Deep Syntax Annotation of the Sequoia French Treebank

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karen Fort, Djamé Seddah, Éric Villemonte de La Clergerie

Deep Syntax Annotation of the Sequoia French Treebank

Marie Candito*, Guy Perrier†, Bruno Guillaume‡,
Corentin Ribeyre*, Karën Fort†, Djamé Seddah◇, Éric de la Clergerie*
* Université Paris Diderot/INRIA †Université de Lorraine/LORIA
‡Inria Nancy Grand-Est/LORIA ◇Université Paris Sorbonne/INRIA

Abstract

We define a deep syntactic representation scheme for French, which abstracts away from surface syntactic variation and diathesis alternations, and describe the annotation of deep syntactic representations on top of the surface dependency trees of the Sequoia corpus. The resulting deep-annotated corpus, named DEEP-SEQUOIA, is freely available, and hopefully useful for corpus linguistics studies and for training deep analyzers to prepare semantic analysis.

1. Introduction

In the last decade, dependency syntax has seen a surge of interest in Natural Language Processing (NLP), partly due to the availability of efficient dependency parsing algorithms (see (Kübler et al., 2009) for an overview) and also to the fact that even for a language with relative fixed word order like English, dependency trees allow for a more direct extraction of predicate-argument structures, which associate a predicate to its semantic arguments.

However, because predicate-argument structures can be realized in many surface syntactic ways, a typical solution for providing more semantically-oriented information is to define "deeper" representations, "deeper" having a different meaning across different linguistic frameworks. In this paper, we focus on providing both a deep annotation scheme and a deep-annotated corpus for French. We define a deep level of representation that abstracts away from surface syntactic variation by making explicit the canonical subcategorization frame of predicates, and which linguistic expressions fill it. The deep annotations were added to the surface annotations of the Sequoia corpus (Candito and Seddah, 2012b), following preliminary work by Bonfante et al. (2011). The resulting corpus is freely available[1].

In section 2., we provide the theoretical characteristics of our Deep Semantic Representations (DSRs), built on top of an existing surface dependency annotation scheme derived from that of the French Treebank (FTB) (Abeillé and Barrier, 2004). Examples taken from the corpus illustrating some of the main difficulties are detailed in section 3. We describe in section 4. the methodology we adopted for the annotation of the corpus. We then compare our work with respect to other deep annotation schemes or corpora and conclude in sections 5. and 6.

2. Deep Syntactic Representations

2.1. Building on the Surface Dependency Scheme of the FTB

We define the DSRs on the Sequoia corpus (Candito and Seddah, 2012b; Candito and Seddah, 2012a), which contains 3,099 sentences covering several domains (news, medical, europarl and fr-wikipedia).[2] These sentences were already annotated for surface dependencies. The underlying surface dependency scheme influenced our DSRs scheme in various ways, it is therefore important to detail how this surface dependency scheme is defined: the Sequoia corpus was first annotated using constituency trees, following the annotation scheme of the FTB (Abeillé and Barrier, 2004), except that a slightly more specific set of dependency labels was used for oblique complements, and that only functional compound words (e.g. *bien que ('although')*) or syntactically irregular compound words are annotated as such (syntactically regular compound words are not marked as such, and are represented with regular internal structure).

It was then automatically converted into projective surface dependencies using the procedure described in (Candito et al., 2010). As such, the obtained dependency trees follow the representation choices made for the FTB (Abeillé et al., 2004; Abeillé, 2004), in the sense that the vast majority of linguistic phenomena are mechanically translated into dependencies. Additional information predicted by the conversion procedure itself concerns some dependency labels, and additional structure for cases in which a single constituent includes several heads. Further, Candito and Seddah (2012a) corrected dependencies in case of wh-extraction, to account for long-distance dependencies, introducing some non-projective arcs.

The resulting surface dependency scheme is our starting point to define the DSRs. We made the pragmatic decision of minimizing the changes introduced at the deep syntactic level, to concentrate on phenomena not representable in surface. This impacted the DSR schemes. For instance, we kept the representation of coordinating structures with the first conjunct as the head[3], although it is clearly problematic for the representation of shared modifiers.

2.2. Linguistic Characteristics

The overall objective of our DSR is to abstract away from syntactic variation by making explicit which expressions fill the *canonical subcategorization frame* of predicates. In order to precisely define this notion, let us first recall that grammatical functions (GF) can be defined as sets of syn-

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

[2] The 4.0 version on top of which we annotated deep syntax contained 3,200 sentences, but we removed a duplicate extract of 101 sentences.

