Conversion et améliorations de corpus du français annotés en Universal Dependencies

Bruno Guillaume, Marie-Catherine de Marneffe, Guy Perrier

Conversion et améliorations de corpus du français annotés en Universal Dependencies

Bruno Guillaume[*] — **Marie-Catherine de Marneffe**[**] — **Guy Perrier**[*]

[*] *Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*
[**] *The Ohio State University, Columbus, Ohio, USA*

bruno.guillaume@loria.fr, mcdm@ling.osu.edu, guy.perrier@loria.fr

RÉSUMÉ. *Cet article décrit l'effort d'amélioration de deux corpus du français annotés en dépendances syntaxiques, qui s'inscrit dans le cadre du projet Universal Dependencies (*UD*) qui vise à élaborer un schéma d'annotation syntaxique permettant d'analyser de façon similaire plusieurs langues différentes. Nous avons cherché à rendre plus conformes au schéma* UD *ces deux corpus du français, et nous avons évalué l'impact des modifications apportées aux corpus sur la conformité avec le schéma* UD *et la cohérence interne de leur annotation.*

ABSTRACT. *This paper describes an effort to improve the consistency of two French corpora annotated with the Universal Dependencies (*UD*) scheme. The Universal Dependencies project aims at building a syntactic dependency scheme which allows similar analyses for several different languages. We improved the annotations of the two French corpora to render them closer to the* UD *scheme, and evaluated the changes done to the corpora in terms of closeness to the* UD *scheme as well as of internal corpus consistency.*

MOTS-CLÉS : *corpus du français, syntaxe en dépendances, Universal Dependencies, correction de corpus.*

KEYWORDS: *French corpora, grammatical dependencies, Universal Dependencies, corpus correction.*

1. Introduction

Le projet Universal Dependencies[1] (UD) a pour but de créer un schéma d'annotation syntaxique qui puisse être utilisé pour un grand nombre de langues différentes (Nivre *et al.*, 2016). Pour que le projet aboutisse pleinement, il importe que les corpus annotés selon le schéma UD soient, d'une part, conformes à ce schéma, et, d'autre part, présentent une annotation cohérente au niveau de chaque corpus, mais également entre corpus, et principalement entre les corpus d'une même famille de langues (Zeman, 2015), et *a fortiori* entre les corpus d'une même langue. Dans cet article, nous décrivons l'effort pour harmoniser deux corpus français existants avec le schéma UD : UD_FRENCH-GSD et UD_FRENCH-SEQUOIA. Disposer de plusieurs corpus annotés de façon cohérente au sein d'une même langue a plusieurs avantages : offrir plus de données qui peuvent être utilisées de la même façon (les méthodes des réseaux neuronaux actuelles nécessitent une quantité de données importante) et faciliter la comparaison entre corpus ; une annotation cohérente entre langues permet également des applications translinguistiques. Nous avons donc choisi d'harmoniser les annotations de nos corpus du français avec le schéma UD, ce qui permet de les intégrer aux efforts d'apprentissage multilingue (Zeman *et al.* 2017, Zeman *et al.* 2018). Un schéma d'annotation qui se veut universel présente également certains inconvénients : les particularités de chaque langue ne peuvent être représentées. Transformer un corpus existant au schéma UD nécessite souvent de perdre de l'information qui existe dans le corpus original. Nous retraçons les décisions d'annotation et les corrections apportées aux corpus depuis leur intégration à UD. Pour minimiser la perte d'information, nous avons opté, pour certaines constructions, pour une analyse divergente du schéma UD, mais le schéma UD strict est chaque fois recouvrable de façon automatique. Nous montrons également, par deux méthodes d'évaluation, que les modifications apportées au corpus UD_FRENCH-GSD ont contribué, globalement, à une amélioration de la cohérence interne au corpus.

Un format d'annotation commun tel que UD permet, en théorie, d'offrir une analyse parallèle aux constructions grammaticales semblables dans différentes langues. L'annotation se fait à plusieurs niveaux : segmentation du texte en tokens, partie du discours et traits morphologiques pour chaque token, et relations syntaxiques entre les tokens. Pour la partie du discours, un jeu de dix-sept étiquettes est fixé, et chaque langue doit préciser comment elle utilise ce jeu, notamment en précisant quels mots sont considérés comme des particules et comment sont distinguées certaines étiquettes (les verbes des auxiliaires, les déterminants des pronoms, etc.). Pour la morphologie, un large éventail de traits est proposé et chaque langue est invitée à spécifier quels traits sont pertinents et comment ils doivent être annotés. Le schéma UD propose un système de trente-sept étiquettes principales pour les relations syntaxiques, communes à toutes les langues, tant celles morphologiquement riches que celles *pro drop*[2]. Le schéma propose également des étiquettes secondaires qui visent à prendre en compte

1. `http://universaldependencies.org`
2. Les langues pour lesquelles certains pronoms peuvent ne pas être réalisés syntaxiquement.

les particularités de certaines langues, et qui peuvent donc varier entre familles de langue, ou d'une langue à l'autre.

