verb in 95 cases; in the 16 remaining cases, the subject of the infinitive is absent of the sentence;
- *main verb in passive voice (or modal introducing a passive):* there are 29 cases; in 11 cases, the subject of the infinitive is the subject of the main verb; in the 18 remaining cases, the subject of the infinitive is a virtual agent of the passive verb, which is not present in the sentence;
- *main verb in medio-passive voice:* there are 3 cases, in which the subject of the infinitive is not present in the sentence.

A heuristic that triggers the sharing for active main verbs only will obtain a 90% recall and 83% precision only.

In a similar construction, a present participle introduced with a preposition (*en* in French and *by* in English) plays the role of a modifier for a main verb. The subject of the participle is generally the subject of the main verb but again, this does not hold if the main verb is in passive voice (or is a modal introducing a passive, as shown in ex. (5).

(5) Ce médicament doit être pris en
This drug should be taken by
mangeant
eating
"This drug should be taken while eating"

In Sequoia, there are 39 such constructions. For all the 30 cases in which the main verb is in active voice, the subject of the main verb is understood as the subject of the participle. For the 9 cases in which the main verb is passive, for 8 of them the subject of the participle is not present in the sentence. Therefore, an automatic procedure taking into account the voice of the main verb should produce only a very small number of errors.

Arbitrary control Arbitrary control is a construction in which the subject of an infinitive can have any position in the sentence (Baschung, 1996).

(6) **Fumer** est dangereux pour la santé
Smoking is dangerous for the health

(7) **Fumer** est dangereux pour lui
Smoking is dangerous for him

In Example (6), the subject of *fumer* is understood as generic while in Example (7), the subject is *lui*. While by definition such control cannot be easily resolved, such constructions are fortunately very rare in corpora and ignoring them produces few missing subjects of infinitives.

4 Neutralizing syntactic alternations

Syntactic alternations (like passive) are known to cause diversity in the observed linking patterns in corpora, i.e. the grammatical functions born by the semantic arguments of a verb. At least some of the existing syntactic alternations are very general and can be identified purely on syntactic grounds, without resorting to semantic disambiguation. In this work, we advocate for neutralizing such variation in an "enhanced-alt UD" representation (enhanced UD representation augmented with syntactic alternation neutralization). Following (Candito et al., 2014; Perrier et al., 2014), we propose to distinguish *canonical* versus *final* grammatical functions, and to normalize syntactically alternated verb instances by making explicit the canonical grammatical functions of their arguments. The objective is to cluster observed subcategorization frames into possibly one canonical frame, with thus one linking pattern between canonical functions and semantic arguments.

We handle the French syntactic alternations for which morpho-syntactic clues are available, namely passive, medio-passive, impersonal and causative. We detail these below, identifying for each what is feasible using morpho-syntactic and lexical clues only, and what requires semantic information.

4.1 Passive

Passive is by far the most frequent syntactic alternation, and it is fortunately rather easy to identify in a language such as French. Note that because the UD scheme uses several labels for the same argumental slot, depending on the argument's category, the basic rule of having the passive's subject being the canonical direct object has to be split. The `nsubj:pass` dependent is considered the canonical `obj`. The `csubj:pass` dependent is the canonical `ccomp` (for full clauses), or `xcomp` (for infinitival phrases).

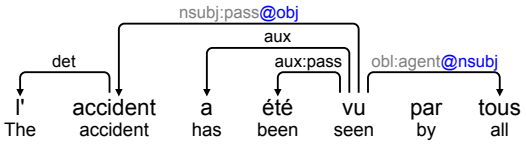

Figure 8: Passive with canonical functions made explicit.

Although passive is identified unambiguously, correctly identifying the argument that is subject in the active form (the "by-phrase" in English) is more problematic given the UD scheme. In French, it is introduced by a PP with preposition *par* (Fig. 8) or for certain verbs, with preposition *de*. But both prepositions can also introduce adjuncts, and the current French version of UD scheme uses the same label `obl` in both cases, leading to an ambiguity concerning the argumental status of the PP. In the following, we use a more specific `obl:agent` label for the *by*-phrases, as is done e.g. in the UD versions of the par-TUT parallel treebank (Sanguinetti and Bosco, 2014) (for English, French and Italian). We detail in section 5 how we can obtain this labeling for the other French UD treebanks.