[3] All the subsequent coordinating conjunctions attach to the first conjunct, and the non-first conjuncts attach to their preceding coordinating conjunction (see e.g. Sentence 3 in Figure 1).

tactic properties imposed by lexemes (in particular verbs) to their semantic arguments. Yet, it is well-known that the same lexeme can occur in constructions that vary in the way semantic arguments are linked to GFs, and this kind of variation also shows regularities known as syntactic alternations or diathesis alternations. In order to capture such regularities without resorting to semantic properties or thematic roles but sticking to syntactic generalizations, we use the distinction, inspired by Relational Grammar (Perlmutter, 1983), between *canonical* grammatical function (canonical GF) and *final* grammatical function (final GF)[4], and between *canonical subcategorization frames* (canonical SF) and *final subcategorization frames* (final SF).
We define the final subcategorization frame of an occurrence of a verb as the list of GFs associated to its expressed arguments, plus the GFs that would be associated with the linguistic expressions that would appear as argument, if the verb were used in finite mode and in a non elliptical construction. This formulation accounts for the subject of infinitives, the subject of coordinated verbs or more generally any argument shared by several predicates. For instance, in *Jean veut partir mais doit rester ('Jean wants to-leave but has to-stay')*, the final subcategorization frame for *partir* is [subject] associated with *Jean*, and the final subcategorization frame of *doit* is [subject, object], filled by *Jean* and *rester*. We call deep syntactic arguments of a verb the set of linguistic expressions that bear a final GF with respect to that verb.

Among the linguistic expressions that bear a final GF, we retain as deep syntactic arguments the ones that are semantically non-empty. To neutralize surface syntactic variation due to diathesis alternations, we view these as redistributions of the grammatical functions associated to the syntactic arguments. Following Relational Grammar (Perlmutter, 1983), we view the final SF as resulting from the application of 0 to n redistributions to a *canonical subcategorization frame*. A simple case is for instance a passive occurrence of a transitive verb: the final subcategorization frame is [subject, by-object] while the corresponding canonical SF is [objet, subject]. The set of alternations and their allowed combinations is language-specific[5]
We only considered redistributions that are morpho-syntactically marked (for instance with an auxiliary for passives, or a void reflexive clitic *se* for middle or neuter alternations). Unmarked redistributions are not accounted for (because disambiguating them, in the absence of marking, resorts to semantic analysis). For instance, for the verb *couler ('to sink')*, the non-marked causative/inchoative alternation gives rise to two canonical SFs: the two constructions *X coule Y (X sinks Y)* and *Y coule (Y sinks)* are not related in the deep syntactic representation. They get the two distinct canonical SF [subject, object] and [subject] respectively, and for both occurrences, the canonical SF is identical to the final SF.[6] On the contrary, for the neuter and middle alternations, which are marked by a void reflexive clitic *se*, are represented in our DSRs using redistributions. For instance, for both *(Paul cassa la vase) Paul broke the vase* and *le vase se brisa* (litt. *the vase SE broke* for *the vase broke*), *vase* is canonical object.
The notion of canonical versus final GF interacts with the explicitation of the deep syntactic arguments. Take for example the case of a control verb introducing a passive infinitival verb: *Paul veut être embauché (Paul wants (to) be hired)*. In the surface syntactic representation, *Paul* is the final subject of *veut*. In the deep representation, it is also the *final* subject of *être embauché (be hired)* and its canonical object. We capture the regularity that control verbs control the *final* subject of the infinitival verb they introduce, independently of the diathesis of the infinitival verb.
Our work currently focuses on verbal and adjectival predicates, leaving the other open classes as future work. For an adjective, we use 'subject' as canonical GF for the first semantic argument of the adjective (that is the noun the adjective modifies when it is attributive), even though this argument never shows as a syntactic dependent of the adjective.

To sum up, our DSRs make explicit three major types of information with respect to the surface representations:

- the semantic status of the words in the sentence, whether semantically void or not;
- the full list of deep syntactic arguments of a predicate, including those arguments that are syntactically dependent of another head (e.g. the subject of infinitival verbs) or that appear as the surface governor of the predicate (e.g. in the case of an attributive participle: *des personnes parlant italien ((some) people speaking italian)*)
- syntactic information that remains stable across diathesis alternations (i.e. we neutralize diathesis alternations).

In our DSRs, the canonical SF of each occurrence of verb or adjective is identified, along with the elements that fill it, which are directly attached to the predicate using the canonical GF as label. More precisely, we take as argument head the higher content word (semantically empty functional words, such as void complementizers and empty prepositions, are shunted off).

2.3. Deep Syntax versus Semantic Representations

The DSRs we define differ from semantic representations in various ways. The most important feature is that in the DSRs, the meaning of ambiguous lemmas is not disambiguated. The semantics of predicates is only used to disambiguate syntactic attachments. Second, while semantically empty nodes do not appear in the DSRs, the remaining nodes do not necessarily form a semantic unit: polylexical expressions (idioms, light verb constructions, syntactically

[4]We use the term *canonical* instead of the RG term *initial*.

[5]See (Candito, 1999) for an account of alternations and their combinations for French. Sentence 7 in figure 2 illustrates a case of passive plus impersonal alternations.