Dans UD, l'objectif d'une analyse parallèle a orienté certains choix d'annotation. Pour une même construction grammaticale, quand on passe d'une langue à une autre, les mots lexicaux (noms, verbes, adjectifs et certains adverbes) sont plus stables que les mots grammaticaux (les autres catégories). Pour offrir une analyse parallèle de constructions semblables, les relations syntaxiques portent donc directement sur les mots lexicaux, les mots grammaticaux étant représentés comme marqueurs de ces mots lexicaux. La figure 1 illustre la différence entre les réalisations syntaxiques de prédication non verbale en français et en russe[3]. Les relations en trait continu sont celles entre mots lexicaux ; les relations en pointillé portent sur les mots grammaticaux. Le français utilise une copule, absente en russe. Le français utilise également des déterminants pour exprimer que le nom est défini, alors que le russe utilise une marque morphologique. Le russe utilise également un système de cas, là où le français utilise une préposition. En donnant priorité aux relations entre mots lexicaux, UD offre une analyse parallèle dans les deux langues.

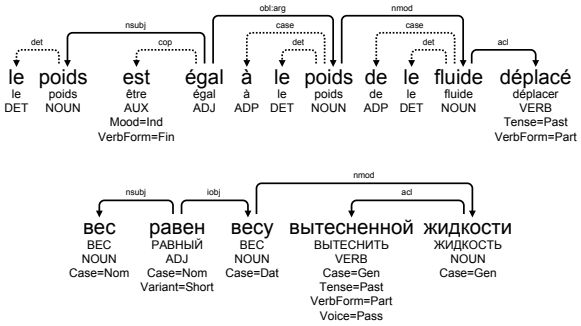

Figure 1. *Annotation en français et russe de la phrase :* le poids est égal au poids du fluide déplacé

Si donner priorité aux relations entre mots lexicaux permet une analyse parallèle entre langues ou différents usages d'une même langue (e.g., langage chez l'enfant qui omet souvent aux premiers stades les mots grammaticaux), certaines particularités linguistiques sont plus difficiles à représenter. On perd en particulier un lien direct entre les verbes et les prépositions qui en dépendent (Osborne et Gerdes, 2019) : par exemple, les prépositions *sur* et *de* dans les phrases verbales *compter sur quelqu'un* ou *dépendre de quelqu'un* seront analysées comme dépendantes de *quelqu'un* alors qu'elles sont sous-catégorisées en fonction du verbe. Un schéma qui permet une analyse parallèle de constructions semblables amène donc à devoir accepter certains compromis d'analyses linguistiques et à des pertes d'information.

3. La phrase russe choisie est dérivée de la phrase `train-s505` du corpus UD_RUSSIAN-GSD.

La version 2.4 de UD contient 146 corpus représentant 83 langues. La taille des corpus diffère toutefois largement, de 292 mots et 55 phrases (UD_TAGALOG-TRG) à environ 1,5 million de mots et 88 000 phrases (UD_CZECH-PDT).

Dans la suite, nous introduisons d'abord les particularités du français dans le projet UD, au niveau de l'annotation ainsi que les corpus disponibles (section 2). Nous décrivons brièvement les méthodes de correction de corpus existantes (section 3). Nous détaillons ensuite les deux corpus qui sont l'objet de cet article, le UD_FRENCH-GSD (section 4) et le UD_FRENCH-SEQUOIA (section 5), et les corrections et transformations que nous y avons apportées pour se conformer au schéma UD. Nous terminons par évaluer l'évolution du corpus UD_FRENCH-GSD en comparant les différentes versions de celui-ci à une annotation de référence ainsi que les performances d'un analyseur syntaxique entraîné et évalué sur les différentes versions.

2. UD pour le français

2.1. *Application du guide d'annotation de UD au français*

Le guide d'annotation général de UD, d'une part, est muet sur certains phénomènes (les clivées par exemple) et, d'autre part, ne traite pas des spécificités de chaque langue. Pour le français, nous avons travaillé avec Marie Candito, Kim Gerdès, Sylvain Kahane et Djamé Seddah à l'établissement d'un guide d'annotation en UD[4] pour les phénomènes ou pour les spécificités du français que ne couvre pas le guide général. Nous présentons ci-après quelques-uns de ces phénomènes et de ces spécificités sur lesquels nous nous sommes penchés plus particulièrement.