4.1.1 English passive and ditransitives

Although our focus is French, we also describe here briefly how to handle passive of English ditransitives, a case that does not exist in French.

Let us first note that the current marking of passive in the UD scheme (`nsubj` versus `nsubj:pass` distinction, and `aux:pass` label for passive auxiliary) is not always directly usable to link syntactic arguments to semantic ones. First, passive forms without auxiliaries are not currently marked as such (e.g. in *the planet reached by astronauts*). Second, even for a passive form with passive auxiliary, the recommended `nsubj:pass` label is ambiguous in case of a ditransitive verb: for instance in *He was given orders* and *Orders were given to him*, the `nsubj:pass` corresponds to different semantic arguments[8]. If we choose the double object frame as canonical frame for ditransitives, then the canonical labels can be made explicit as shown in figure 9. Note that the canonical function of the

[8]This is already identified by Gerdes and Kahane (2016), who advocate for directly adding the semantic argument rank (1,2,3...) on top of the syntactic label.

nsubj:pass argument is iobj if the verb has a direct object (Fig. 9a) or obj otherwise (Fig. 9b).

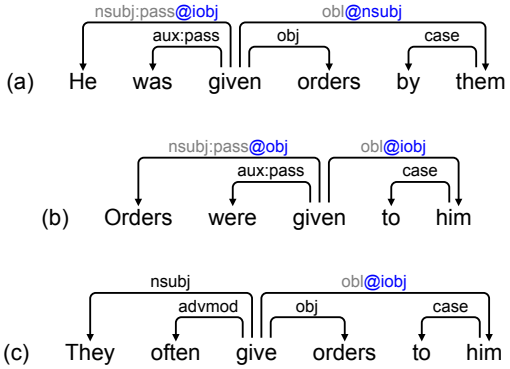

Figure 9: Syntactic alternation normalization for ditransitives.

4.2 Medio-passive

The French reflexive clitic *se* has various status. Roughly, it can mean true reflexivization (*Jean se voit (Jean SE sees) "Jean is seeing himself"*), be part of a compound verb (*s'apercevoir (to realize)*), or mark a valency alternation in which the object is promoted to subject. In the latter case, the canonical subject argument cannot be realized locally, but from the semantic point of view, an agent is either understood (Fig. 10b) or not (Fig. 10a). Disambiguating the status of a given *se* instance is a difficult task requiring semantic information. Note though the phenomenon is not massive. For instance in the Sequoia corpus (Candito et al., 2014), about 5.7% of verbs bear a *se* clitic, among which 16% correspond to a syntactic alternation.

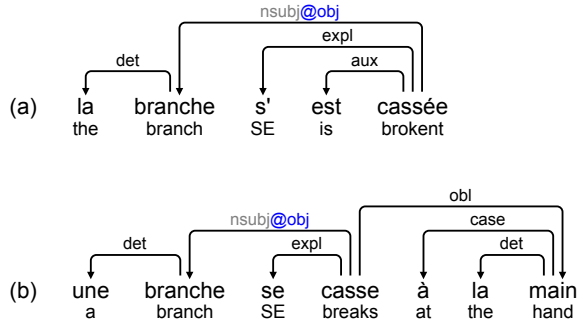

Figure 10: Medio-passive, with or without understood agent (*The branch broke* and *One can break a branch by hand*)

4.3 Impersonal

Impersonal constructions can also be viewed as syntactic alternations: in French the postverbal complement has object-like properties (in particular the pronominalization with the quantitative clitic *en (of-it)*).

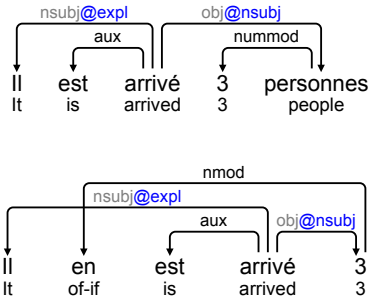

Figure 11: Impersonal construction for sentences *"There arrived 3 people"* (top) and *"Three (of them) arrived"* (bottom).

The representation of such constructions in UD is subject to debate. In the French-UD v2.0 treebank, the non-referential *il* clitic is treated as a nsubj, and the post-verbal argument as an object. We thus handle impersonal constructions as syntactic alternations (Fig. 11): the *il* receives an expl label, and the post-verbal dependent receives a canonical nsubj or csubj label (unless the verb is passive).