[6]In other words, we do not retain unaccusative properties in order to define canonical SF.

regular compounds) are not marked as such, and are represented with regular internal syntactic structures.
Moreover, the elements filling the canonical SF of a predicate are semantically non-void (hence the expletive *il ('there')* does not bear a canonical grammatical function), but are not necessarily semantic arguments of the predicate. So for instance, the surface subject of a raising verb belongs to its canonical SF, although it is not a semantic argument of the raising verb.
Further, DSRs use canonical grammatical functions, instead of either semantic roles or argument numbering that are typically found in semantic representations. This is coherent with the fact that the lemmas are not disambiguated. Grammatical functions are important clues for (further) disambiguation, and cannot be discarded at this stage. For instance, consider the two related senses of *parler ('to speak')* in *Paul parle italien (Paul speaks italian)* and *Paul parle de l'Italie (Paul is speaking about Italy)*: while a semantic-oriented representation focusing on argument structures (such as PropBank (Palmer et al., 2005)) would distinguish *parler_sense1(arg0=Paul, arg1=italien)* and *parler_sense2(arg0=Paul, arg1=Italie)*, at the level of deep syntax, we only make explicit two different canonical subcategorization frames: *parler(subject=Paul, direct_object=italien)* versus *parler(subject=Paul, de_object=Italie)*.
In order to obtain argument structures from of our DSRs, it is necessary to disambiguate predicative lemmas, keeping only true semantic arguments of predicates, and to associate them with semantic roles or a simple number. Argument numbers can be obtained using the oblicity order of canonical grammatical functions (tipically subject $<$ object $<$ indirect objects etc...). Note though that if one wanted to obtain PropBank-style argument structures, canonical subjects of intransitive verbs should be further disambiguated into Arg0 or Arg1, following the proto-Agent versus proto-Patient distinction proposed by Dowty (1991).

2.4. Formal Properties

We define a *complete* representation as a dependency graph containing both the surface syntactic representation (SSR) and the DSR. Examples from the corpus are shown in Figures 1 and 2. Nodes are the words of the sentence, typed as semantically void (red words in the figures) or not (black words). Arcs carry:

- two boolean types, for surface or not, and deep or not, which combine into three values only: an arc can be surface but not deep (red arcs), deep but not surface (blue arcs) or both deep and surface (black arcs);
- a final grammatical function (final GF);
- a canonical GF, in the case of a final GF that can be involved in diathesis alternations.

The DSR is made of the semantically non-void nodes only, and the deep arcs. It may contain cycles. The SSR is a tree made of all the word nodes and all the surface arcs, labeled with the final GFs only.
In all the following, as well as in the figures, we refer to an arc with final GF x and canonical GF y as an "x:y" arc. For each sentence, the SSR is shown above the sentence, and comprises the surface-and-deep arcs (in black) and the surface-only arcs (in red).[7] The DSR is made of the deep-only arcs (in blue below the sentence) and the black arcs above.

3. Examples

The examples in Figures 1 and 2 focus on significant phenomena and illustrate the complex interaction between the addition of deep syntactic arguments, the redistribution between canonical and final GFs and the deletion of semantically empty words.

Auxiliaries The DSR scheme contains deep features on top of the morphological features defined in the surface scheme. In particular, all auxiliaries (for compound tenses, passives and causatives) are attached to the content verb in the SSR (e.g. the tense auxiliary *a* and passive auxiliary *été* attach to the participle *observé* in sentence 7) and are replaced in SSR by deep tense and/or diathesis features on the content verb.

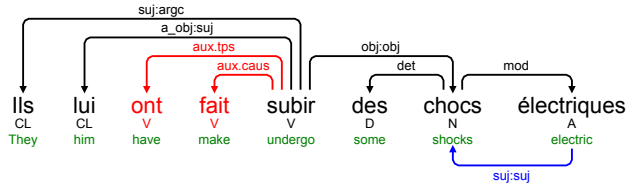
Causatives Sentence 1 illustrates a causative construction, analysed as a diathesis alternation that adds a 'causer' semantic argument to a verb *ils (they)*, with subject as final GF, and demotes the canonical subject to object, a-object or par-object of the verb, depending on its transitivity. In sentence 1, the causer is *ils (they)*, final subject and canonical *argc* of the verb *subir*, while the canonical subject *lui (him)* is final indirect object a-object, hence the label a-obj:suj.

Reflexives The reflexive clitic *se* was annotated using four main classes. We only detail two of these, due to a lack of space. Sentence 2 examplifies a true reflexive. In the SSR, *se* is object of *déclare*, the co-reference with the subject *juge* is not marked. In the DSR, *se* is shunted off, and any incoming arcs are redirected to *juge*.
The clitic *se* can also mark diathesis alternations in which a canonical object shows up as final subject. The semantic argument corresponding to the canonical subject is either semantically interpretable (middle constructions) or absent (neuter constructions, as for *se cassent* in sentence 3). A clear-cut distinction between both revealed difficult to annotate, so we currently use the same DSR for both, in which *se* is discarded and replaced by a specific diathesis feature on the verb, and the final subject of the verb is its canonical object. For instance in sentence 3, *os* is the suj:obj of *cassent*. Moreover, that suj:obj arc is only present in the DSR, because the final subject *os* is shared between the two coordinated verbs at the deep level, but attached to the first verb conjunct only at the surface level.