2.1.1. *La copule*

Le verbe *être*, lorsqu'il n'est pas auxiliaire de temps ou du passif, est seulement considéré comme copule lorsqu'il a un attribut du sujet qui n'est pas verbal, sinon il est considéré comme un verbe ordinaire. La frontière entre les deux n'est pas toujours évidente. On peut imaginer des critères pour aider à faire la distinction s'inspirant de l'approche pronominale introduite par Claire Blanche-Benveniste (Blanche-Benveniste *et al.*, 1987) et mise en œuvre dans le lexique Dicovalence[5]. Par exemple, si on peut poser la question *est comment ? est quoi ?*, on considérera le verbe *être* comme copule. La première question renvoie à un attribut du sujet qui est une propriété tandis que la seconde renvoie à un attribut du sujet qui est une entité. Si on peut poser la question *est où ?*, le verbe *être* sera analysé comme un verbe ordinaire avec le sens *être situé* et requérant un argument locatif. Ces critères ne permettent pas de trancher tout le temps comme le montrent les exemples (1) et (2). Dans ceux-ci, nous avons choisi de traiter le verbe *être* comme copule mais sans utilisation d'un critère décisif. Les expressions *en poste* et *au pouvoir* sont alors vues comme des propriétés,

4. http://universaldependencies.org/fr
5. https://www.ortolang.fr/market/lexicons/dicovalence

mais on pourrait contester ce choix, en disant que le verbe *être* signifie *être situé* mais dans un sens figuré.

(1) *Il a été en poste de 1934 à 1941.*

(2) *[. . .] cette ville était encore au pouvoir des Ligueurs.*

Lorsque la copule a comme attribut du sujet une proposition, nous avons choisi de ne pas la considérer comme dépendant de la tête de cette proposition dans une relation `cop`. Si on le faisait, cette tête qui est un verbe aurait deux sujets, un sujet propre et un sujet lié à la copule. Prenons un exemple.

(3) *Le seul problème est qu'il n'a pas de super-pouvoirs [. . .]*

Dans la phrase (3), l'attribut du sujet est toute la proposition *qu'il n'a pas de super-pouvoirs* avec comme tête le verbe *a*. Si, comme on fait généralement dans UD, on considérait qu'il y a une dépendance `cop` de *a* vers *est*, le verbe *a* serait le gouverneur de deux dépendances `nsubj`, l'une pour le sujet propre *il* dans la subordonnée et l'autre pour le sujet de la principale *problème*. Pour éviter cette difficulté, on traite dans ce cas le verbe *être* comme un verbe ordinaire (figure 2).

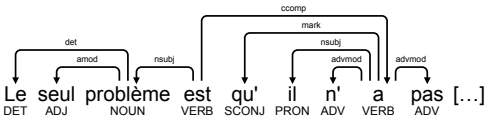

Figure 2. *Annotation* UD *de la phrase (3)*

L'inconvénient est d'avoir deux modélisations différentes de la copule selon la forme de l'attribut du sujet. En anglais et en allemand, par exemple, certains ont choisi de considérer que la copule est dépendante de l'attribut du sujet dans tous les cas mais avec le problème d'avoir deux sujets lorsque l'attribut est une proposition.

2.1.2. *Les dates*

Pour les dates, la hiérarchie suivante dans l'établissement des dépendances a été établie : jour de la semaine → date du jour → mois → année. Dans l'expression *le 25 janvier 2010*, le fait que l'on puisse l'élider en *le 25* est un argument décisif pour que soit choisie comme tête de l'expression *25* et pas *janvier*.

2.1.3. *Les clivées*

Dans une clivée comme (4), la subordonnée n'est pas un argument du foyer de cette clivée, mais pour la distinguer d'une subordonnée ordinaire qui modifie une proposition principale, nous utilisons la relation `advcl:cleft` de la tête du foyer vers la tête de la subordonnée. Une clivée peut avoir une forme interrogative. Dans l'exemple (5),

le foyer de la clivée est le pronom interrogatif *Qu'*. Cette interrogative est une clivée car on peut la paraphraser : *C'est quoi qui va augmenter ?*

(4) *C'est la troisième fois que nous venons, […]*

(5) *Qu'est-ce qui va augmenter ?*

2.1.4. *Les sujets explétifs*

Le guide d'annotation de UD préconise d'utiliser la relation `expl` pour les arguments syntaxiques qui n'ont pas de rôle sémantique. Récemment, Bouma *et al.* (2018) ont proposé des critères plus précis pour annoter les différentes sous-catégorisations des relations `expl`. Les versions actuelles des corpus présentés sont conformes à ces préconisations pour le pronom réfléchi *se* servant à marquer les verbes essentiellement pronominaux (*s'enfuir*) ou le passif pronominal (*le bruit s'entend de loin*) qui sont bien annotés `expl`.

En revanche, pour d'autres phénomènes, les corpus ne sont pas encore complètement conformes à ces préconisations. Le token *t* dans *va-t-il* est annoté `expl` alors qu'il n'est pas argument. Les annotations actuelles s'écartent également du guide pour les constructions impersonnelles. Dans l'exemple *il ne manque plus que le soleil*, la dépendance de *manque* vers *il* est étiquetée `nsubj`, alors que la dépendance de *manque* vers *soleil* est étiquetée `obj`. Selon le guide d'annotation de UD, la dépendance de *manque* vers *il* devrait être étiquetée `expl` et celle de *manque* vers *soleil* devrait être étiquetée `nsubj`[6]. Ces questions seront gérées dans les prochaines versions du corpus.