4.4 Causative

Causative is another construction that can be viewed as a syntactic alternation in French. It is formed syntactically with a *faire (to do)* verb followed by the infinitive of the "caused" verb. It has complex properties described in a vast litterature. For instance Abeillé et al. (1997) advocate for two competing analyses, the main one representing the *faire* + Vinf as a complex predicate, with the arguments of Vinf plus an argument for the causer, which shows as final subject (we use nsubj:caus as canonical function to mark it in the enhanced UD representation). The causee, which corresponds to the canonical subject of the Vinf, can show as a direct object, an oblique with preposition *à* or preposition *par*, depending on the transitivity of the Vinf, and other complex factors. So though detecting a causative construction is trivial, detecting which surface argument of the complex predicate corresponds to the causee is not. We provide in Fig.12 an example of ambiguity: *Zola* can

be understood as the author that is read or the person who reads. The phenomenon is rather rare, e.g. occurring roughly once every 100 sentences in the Sequoia treebank.

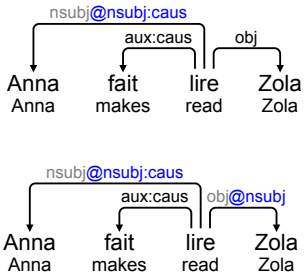

Figure 12: Ambiguous causative sentence, meaning either *"Anna makes someone read Zola"* (top) or *"Anna makes Zola read"* (bottom, *Zola* is the canonical subject).

4.5 Interaction

Syntactic alternations can interact with all the other "UD-enhanced" phenomena. For ease of reading, we provide an English example in Fig. 13, where coordination interacts with passive and a secondary predicate construction[9]. We further focus on interaction between passive and added dependents of verbs. For all the cases listed in sections 2 and 3 in which a subject is added to a non-finite verb, the syntactic regularity concerns the *final* grammatical subject, which does or doesn't correspond to the *canonical* subject, depending on the voice of the verb. We develop below two examples: (i) noun-modifying particial phrases and (ii) control verbs.

Passive and noun-modifying participial phrases: We wrote in section 3 that a noun modified by a participle corresponds to the subject of the participle (Fig. 5). Yet, this generalization only holds if subject is intended as *final* subject. Fig. 14 shows examples of past participles, with or without auxiliaries, that modify a noun. The noun is the semantic first actant of the intransitive participle (a), but the semantic second actant of the transitive participle (b). Using the notion of final versus canonical grammatical functions, we can uniformely state that in all cases, the modified noun is the final subject of the participle (whether past or present participle), and consider (i) all present participles as active, (ii) the intransitive participles as active, but (iii) the transitive participles as passive. For the latter, the final subject is the canonical object, as usual for passives.

Note that from a practical point of view, it is rather easy to decide whether a given noun-modifying past participle falls under case (ii) or (iii). Indeed, only a few intransitive verbs[10] can function as noun-modifying past participle phrases (case (ii)), all other instances necessarily fall under the passive case (iii).

Passive and control verbs: For control verbs we have both a syntactic constraint and a semantic (or lexical) constraint: a control verb controls which of its *semantic* argument will necessarily be the *(final) subject* of the infinitive. For instance, let's consider first the so-called "subject control verbs" (e.g. *vouloir (to want)*) or movement verbs (e.g. *venir (to come)*). The canonical subject of such verbs (*ceux (those)* in Fig. 15) is the final subject of the infinitive, but its canonical subject for active infinitives (Fig. 15a and Fig. 15c) and canonical objects for passive infinitives ((15b).

For "object control verbs", the controller (final subject of the infinitive) is their canonical object. This holds both for active (Fig. 16a) or passive object control verbs (Fig. 16b). For instance in Fig. 16b, *forcer (to force)* is passive, the controller (*ceux (those)*) is always its canonical object, but shows as its final subject.

5 Producing enhanced graphs for French UD treebanks

We have experimented the proposed enhanced scheme on two French corpora of the UD project: UD_FRENCH and UD_FRENCH-SEQUOIA. UD_FRENCH is in the UD projet since the version 1.0 (January 2015); data are taken from the Google dataset (McDonald et al., 2013) where annotations where verified by one annotator. It was later converted into a UD version which has not been manually corrected systematically. Nevertheless, the data were corrected and enriched in later versions. UD_FRENCH-SEQUOIA is part of the UD project since version 2.0 (March 2017). It was automaticaly converted from the Sequoia

[9]Note that for the secondary predicate construction *X demonstrates Y to be Z*, the direct object *Y* is not a semantic argument of the verb. Hence the dependency between *demonstrated* and its canonical object *charges* should be dropped in a semantic representation.