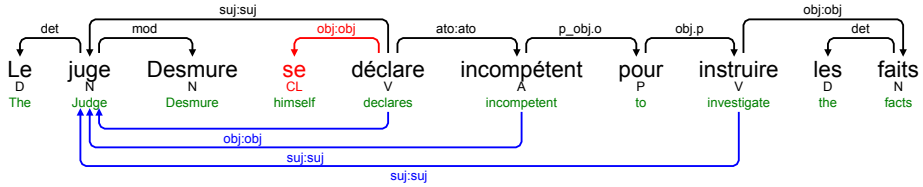
Adjectives All adjectives but cardinal ones get a subject in the DSR, which is both canonical and final (no diathesis alternation for adjectives). The subject of an attributive adjective is the noun it modifies. Predicative adjectives are attached (in SSR and DSR) to a verb, with the lable ats or atoA, depending on whether their subject is the subject or the object of the verb. For instance, in sentence 2, *incompétent* has a deep subject *juge* and it is a predicative complement of the verb *déclare* referring to the object *se*. Since

[7] More precisely, the canonical GFs (the :y suffixes of the labels) do not belong to the surface dependency trees.

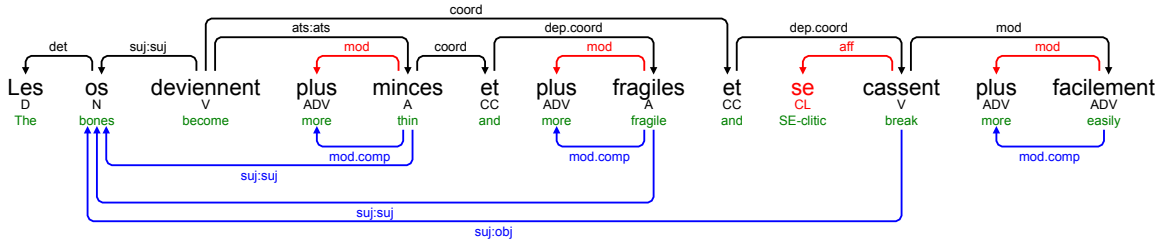
1: (Causative alternation) "*They subjected him to electric chocs* "



2: (Reflexive; infinitival verb subject) "*Judge Desmure declares himself incompetent to investigate the facts*"



3: (Subject ellision; Comparatives; Neuter diathesis alternation) "*The bones become thiner and more fragile and they break more easily*"



4: (Passive; Raising verb) "*The structural works package will probably have to be declared unsuccessful*"

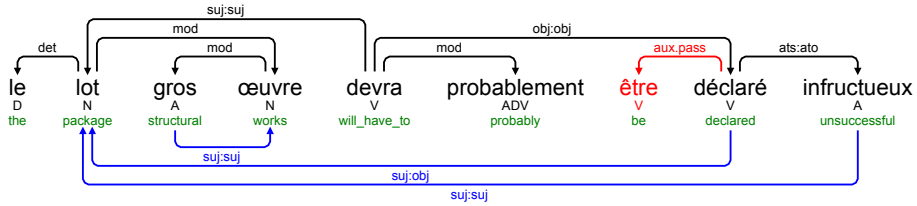


Figure 1: Examples of annotated sentences from the corpus (features are omitted). For each sentence, the surface syntactic representation appears above the sentence, and the deep syntactic representation is made of the top black and bottom blue arcs. Single labels are final GFs, whereas in a double label x:y, x and y are the final GF and canonical GF, respectively. Semantically void tokens are in red.

there is no redistribution, the canonical and the final GFs of the adjective are the same (GF *ato*).

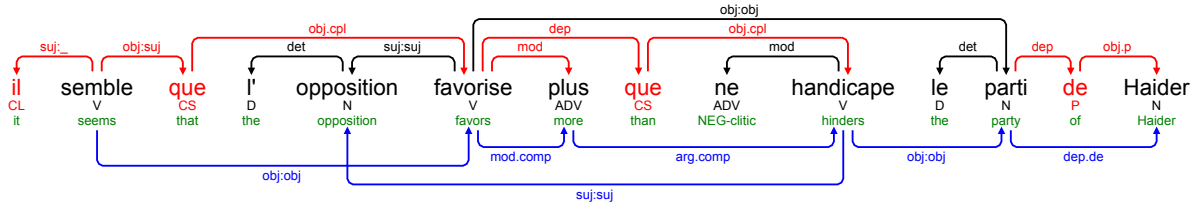
In sentence 4, the adjective *infructueux* is a predicative complement of *déclaré* but, because of a passive redistribution, there is a change in the GF of the adjective: in its canonical GF, it refers to the object of the verb (GF *ato*), and in its final GF, it refers to the subject (GF *ats*). This final subject is a deep subject because of the presence of the modal auxiliary *devra*.