2.2. *Les expressions polylexicales*

Pour l'annotation des expressions polylexicales, nous proposons une annotation plus riche[7] que celle prévue dans le guide UD. Cependant, pour fournir des données le plus proche possible de ce qui est préconisé par le guide, nous avons mis en place une conversion automatique. Les données de la distribution officielle sont celles obtenues après conversion (et donc conformes au guide), les données enrichies sont disponibles directement sur le Github du projet (pour le corpus UD_FRENCH-GSD uniquement).

Une définition très générale de ce qu'est une expression polylexicale (EP) consiste à dire que c'est une expression qui ne respecte pas le principe de compositionnalité du sens. Le projet PARSEME-FR[8] propose pour le français des critères permettant d'identifier les EP[9]. Par exemple, le critère [MORPHO] indique que si changer les

6. L'inconvénient d'une telle annotation est qu'elle ne permet pas de distinguer les explétifs sujets des autres explétifs. Par ailleurs, pourquoi annoter une alternance syntaxique pour les constructions impersonnelles et pas pour les constructions passives ?

7. Cette proposition est issue des discussions avec les collègues cités *supra* ainsi qu'une collaboration dans le cadre du projet PARSEME-FR porté par Mathieu Constant.

8. `http://parsemefr.lif.univ-mrs.fr`

9. `https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Criteres`

traits morphosyntaxiques d'une expression est agrammatical ou provoque un changement de sens inattendu, elle doit être considérée comme une EP. C'est le cas pour *perdre les pédales* car *perdre la pédale* a bien un sens, mais la différence de sens n'est pas celle attendue pour le changement de nombre d'un objet direct. Comme le reconnaissent les auteurs de ces critères, assigner un statut binaire (EP *versus* expression compositionnelle) est parfois difficile. Leur choix a été de prendre chaque critère comme une condition suffisante d'EP.

Le guide UD impose la relation `fixed` pour représenter les EP : le mot le plus à gauche d'une EP est relié à tous les autres par des dépendances `fixed`. Cette représentation ignore donc la structure syntaxique interne éventuelle d'une EP. Or, quand cette structure est manifeste, il peut être utile de la faire apparaître. Dans notre version enrichie, nous avons choisi de la représenter comme une structure syntaxique ordinaire et nous utilisons des traits pour marquer l'EP. D'une part, la tête de l'EP porte un trait *MWEPOS* qui donne la catégorie grammaticale de l'EP considérée comme un tout par rapport au reste de la phrase. D'autre part, tous les autres composants sont marqués avec le trait *INMWE=Yes*. Nous reprenons ici une proposition de représentation faite par Candito et Constant (2014).

L'exemple (6) contient l'EP *en même temps* et la figure 3 illustre les deux annotations.

(6) *Ils furent créés en même temps que les tribuns de la plèbe.*

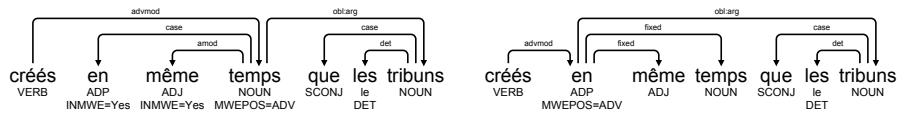

Figure 3. *Annotation enrichie et annotation* UD *pour la phrase (6)*

Dans l'annotation enrichie, comme l'EP joue le rôle d'un adverbe, sa tête *temps* porte un trait *MWEPOS=ADV*. Les composantes de l'EP, *en* et *même*, portent, elles, le trait *INMWE=Yes*. Le caractère facultatif de la conjonction *que* et les coordinations de la forme *en même temps que A et que B* plaident pour dissocier la conjonction *que* de l'EP. Dans l'exemple ci-dessus, la conjonction *que* introduit le complément *que les tribuns de la plèbe* qui est un argument de l'EP, exprimé par une dépendance `obl:arg` de la tête de l'EP *temps* vers *tribuns*.

La seconde annotation de la figure 3, conforme au guide UD est produite automatiquement à l'aide d'un système de réécriture de graphes (sept règles). Il suffit de déplacer la tête de l'EP sur le mot le plus à gauche et de remplacer toutes les dépendances internes à l'EP par des dépendances `fixed` issues du mot le plus à gauche.

Pour les EP qui n'ont pas de structure syntaxique interne évidente (par exemple *ou bien*, *parce que*, *tandis que*), nous utilisons la relation `fixed` prévue par le guide UD. Dans la version 2.4 de UD_FRENCH-GSD, il y a 2 841 EP annotées : 1 274 (45 %) sont annotées en `fixed` et 1 567 (55 %) le sont avec une annotation enrichie.