[10]These are the unaccusative verbs, which use *être (to be)* tense auxiliary instead of *avoir (to have)*.

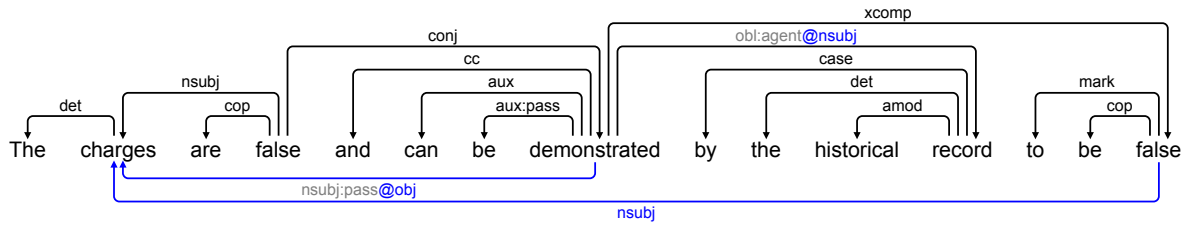

Figure 13: Enhanced UD graph, with neutralization of syntactic alternation: example with interaction of coordination, passive and predicative complement.

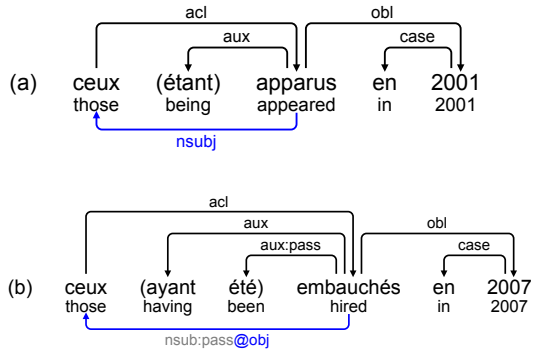

Figure 14: Noun modified by a participial phrase, with or without auxiliary

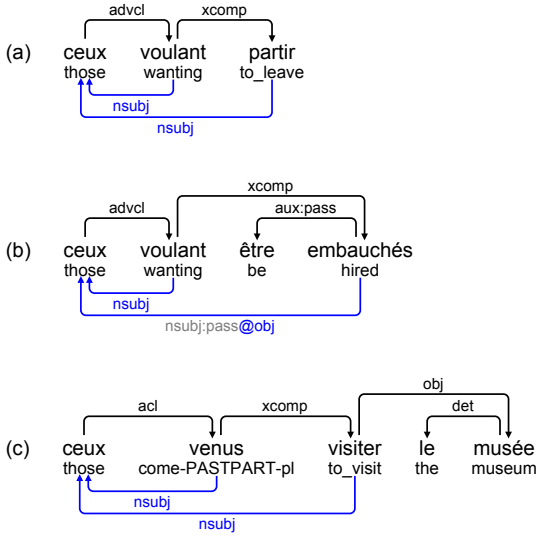

Figure 15: Subject-control verbs (necessarily active): their canonical subject is the final subject of the infinitive.

corpus[11] (Candito and Seddah, 2012) but the result was not manually corrected.

We developed two sets of rules, using two conceptually different graph rewriting systems[12], so that an adjudication of two outputs could be done.

As pointed in section 4, the full processing of syntactic alternations requires to disambiguate the argumental status of some complements: (a) which *par*-phrases are agents of passives, (b) which instances of the reflexive clitic *se* correspond to an alternation promoting object to subject, and (c) which complement of a causative complex predicate *faire*+Infinitive correspond to the subject of the infinitive.

For the Sequoia corpus, all this information is already annotated in the original corpus, and we simply had to report it on UD_FRENCH-SEQUOIA. For UD_FRENCH, we manually annotated our TEST data for the three kinds of information listed above. In the full UD_FRENCH, the number of occurences to disambiguate are: 766 for (a), 635 for (b) and 519 for (c).