Impersonal subjects The semantically void expletive *il* can appear either as (final) subject of 'essentially impersonal verbs' or in case of impersonal diatheses of verbs that can also appear with referential subjects. Sentence 5 illustrates the former case, with *semble* as essentially impersonal verb. No redistribution is involved, therefore the other dependents have identical final and canonical GFs.

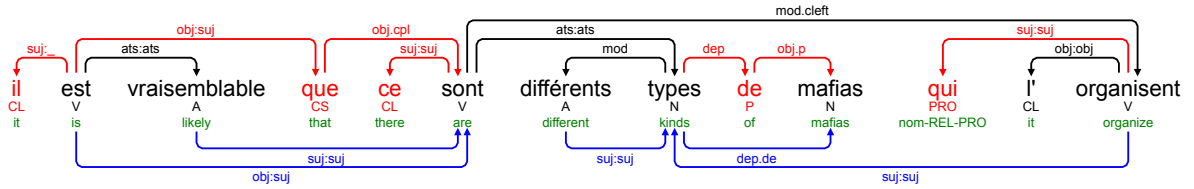
Sentences 6 and 9 illustrate two cases of impersonal diathesis. In sentence 6, the impersonal alternation demotes the canonical subject of the copula *est* to final object. That final object is sentential introduced by the semantically void complementizer *que*, which is ignored in SSR, hence the deep arc obj:suj between *est* and the verb introduced by *que* (*favorise*). Sentence 9 illustrates an impersonal passive, analyzed as the sequence of passive and impersonal alternations. The canonical object *effet* is promoted to subject (by passive) and re-demoted to object (by impersonal), hence the label obj:obj.

Coordination In accordance with the FTB scheme, the head of a coordinating structure is the head of the first conjunct. The coordinating conjunction is linked to the head with a *coord* label and the head of the second conjunct is linked to the conjunction with a *dep.coord* label. These de-

5: (Impersonal verb; Comparative) "*It seems that the opposition favors rather than hinders Haider's party*"



6: (Impersonal diathesis alternation; Cleft clause) "*It is likely that there are different kinds of mafia organizing it*"



7: (Passive plus impersonal diathesis alternation; noun-modifying participial clause) "*No unwanted effect resulting from physiological stress was observed*"

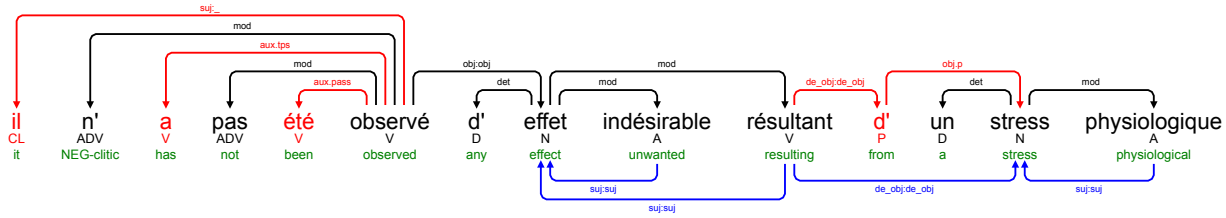


Figure 2: Examples of annotated sentences from the corpus (see legend of Figure 1).

pendencies are present in the SSR as well as in the DSR. Now in the DSR, we deal with incoming dependencies of a coordination differently from the outgoing dependencies, for semantic considerations. With respect to the incoming dependencies, a coordination is regarded as a whole and the dependencies are not distributed between the conjuncts (i.e. disambiguation between collective and distributive readings is not done). On the contrary, *dependents* of a coordinated elements, which are linked to the first conjunct in the SSR, are distributed to the other conjuncts in the DSR, except for cases not detailed here. For instance, in sentence 3, the (deep) subject of the adjective *mince* is distributed to the coordinated adjective *fragile*, while only the first conjunct *mince* depends on the verb *deviennent* (as a predicative complement, with label ats). Sentence 3 illustrates coordination of two finite verbs (i.e. subject ellision or coordination of VPs): *os* is added as suj:obj of the second verb conjunct *cassent*.

Cleft clauses Sentence 6 includes the cleft clause *ce sont différents types de mafias qui l'organisent* where the subject of *organisent*, the noun phrase *différents types de mafias* is extracted from the canonical sentence *différents types de mafias l'organisent* and put as predicative complement of *sont* (GF *ats*) to be highlighted. In the canonical sentence, the trace of the extracted subject is represented with the relative pronoun *qui*, and the sentence becomes a relative clause depending on the verb *sont* with a dependency *mod.cleft*. In the DSR, the canonical sentence is not totally restored and the dependency *mod.cleft* is preserved to allow for the expression of modifier adjunction to the verb *sont* (as for instance in *ce ne sont pas différents types de mafias* . . .) . Only the pronouns *ce* and *qui* are removed and *types* becomes the deep subject of *organisent*.