2.3. *Les corpus du français disponibles au format* UD.

Dans le projet UD (version 2.4 de mai 2019), il y a sept corpus pour le français. Deux de ces corpus (UD_FRENCH-GSD et UD) sont détaillés ailleurs dans l'article. Nous listons les cinq autres ci-dessous. Le tableau 1 reprend la taille des corpus, le nombre de relations grammaticales utilisées, et précise si les corpus disposent de lemmes et de traits morphologiques.

– UD_FRENCH-PARTUT est disponible depuis la version 2.0. Ce corpus est la conversion d'un corpus existant PARTUT (Sanguinetti et Bosco, 2015) et il contient des genres variés (Wikipédia, textes légaux, parole…). Toutes les phrases du corpus sont alignées avec leurs équivalents en anglais (UD_ENGLISH-PARTUT) et en italien (UD_ITALIAN-PARTUT)[10].

– UD_FRENCH-FTB est disponible depuis la version 2.0[11]. La conversion vers le schéma UD a été faite avec les outils de conversion présentés en 5.2 complétés par une étape d'analyse syntaxique automatique pour récupérer les erreurs de conversion par règles (Seddah *et al.*, 2018).

– UD_FRENCH-PUD (disponible depuis la version 2.1) fait partie du projet Parallel Universal Dependencies (PUD) qui propose un ensemble de 1 000 phrases alignées dans différentes langues (Zeman *et al.*, 2017). Les phrases proviennent de Wikipédia ou de textes journalistiques dans cinq langues sources et ont été traduites dans quinze langues au total. Les phrases ont été annotées en suivant le schéma de McDonald *et al.* (2013), et transformées ensuite selon le schéma UD par des membres de la communauté UD. Pour le français, l'annotation de ce corpus diffère quelque peu des autres en restant souvent plus proche de l'anglais : par exemple, les possessifs sont annotés comme des pronoms avec la relation `nmod:poss` (les autres corpus français les annotent comme des déterminants avec la relation `det`).

– UD_FRENCH-SPOKEN (Gerdes et Kahane, 2017), disponible depuis la version 2.2, est une conversion du *treebank* Rhapsodie[12] (MoDyCo *et al.*, 2018) qui contient des annotations en prosodie et en syntaxe de transcriptions de l'oral.

– UD_FRENCH-FQB est disponible depuis la version 2.4. Ce corpus provient d'une conversion automatique du French QuestionBank v1 (Seddah et Candito, 2016) au schéma UD. La conversion a été faite avec la procédure décrite section 5.2.

10. Les corpus italiens et anglais contiennent d'autres phrases absentes de la partie française.
11. Dans les données du projet UD, le texte des phrases et les lemmes ne sont pas fournis car le corpus FTB original est soumis à une licence spécifique (gratuite pour la recherche).
12. `http://www.projet-rhapsodie.fr`

	# phrases	# tokens	lemmes	morpho	# rel
UD_FRENCH-GSD	16 342	400 387	Oui	Oui	50
UD_FRENCH-SEQUOIA	3 099	70 567	Oui	Oui	45
UD_FRENCH-FTB	18 535	573 370	Oui	Oui	36
UD_FRENCH-PARTUT	1 020	28 594	Oui	Oui	45
UD_FRENCH-PUD	1 000	24 734	Non	Oui	42
UD_FRENCH-SPOKEN	2 786	34 972	Oui	Non	52
UD_FRENCH-FQB	2 289	24 135	Oui	Oui	39

Tableau 1. *Information sur les sept corpus du français disponibles en* UD *(version 2.4)*

3. Les méthodes de correction de corpus

3.1. *État de l'art*

Plusieurs méthodes ont été développées pour identifier des erreurs systématiques d'annotation dans les corpus. Il y a deux types de méthodes : celles qui se fondent sur la recherche de motifs définis *a priori* (par exemple De Smedt *et al.* (2016) qui se concentrent sur les erreurs d'annotation dans les expressions figées), et celles qui se fondent sur la localisation de contextes susceptibles de contenir des erreurs (Boyd *et al.*, 2008 ; Alzetta *et al.*, 2018a). Ces deux types de méthodes nécessitent néanmoins une inspection manuelle, pour distinguer les annotations qui résultent d'une réelle différence de celles qui proviennent effectivement d'erreurs.

Boyd *et al.* (2008) se fondent sur le concept de *variation nuclei* développé par Dickinson et Meurers (2003, 2005). Un *variation nucleus* est un élément qui apparaît plusieurs fois dans un corpus avec une annotation différente. Par exemple, dans la figure 4, la construction *ce qui* reçoit deux analyses différentes. Pour Boyd *et al.* (2008), un élément de variation est une paire de mots (*ce* et *élevé* dans l'exemple) apparaissant dans un même contexte (même mot à gauche et à droite du *nucleus*) mais liés par une relation différente.