[11] http://deep-sequoia.inria.fr

[12] The GREW system (Guillaume et al., 2012) and the OGRE system (Ribeyre et al., 2012)

5.1 Evaluation gold corpus

For evaluating the rule-based systems, we produced a reference evaluation corpus, containing 200 sentences not used for tuning the rules (half from UD_FRENCH (UD_{test}) and half from UD_FRENCH-SEQUOIA (SEQ_{test})). The gold enhanced graphs were obtained in three steps: (1) application of the two rule-based systems on the gold UD trees, (2) manual adjudication of the two ouputs and (3) systematic check of infinitive verbs, past or present participles and coordinations.

Below, we consider two sets of edges: N is the set of new edges, mostly argument of verbs (drawn in blue and above words in our figures) and A the set of edges impacted by an alternation (namely with a canonical function different from the final grammatical function and labeled with the '@' symbol in figures). Note that these two sets are not disjoint (see for instance, Fig. 14b).

In the reference data, N represents 5.72% of the

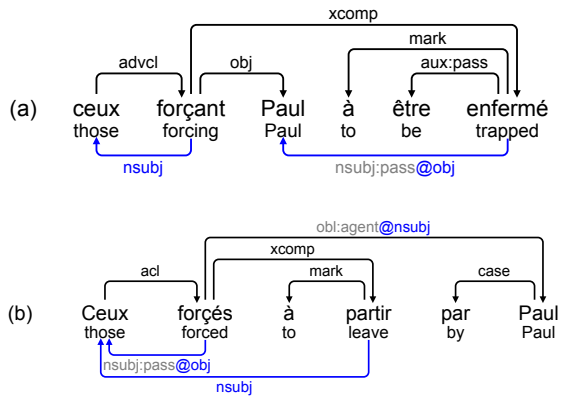

Figure 16: Object-control verb used in active and passive voice: their canonical object is the final subject of the infinitive

		PA−		PA+	
		SEQ_{test}	UD_{test}	SEQ_{test}	UD_{test}
All edges	OGRE	98.81	99.17	99.46	99.40
	GREW	99.44	99.54	99.69	99.66
$N \cup A$ edges	OGRE	86.20	89.89	92.51	91.71
	GREW	93.42	94.31	95.77	95.39

Table 1: Evaluation of rule-based systems producing enhanced graphs: F-measures computed on all edges (top) or only on edges in N or A (bottom); PA− and PA+ are respectively without and with manual pre-annotation to help syntactic alternation disambiguation.

total number of edges in the 200 test sentences. If we consider arguments of verbs only (the set of core arguments of verbs and the `obl` relation), edges in N represents 18.93% of the total number of verb arguments. The edges in set A are 2.77% of the total number of edges the full test data. Again, if we consider arguments of verbs only, these edges represent 13.91% in the 200 test sentences.

5.2 Results and Error Analysis

We evaluated the production of enhanced UD graphs in two settings, depending on whether the input UD trees do (PA+) or do not (PA−) contain manual disambiguation of cases (a), (b) and (c) described above. For the PA− case, we applied basic default rules instead, known to use insufficient information. Table 1 reports the F-measures (computed considering all edges or $N \cup A$ edges only). These results confirm the validity of our approach and highlight the consistency of the resulting graphbanks. Moreover, even if manual pre-annotations are required in theory, we empirically observe that they concern a small number of cases and their effect is marginal (the difference between PA− and PA+ settings is low).

The error analysis shows that the GREW and OGRE systems have different weak points. Of the 52 errors produced by OGRE, 30 were due to a lack of distribution of the governor or dependents on the conjuncts of a coordination, while it missed 5 subjects of infinitives only. For GREW, the result is opposite. Only 4 errors out of 28 relate to the distribution of dependencies within a coordinated structure but 14 correspond to missing subjects of infinitives. These divergences indeed helped to improve the adjudicated gold version, and were further used to improve both rule sets.