Comparatives Sentence 5 includes a comparative construction. According to the FTB scheme, in the SSR, the head of the second term of the comparison, the complementizer *que*, depends on the head of the constituent including the first term of the comparison, the verb *favorise*, with a dependency *dep*. In the DSR, the dependency is renamed in *arg.comp* and its governor moved to the adverb *plus* expressing the degree of comparison. The complementizer is removed and the target of the dependency is moved to the head introduced by *que* (the verb *handicape*). Like for coordinated structures, the dependents of a comparative construction are distributed over the two compared terms. This is the case here for the subject *opposition* and the object *parti*.

4. Corpus Annotation Methodology

The deep annotation scheme described and illustrated in sections 2. and 3. was devised to annotate the deep syntactic layer of the Sequoia corpus, starting from version 4.0.[8] During the deep annotation phase, the annotators were allowed to correct the surface annotation, assuming they followed the surface annotation scheme. Systematic "errors" inherent to the surface annotation scheme are captured at the deep level, without modifying the surface annotation.

[8] https://gforge.inria.fr/frs/?group_id=3597

The different steps of the annotation process were conducted in a collaborative way. As the members of the project were located in two different French towns (Paris and Nancy), we decided to (i) iteratively and collaboratively produce the annotation guide and annotations for a small subset of the corpus, (ii) to independently produce a complete annotation of the corpus in both towns and (iii) to collaboratively adjudicate the two results.

4.1. Mini Reference and Annotation Guide

At the beginning of the project, we produced a "mini reference" deep annotated corpus: we randomly selected 250 sentences from the Sequoia corpus, and annotated them in parallel to the production of the annotation guide, in order to get feedback for the guide, following the methodology presented in (Fort, 2012). Each team separately produced an annotated version of the mini reference. Then, several iterations of discussions (with phone meetings) and revisions of the annotation led to the production of the final version.

4.2. Tools used for Annotation

Three tools were used in the project: O.G.R.E. for the pre-annotation in Paris, GREW for the pre-annotation in Nancy (using a set of rules adapted from (Bonfante et al., 2011)) and DepAnnotator on both sides, for manual annotation.

- O.G.R.E.[9] is a two-stage graph rewriting system which addresses the problems of rules interaction, confluence and long-distance dependencies rewriting by using propagation constraints on edges (Ribeyre et al., 2012).
- GREW[10] (Guillaume et al., 2012) is a graph rewriting tool dedicated to NLP applications: rules of the system are organized in modules that allow for a finer control on rule applications.
- DepAnnotator[11] is a cross-platform annotation tool developed to fit the requirements of the new deep syntax annotation scheme. DepAnnotator also allows for the use of any other annotation schemes thanks to configurable XML files.

4.3. Annotation Campaign

As mentioned earlier, the complete resource, minus the mini-reference, was then annotated both in Paris and in Nancy. The annotation process for each location differs on several aspects. First, the pre-annotation used two different tools relying on two different sets of rules. Second, the level of expertise of the annotators was different:

(i) In Nancy, three students in linguistics without any previous annotation experience worked on the annotations (200h each). Starting from the pre-annotated version of the corpus, the work of the annotators was split into several steps. At each step, the focus was put on a particular phenomenon (locatives, subjects of adjectives, etc), with an expert giving the annotators a specific training on the phenomenon. Once the corpus annotated, the students were asked to produce a document about their annotations, the main problems they encountered, and the way they actually worked.
(ii) In Paris, annotation was performed by skilled annotators, who had already worked on the surface annotation of the same corpus. The causatives and all the reflexive clitics were manually annotated before pre-annotation with O.G.R.E. Then, the annotators checked the result of the pre-annotation, working phenomenon-by-phenomenon (e.g. subjects of non-finite verbs, cleft constructions, coordination, etc...).
We believe that diversity, in terms of pre-annotation tools and expertise of the annotators, limits the risk of biases in the final annotation, in particular those introduced by pre-annotation. A crucial point was not to communicate during that phase, so that inconsistencies could show up at adjudication time.
The two annotated versions were then adjudicated: we first adjudicated together a subset, after having studied the typology of annotation divergences. Then each team in Nancy and Paris adjudicated half of the corpus.

4.4. Evaluation

To assess the reliability of the annotation, we calculated the inter-annotator agreement between the annotated set of parses (without the mini-reference) from Nancy and Paris. We computed the labeled and unlabeled F-measure on arcs, considering one annotation as reference and the other as predicted. Note first that we cannot use (un)labeled attachment scores, because we are dealing with graphs. Second, the F-measure is identical whatever annotation is chosen as reference (precision and recall are inverted).

	ALL	SURF.	SURF. *(no canon.)*	SURF. ONLY	DEEP	DEEP ONLY
LF	95.91	96.27	97.35	95.57	96.00	94.23
UF	98.33	97.79	97.79	97.28	97.48	95.77
Edges	74,051	61,173	61,173	14,816	59,235	12,878

Table 1: Inter-annotator agreement on the Sequoia Deep Treebank (minus the mini-reference), broken down by edge types, and average number of edges for each type.