Figure 4. *Exemple de paire de mots repérée avec la méthode de Boyd* et al. *(2008)*

La méthode de Boyd a été utilisée sur les corpus UD par de Marneffe *et al.* (2017) et Wisniewski (2018) ; de Marneffe *et al.* (2017) l'étendent aux lemmes (et non aux formes de surface), ce qui permet d'extraire plus d'erreurs potentielles (figure 5).

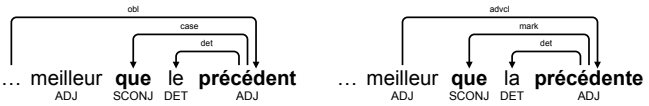

Figure 5. *Exemple d'annotation repérée par la méthode de Marneffe* et al. *(2017)*

Pour UD_FRENCH-GSD 2.0, cette méthode extrait 474 paires potentiellement erronées. Sur 100 paires analysées, 65 % sont erronées.

Alzetta *et al.* (2018b) et Alzetta *et al.* (2018a) utilisent la fiabilité de relations grammaticales produites automatiquement pour identifier des erreurs d'annotation dans les corpus. Leur méthode est fondée sur l'algorithme de Dell'Orletta *et al.* (2013) qui assigne un score de qualité à chaque arc produit par un analyseur, et classe donc les arcs, des plus corrects à ceux potentiellement incorrects. L'algorithme se fonde sur l'hypothèse suivante : plus une structure est fréquemment générée par un analyseur, plus elle est considérée comme correcte. Les relations qui obtiennent un score bas sont en effet souvent susceptibles d'être incorrectes. Une inspection manuelle de ces arcs de pauvre qualité a permis d'identifier des motifs d'erreurs (par exemple, des cas où l'auxiliaire est la tête). Ces motifs d'erreurs, projetés sur tout le corpus italien de UD, reflétaient une réelle erreur dans 86 % des cas.

3.2. *La réécriture de graphes*

La réécriture de graphes est un modèle de calcul puissant qui est utilisé dans de nombreux domaines de l'informatique. Même si les données manipulées dans les modélisations linguistiques sont souvent des arbres, le fait de les voir comme des graphes a de nombreux avantages. D'une part, cela permet de considérer dans un cadre unifié toutes les représentations et notamment celles qui ne sont pas des arbres comme les *enhanced dependencies*[13]. D'autre part, cela permet de gérer de façon homogène des informations périphériques comme l'ordre des mots, par exemple.

Si l'on considère toutes les structures manipulées comme des graphes, la réécriture de graphes est un cadre naturel pour décrire les transformations que l'on souhaite effectuer sur ces structures. Le principe est de décomposer une transformation globale en une succession de transformations élémentaires qui ne font chacune intervenir

13. Les *enhanced dependencies* (`http://universaldependencies.org/workgroups/enhanced.html`) comprennent notamment toutes les dépendances argumentales qui ne sont pas représentées dans UD, telles que les sujets d'infinitifs. On les appelle des dépendances profondes par opposition à celles de UD qui sont considérées comme des dépendances de surface.

qu'un nombre fixé de nœuds du graphe. Chaque transformation élémentaire est décrite par une règle qui est composée de deux parties : un motif (qui décrit le contexte dans lequel une transformation peut s'appliquer) et des commandes (qui décrivent les modifications locales qu'il faut apporter au graphe quand le motif est repéré).

La réécriture de graphes a été utilisée ponctuellement en traitement des langues (Hyvönen, 1984 ; Bohnet et Wanner, 2001 ; Jijkoun et de Rijke, 2007 ; Bedaride et Gardent, 2009 ; Chaumartin et Kahane, 2010 ; Ribeyre, 2016). Nous verrons plus loin des exemples d'utilisation de la réécriture de graphes pour des transformations de corpus, mais on utilise également de façon indépendante le repérage de motifs pour la recherche d'erreurs ou d'incohérences dans les annotations.

Nous utilisons le logiciel GREW[14] pour décrire et appliquer les règles de nos transformations et l'outil dérivé GREW-MATCH[15] pour le repérage de motifs. Nous avons choisi d'utiliser le système GREW, que nous maîtrisons et pouvons faire évoluer selon les besoins, deux des auteurs du présent article étant impliqués dans la création et la maintenance de cet outil (Bonfante *et al.*, 2018).

3.3. *Corrections utilisant une double conversion*

Convertir une annotation d'un format A dans un format B puis convertir l'annotation obtenue en sens inverse de B vers A permet de détecter des incohérences dans l'annotation initiale. Si l'on ne retrouve pas l'annotation de départ après la double conversion, il y a nécessairement une erreur dans l'annotation de départ ou dans le processus de conversion.