6 Discussion and Related Works

Since the rise of large annotated corpora and given the cost of annotations of large scale project such as the PDT (Böhmová et al., 2003), methods aiming at automatically enriching syntactic trees with deeper structures have peaked a decade ago (Hockenmaier, 2003; Cahill et al., 2004; Miyao and Tsujii, 2005) but have then been subsumed by purely data-driven methods when corpora with richer annotation have been made available (Hajic et al., 2006; Oepen et al., 2014; Mille et al., 2013). Space is missing for an in-depth comparison between these different annotation scheme, we refer the reader to (Rimell et al., 2009; Ivanova et al., 2012; Candito et al., 2014; Kuhlmann and Oepen, 2016) for a more complete overview. Here, we will focus on the differences between the Meaning Text Theory (MTT, (Melčuk, 1988)), as instanced in the recent AnCora-UPF treebank (Mille et al., 2013; Ballesteros et al., 2016), and our proposal.

The MTT defines an explicit deep syntactic representation level[13], hereafter DSyntS. The AnCora-UPF Treebank follows its four layer model: morphological, surface-syntactic, deep-syntactic and semantic. The method used for annotating that corpus is similar to the procedure we used. Starting from the surface-syntactic level, the two other levels are automatically pre-annotated step by step: the annotation of a given level is rewritten to the next level using the MATE tools (Bohnet et al., 2000).

[13] Kahane (2003) proposed to view the deep syntactic representation as a derivation step between surface syntax and semantic representation.

The DSyntS produced by Ballesteros et al. (2016) share important properties with our extented enhanced UD graphs, in that they neutralize syntactic alternations. However, they do not contain additional arcs for argument sharing, as subjects of infinitives for instance, as they stick to tree structures. Besides the choice of representation structures, graphs in our cases, trees in the other, important differences remain: Another difference concerns the dependency labels for arguments: canonical function labels (nsubj, obj etc...) in our case versus "argument relations" for MTT, namely numbers (I, II, III etc...), ordered using a "growing obliquity" order (Iordanskaja and Melcuk, 2000). These numbers do not have a meaning per se, and are intended to be read within a lexical entry linking them to syntactic realizations. We note that using argument numbering in a deep syntactic representation, hence in the absence of word sense disambiguation, leads to the loss of plain syntactic information useful for disambiguation. For example in French: *apprendre* is ambiguous between *to learn* as in *X apprend Y de Z*, and *to teach*, as in *X apprend Y à Z*. Both senses entail different subcategorization frames *(subj, obj, obl:de)* vs *(subj, obj, obl:à)*, but bear the same argument numbers in the MTT (I, II, III), the meaning of III being too underspecified in the absence of semantic disambiguation[14].

7 Conclusion

We proposed extensions of the current enhanced universal dependencies scheme. We advocated in particular for neutralizing syntactic alternations, in order to limit the diversity of observed subcategorization frames for a given verb, while staying at the syntactic level, without resorting to word sense disambiguation. We implemented rule-based modules to obtain enhanced graphs from French UD trees. Evaluation on a 200-sentence sample shows we obtain over 90% of F-measure on the enhanced edges (edges not present in the input UD tree). Moreover, we report a 19% proportion of enhanced edges among the edges for arguments of verbs, meaning that the saturation of predicate-argument structures for verbs concerns a non negligible amount of arguments. We hope this proposal can be tested on other languages, the most obvious ones being the Romance languages, which show very similar syntactic alternations.
We position this proposal within the UD framework and remain compatible with all choices already made by the current specifications (Nivre et al., 2016; Schuster and Manning, 2016). Moreover, our de-facto adhesion to the CoNLL-U representation format allows for a straight-forward use by current data-driven graph parsers. We leave this promising path of study to further work.

[14] One of the anonymous reviewers pointed that because in UD some labels are distinguished according to the category of the dependent (e.g. `nsubj` vs. `csubj`), the MTT labels would still better account for linking regularities. While we do agree that the UD label distinctions multiply linking patterns maybe uselessly, we believe that on the other hand, the MTT deep labels do add ambiguity, and are thus insufficient per se.

Acknowledgments

We warmly thank our anonymous reviewers for their insightful comments. The first and the last authors were partly funded by the ANR projects ParSiTi (ANR-16-CE33-0021), SoSweet (ANR-15-CE38-0011-01) and supported by the Program *Investissements d'avenir* managed by the Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

References

Anne Abeillé, Danielle Godard, and Philip Miller. 1997. Les causatives en français, un cas de compétition syntaxique [in french]. *Langue française*, 115(1):62–74.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Miguel Ballesteros, Bernd Bohnet, Simon Mille, and Leo Wanner. 2016. Data-driven deep-syntactic dependency parsing. *Natural Language Engineering*, 22(6):939–974.