Table 1 shows the evaluations, broken down by arc types: *All* stands for all edges, whether deep or not, surface or not, *Surf.* and *Deep* for the surface arcs and deep arcs respectively, and *Surf. only* / *Deep. only* for the surface non-deep arcs (red arcs) and non-surface deep arcs (blue arcs) respectively. Furthermore, evaluations in column *Surf. only (non canon.)* do not take into account the $:y$ suffixes in labels, for the canonical GFs. It thus corresponds to the pure surface dependency trees. In the last row, we provide the average number of edges in the two annotated versions, for each type of evaluation, to give a clearer view of surface and deep edges proportion. This shows that in the DSRs, deep-only arcs correspond to roughly one fifth of the edges ($12,878$ out of $59,235$). Overall, the F-measures reach around 97 and 96, which we believe shows that our thorough methodology led to a resource of acceptable quality, despite the use of pre-annotation tools. The evaluation on deep-only arcs is a bit lower (LF=94.23 / UF=95.77). This might be explained by the novelty of the deep annotation

[9] http://www.corentinribeyre.fr/projects/view/OGRE
[10] http://grew.loria.fr
[11] http://yquem.inria.fr/~ribeyre/deploy/DepAnnotator

scheme for annotators, and also by the fact that disagreement on a surface arc carries over to the deep analysis, often impacting several deep arcs.

5. Related Work

Making explicit some "deep" syntactic information in treebanks, either manually or automatically, is a usual solution for providing more semantically-oriented information likely to be useful for preparing any syntax to semantic interface. To this end, two main theoretical frameworks dominate the dependency syntax landscape and are often used to express the links between surface syntax and deep representation, namely the Prague Dependency school and the Meaning Text Theory (MTT, (Melčuk, 1988)). The former is instanced through the well-known Prague Dependency treebanks (for Czech, English and Arabic), and the latter has recently been used as the basis of a new Spanish treebank (Mille et al., 2013).

The Czech version of the Prague Dependency Treebank (PDT, (Hajic et al., 2006)), in its 2.0 version, contains about two million words. It distinguishes between analytical and tectogrammatical levels. While the former corresponds to the surface level, the former is more semantically-oriented: it makes use of semantic labels, called *functors*, such as *Actor/Bearer, Addressee,etc.*, and comprises the representation of the topic-focus articulation of a sentence, which we do not represent in our DSR. So, a key difference in our deep representation scheme is that we make central use of the canonical GFs as a proxy between surface realization and semantic roles.

MTT defines an explicit deep syntactic representation level[12], and the recent AnCora-UPF Treebank (Mille et al., 2013) follows the MTT model with four layers: morphological, surface-syntactic, deep-syntactic and semantic. The method used for annotating the corpus is similar to ours. 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). The size of the AnCora-UPF Treebank, 3,513 sentences, is of the same order as our Sequoia corpus.

For English, following the first release of the PTB (Marcus et al., 1993), further releases have been manually augmented with non-surface information (Marcus et al., 1994) on top of its context-free backbone: co-indexed null elements make explicit long-distance dependencies as well as subjects of infinitives. In addition, some diathesis alternations are captured: using our terminology, final and canonical subjects of passives bear different functional tags (SBJ and LGS respectively), and a null element in post verbal position is co-indexed with the final subject. Although less directly available than in a deep dependency representation, this deep information layer has been used to extract wide coverage deep grammars, and to build deep syntactic parsers for the CCG, LFG and HPSG frameworks (Hockenmaier, 2003; Cahill et al., 2004; Miyao and Tsujii, 2005).

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

Besides these theoretically-oriented deep syntax parsers, the Stanford Dependency annotation schemes (De Marneffe et al., 2006; De Marneffe and Manning, 2008) aim at proposing a more agnostic view of treebank-based syntactic annotations. Those are available through the Stanford parser and propose various degrees of syntactic representation, ranging from a purely surface-oriented representation to a deep syntax framework where empty nodes and copied nodes (for gapping) are allowed. As in most other representations, subjects of infinitives are annotated in the "deepest" syntactic scheme. In the deeper syntactic layers, namely the *collapsed* instances, only semantically non-empty elements are represented and the representation itself can be a directed graph.

Regarding French, deep syntactic information such as diathesis information or argument elision was not natively annotated in the FTB. However, earlier attempts at deriving deeper representations were carried out by Schluter and Van Genabith (2009), who proposed a semi-automatic method to produce LFG f-structure representations on top of a heavily modified subset of the FTB, called the MFT, in order to facilitate the production of the f-structures. These comply to the principles of the LFG theory as expressed by the Pargram group and exposed in (Dalrymple, 2001). It covers e.g. subject of infinitives and argument sharing.

While it could have been possible to start our annotation process from the MFT (after conversion to dependencies), we chose not to because we wanted to remain more neutral with respect to the underlying linguistic theory, and we also wanted to annotate a freely available corpus.

Table 2 presents an overview of the main differences between the annotation schemes we mentionned in this paper.