Dans les expériences que nous avons menées, nous avons utilisé comme format B le format SUD (*Surface-Syntactic Universal Dependencies*) (Gerdes *et al.*, 2018). Le projet UD se propose de fournir un cadre commun d'annotation syntaxique à des corpus, quelle que soit la langue utilisée par ces corpus. C'est pourquoi le format tient compte du niveau sémantique, qui présente une moindre différence entre les langues. Les têtes des dépendances sont les mots lexicaux, et les mots grammaticaux sont dépendants des mots lexicaux. Les structures syntaxiques de UD sont donc plutôt plates[16]. Par ailleurs, les noms des relations dépendent des fonctions syntaxiques qu'elles représentent, mais aussi des catégories des mots qu'elles font intervenir. Ainsi, une relation *modificateur*, quand le gouverneur est un verbe, est notée `obl`, `advcl` ou `advmod`, selon que le dépendant est un nom, un verbe ou un adverbe. Quand le gouverneur est un nom, la relation est notée `nmod`, `amod`, `acl` ou `advmod` selon que le dépendant est un nom, un adjectif, un verbe ou un adverbe.

14. `http://grew.fr`

15. `http://match.grew.fr`

16. Dans UD_FRENCH-GSD, 4,9 % des phrases ont une annotation non projective contre 14,3 % pour la version SUD.

Le format SUD se veut plus fidèle à la tradition de la syntaxe en dépendances (Mel'čuk, 1988 ; Hajič *et al.*, 2017) qui met au centre les fonctions syntaxiques. Les noms des relations représentent strictement des fonctions syntaxiques indépendamment de la catégorie des mots qu'elles relient (comme également pour le format FTB décrit plus bas). Ainsi, il y a une seule relation `mod` pour les modificateurs qui vaut, quelle que soit la catégorie du mot qui modifie et du mot qui est modifié. Les mots fonctionnels sont la tête des expressions qu'ils introduisent. Cela concerne les prépositions, les conjonctions de subordination, les auxiliaires et les copules. Les seules exceptions concernent les déterminants et les conjonctions de coordination qui sont dépendants de mots auxquels ils s'appliquent.

Comme dans UD, les étiquettes peuvent se présenter en deux parties *(étiquette principale : étiquette secondaire)*, ce qui permet de varier la granularité de l'étiquetage. Par exemple, la relation `comp` s'applique à tous les arguments verbaux mais on peut préciser `comp:obj` pour les objets directs. Le lecteur trouvera plus bas un exemple de phrase annotée en UD et en SUD dans la figure 8.

4. Le corpus UD_FRENCH-GSD

La dernière version en date du corpus UD_FRENCH-GSD est la version 2.4 (novembre 2018). Suivant les conventions du projet UD, le corpus est divisé en trois parties dont les tailles sont reprises dans le tableau 2.

	`train`	`dev`	`test`	Total
Nombre de phrases	14 450	1 476	416	16 342
Nombre de tokens	354 655	35 714	10 018	400 387

Tableau 2. *Taille des parties du corpus* UD_FRENCH-GSD

La maintenance et la correction de corpus sont une tâche de longue haleine que nous menons en continu depuis 2015 sur le corpus UD_FRENCH-GSD dans le contexte du projet UD qui propose deux nouvelles distributions des corpus chaque année.

Les modifications apportées sont de deux types : l'application au corpus de nouvelles décisions d'annotation (qui peuvent être générales au projet UD, spécifiques au français ou internes au corpus lui-même) et la correction d'erreurs ou d'incohérences.

4.1. *Historique du corpus*

Les données actuelles du corpus UD_FRENCH-GSD proviennent du *universal dependency treebank v2.0* de Google (McDonald *et al.*, 2013). Le corpus UD_FRENCH-GSD a été initialement annoté manuellement par deux équipes différentes dans le

cadre du projet de Google d'harmonisation de dépendances grammaticales pour six langues (l'anglais, l'allemand, le français, l'espagnol, le suédois et le coréen). La stratégie adoptée par Google pour harmoniser le schéma d'annotation était de garder l'ensemble des relations grammaticales et les analyses le plus proche possible de celles du corpus anglais SD (de Marneffe *et al.*, 2006). Les annotateurs ne pouvaient ajouter de relations grammaticales que pour des phénomènes non présents en anglais. Cette stratégie rigide d'annotation a, d'une part, poussé à analyser certaines constructions grammaticales parallèlement à l'analyse proposée pour l'anglais au lieu de tenir compte de différences systématiques entre langues et, d'autre part, a gardé des étiquettes différentes pour exprimer une distinction qui est réalisée au niveau morphologique (comme marquer différemment un modificateur infinitif (*un spectacle à mourir d'ennui*) ou participial (*un spectacle barbant au possible*)) et qui ne se manifeste pas dans toutes les langues.