Karine Baschung. 1996. Une approche lexicalisée des phénomènes de contrôle [in french]. *Langages*, 30(122):96–123.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Bernd Bohnet, Andreas Langjahr, and Leo Wanner. 2000. A development environment for an mtt-based sentence generator. In *Proc. of the First International Conference on Natural Language Generation*, INLG '00, pages 260–263.

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith, and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired

Wide-Coverage PCFG-Based LFG Approximations. In *Proc. of ACL*, pages 320–327.

Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN*.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric De La Clergerie. 2014. Deep Syntax Annotation of the Sequoia French Treebank. In *In Proc. of LREC*, Reykjavik, Islande, May.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets universal dependencies. In *15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, Indiana University, US.

Marie-Catherine de Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. In *Proc. of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140, Berlin, Germany, August.

Bruno Guillaume, Guillaume Bonfante, Paul Masson, Mathieu Morey, and Guy Perrier. 2012. Grew : un outil de réécriture de graphes pour le TAL. In *Proc. of TALN*, Grenoble, France.

Jan Hajic, Jarmila Panevová, Eva Hajicová, Petr Sgall, Petr Pajas, Jan Štepánek, Jiří Havelka, Marie Mikulová, Zdenek Zabokrtskỳ, and Magda Ševcıková Razımová. 2006. Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

Julia Hockenmaier. 2003. *Data and models for statistical parsing with Combinatory Categorial Grammar*. Ph.D. thesis.

Lidia Iordanskaja and Igor Melcuk. 2000. The notion of surface-syntactic relation revisited (valence-controlled surface-syntactic relations in french). *Slovo v tekste i v slovare. Sbornik statej k semidesjatiletiju Ju.D. Apresjana, Moskva: Jazyki russkoj kul'tury*, pages 391–433.

Angelina Ivanova, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger. 2012. Who did what to whom?: A contrastive study of syntacto-semantic dependencies. In *Proc. of the 6th Linguistic Annotation Workshop (LAW-VI 2012)*, pages 2–11.

Sylvain Kahane. 2003. On the status of deep syntactic structure. In *Proc. of the First Meaning-Text Theory conference*, Paris, France.

Kevin Knight, Lauren Baranescu, Claire Bonial, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, and Nathan Schneifer. 2014. Abstract meaning representation (amr) annotation release 1.0. *Web download*.

Marco Kuhlmann and Stephan Oepen. 2016. Towards a catalogue of linguistic graph banks. *Computational Linguistics*, Volume 42, Issue 4, December.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97.

Igor Melčuk. 1988. *Dependency syntax: theory and practice*. State University Press of New York.

Olivier Michalon, Corentin Ribeyre, Marie Candito, and Alexis Nasr. 2016. Deeper syntax for better semantic parsing. In *Proc. of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 409–420, Osaka, Japan, December.

Simon Mille, Alicia Burga, and Leo Wanner. 2013. AnCoraUPF: A Multi-Level Annotation of Spanish. In *Proc. of DepLing 2013*.

Yusuke Miyao and Jun'ichi Tsujii. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proc. of ACL 2005*, pages 83–90.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proc. of LREC 2016*, pages 1659–1666.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. 2014. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proc. of the 8th International Workshop on Semantic Evaluation*, pages 63–72.

Guy Perrier, Marie Candito, Bruno Guillaume, Corentin Ribeyre, Karën Fort, and Djamé Seddah. 2014. Annotation scheme for deep dependency syntax of french (un schéma d'annotation en dépendances syntaxiques profondes pour le français) [in french]. In *Proc. of TALN 2014 (Volume 2: Short Papers)*, pages 574–579, Marseille, France, July.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. *arXiv preprint arXiv:1702.03196*.

Corentin Ribeyre, Djamé Seddah, and Éric Villemonte De La Clergerie. 2012. A Linguistically-motivated 2-stage Tree to Graph Transformation. In Chung-Hye Han and Giorgio Satta, editors, *Proc. of TAG+11*, Paris, France. INRIA.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proc. of EMNLP*, pages 813–821.

Manuela Sanguinetti and Cristina Bosco. 2014. Parttut: The turin university parallel treebank. In Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, editors, *Harmonization and development of resources and tools for Italian Natural Language Processing within the PARLI project*. Springer Verlag.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proc. of LREC 2016*. Portorož, Slovenia.