	Ftb	SD	SD.D	PDT	MTT	MFT	Sequoia
graph	no	no	yes	yes	yes	yes	yes
subj. Inf.	no	no	yes	yes	yes	yes	yes
diathesis ch.	no	no	no	yes	yes	part.	yes
sbj. ellipsis	no	n/a	yes	yes	yes	yes	yes
RNR	no	no	yes	yes	yes	yes	yes
gapping	no	no	yes	yes	ukn	yes	yes

Table 2: Synthesis of some deep annotation schemes
SD: Stanford Basic Dependencies, SD.D: Stanford deep, MTT: Anchora-UFP Treebank. Subj. Inf: subject of infinitives, diathesis ch.: diathesis change, RNR: right-node raising

6. Conclusion

We have presented a deep syntactic representation scheme for French and its instantiation on the Sequoia corpus (Candito and Seddah, 2012b). We described its annotation process, which bears the originality of having being done by two distant teams so that annotation biases could be alleviated. We assessed the quality of the annotation by computing the labeled and unlabeled F-measure on the versions annotated by each team, before adjudication. The proposed representation is an intermediate step toward a full semantic analysis, and paves the way for building deep syntactic parsers. The resulting corpus contains 3k sentences, with updated surface dependency trees and deep syntactic

representations. It is freely available[13] under the LGPL-LR license[14].

Acknowledgements

This work was partially funded by the French *Investissements d'Avenir* - Labex EFL program (ANR-10-LABX-0083). We are grateful to Sylvain Kahane for useful discussions and to Alexandra Kinyon for proofreading this paper. All remaining errors are ours.

7. References

Abeillé, A. and Barrier, N. (2004). Enriching a French treebank. In *Proc. of LREC*, Lisbon, Portugal.

Abeillé, A., Toussenel, F., and Martine, C. (2004). Corpus le monde, annotations en constituants, guide pour les correcteurs. *LLF Annotation Guide*.

Abeillé, A. (2004). Guide des annotateurs, annotation fonctionnelle. *LLF Annotation Guide*.

Bohnet, B., Langjahr, A., and Wanner, L. (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.

Bonfante, G., Guillaume, B., Morey, M., and Perrier, G. (2011). Enrichissement de structures en dépendances par réécriture de graphes. In *Proc. of TALN*, Montpellier, France.

Cahill, A., Burke, M., O'Donovan, R., van Genabith, J., and Way, A. (2004). Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *Proc. of ACL*, pages 320–327, Barcelona, Spain.

Candito, M. and Seddah, D. (2012a). Effectively long-distance dependencies in French: annotation and parsing evaluation. In *Proc. of TLT 11*, Lisbon, Portugal.

Candito, M. and Seddah, D. (2012b). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN*, Grenoble, France.

Candito, M., Crabbé, B., and Denis, P. (2010). Statistical french dependency parsing: Treebank conversion and first results. In *Proc. of LREC*, Valletta, Malta.

Candito, M. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées application du français et a l'italien*. Ph.D. thesis, Université Paris Diderot.

Dalrymple, M. (2001). *Lexical-Functional Grammar*. Wiley Online Library.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford University.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proc. of LREC*, volume 6, pages 449–454.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Ph.D. thesis, Université Paris XIII, LIPN, INIST-CNRS.

Guillaume, B., Bonfante, G., Masson, P., Morey, M., and Perrier, G. (2012). Grew : un outil de réécriture de graphes pour le TAL. In *Proc. of TALN*, Grenoble, France.

Hajic, J., Panevová, J., Hajicová, E., Sgall, P., Pajas, P., Štepánek, J., Havelka, J., Mikulová, M., Zabokrtskỳ, Z., and Razımová, M. Š. (2006). Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

Hockenmaier, J. (2003). *Data and models for statistical parsing with Combinatory Categorial Grammar*. Ph.D. thesis.

Kahane, S. (2003). On the status of deep syntactic structure. In *Proc. of MTT 2003*, Paris, France.

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The penn treebank: annotating predicate argument structure. In *Proc. of the workshop on Human Language Technology*, pages 114–119, Stroudsburg, USA.

Melčuk, I. (1988). *Dependency syntax: theory and practice*. State University Press of New York.

Mille, S., Burga, A., and Wanner, L. (2013). AnCoraUPF: A Multi-Level Annotation of Spanish. In *Proc. of DepLing 2013*.

Miyao, Y. and Tsujii, J. (2005). Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proc. of ACL 2005*, pages 83–90.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Perlmutter, D. (1983). *Studies in Relational Grammar 1*. Studies in Relational Grammar. University of Chicago Press.

Ribeyre, C., Seddah, D., and Villemonte De La Clergerie, É. (2012). A Linguistically-motivated 2-stage Tree to Graph Transformation. In Han, C.-H. and Satta, G., editors, *Proc. of TAG+11*, Paris, France. INRIA.

Schluter, N. and Van Genabith, J. (2009). Dependency parsing resources for french: Converting acquired lexical functional grammar f-structure annotations and parsing f-structures directly. In *Proc. of NODALIDA 2009*, Odense, Denmark.

[13] At: `http://deep-sequoia.inria.fr`

[14] `http://deep-sequoia.inria.fr/license`