Début 2015, le corpus a été converti dans le format UD sans que cette transformation ne soit documentée. Depuis lors, nous avons travaillé régulièrement pour faire évoluer ces données. Pour la production de la version 2.0, nous avons dû convertir les annotations pour tenir compte des nouvelles conventions d'annotation. Les autres évolutions ont été internes au corpus ou en collaboration avec d'autres corpus du français. Le tableau 3 décrit les principales évolutions du corpus. Le tableau ne fait pas figurer les corrections diverses des annotations et les homogénéisations qui ont été continues pendant les trois ans de travail sur le corpus. Nous revenons dans la section suivante sur ces corrections plus ponctuelles.

Il est important de noter que les données initiales disponibles ne comportaient aucune métadonnée. Pour les phrases du corpus UD_FRENCH-GSD, nous ne disposons ni du texte original ni d'information sur la source dont est issue la phrase. Ainsi, lors de la production de la version 1.2, le champ `sent_id` a été introduit avec une numérotation interne au corpus et le champ `text` a été reconstruit à partir des annotations.

Par ailleurs, dans les données initiales, nous avons trouvé de nombreuses phrases tronquées dans lesquelles une partie du texte original était manquante. La quasi-totalité des phrases tronquées provient de Wikipédia et les parties manquantes sont souvent des unités ou des dates. Nous avons donc décidé de compléter les phrases du corpus en reprenant le texte original de Wikipédia. En effet, ces erreurs sont dues à des traitements précédents d'extraction des données et ne reflètent pas des erreurs d'usage de la langue. En revanche, nous n'avons évidemment pas corrigé d'autres types d'erreurs de grammaire ou de syntaxe qui sont présentes dans la version Wikipédia utilisée pour construire le corpus (même si souvent ces erreurs sont corrigées dans la version actuelle de Wikipédia) car ces erreurs font partie de l'usage de la langue qu'un corpus souhaite refléter. Il arrive encore que l'on trouve de nouvelles phrases tronquées qui sont alors corrigées au fil de l'eau.

Dans l'annotation UD, il n'est pas requis de faire de distinction entre les deux types de compléments obliques du verbe : les arguments et les modificateurs. Cependant, pour les applications et notamment une analyse sémantique, la distinction entre argument et modificateur est essentielle. Nous avons donc décidé de l'ajouter à la res-

source UD_FRENCH-GSD en sous-typant `obl` en `obl:mod`, `obl:arg` ou `obl:agent` (pour les compléments d'agent).

Une petite partie de ces distinctions a pu être prédite automatiquement (par exemple, les compléments antéposés, séparés par une virgule sont systématiquement `obl:mod`). Pour le reste, l'annotation a été faite manuellement par un annotateur à l'aide d'un outil dédié qui présente à l'annotateur les cas à décider, classés par préposition. Ainsi, 39,87 % des `obl` ont pu être sous-typés dans la version 2.2 et 67,40 % dans la version 2.3.

Version	Date	Description
1.0	janvier 2015	Version initiale construite depuis le Google Dataset
1.1	mai 2015	Corrections de quelques segmentations en phrases
1.2	novembre 2015	Ajout de métadonnées, corrections de phrases tronquées
1.3	mai 2016	Ajout des lemmes et de la morphologie
1.4	novembre 2016	Pas de changement majeur
2.0	de mars à mai 2017	Adaptation des données aux nouvelles directives
2.1	novembre 2017	Revue systématique des auxiliaires
2.2	juillet 2018	Sous-typage partiel de la relation `obl`
2.3	novembre 2018	Applications des nouvelles décisions : date, EP (partiel)
2.4	mai 2019	Corrections imposées par le nouveau validateur UD

Tableau 3. *Historique des versions du corpus* UD_FRENCH-GSD

4.2. *Méthodes utilisées pour corriger le corpus*

Une partie des modifications induites par de nouvelles décisions peuvent être faites automatiquement. En revanche, d'autres nécessitent un travail manuel d'annotation. Pour la correction d'erreurs ponctuelles ou d'incohérences, le recours à l'annotation manuelle est systématique.

4.2.1. *Corrections automatiques*

Nous donnons à titre d'exemple une liste (loin d'être exhaustive) des modifications automatiques apportées au corpus UD_FRENCH-GSD.

– Ajout des lemmes et de la morphologie. Dans la version 1.1, pour chaque mot, seule la catégorie était renseignée. Nous avons appliqué systématiquement une prédiction du lemme et de la morphologie à partir de la forme fléchie et de la catégorie en utilisant le lexique *Lefff* (Sagot, 2010). Les annotations ont été faites manuellement pour les mots absents de *Lefff* et pour les lemmes ambigus (par exemple, la forme verbale *suis* peut correspondre aux lemmes *être* ou *suivre*). Pour les ambiguïtés morphologiques, des règles ont été utilisées et les ambiguïtés morphologiques qui n'ont pas pu être levées automatiquement l'ont été manuellement.

– Changement de l'annotation des coordinations pour la version 2 de UD. Dans la version 1, les conjonctions de coordination sont rattachées à la tête du premier conjoint, alors que dans la version 2, elles sont rattachées à la tête du second conjoint.

– Réduction des verbes auxiliaires aux lemmes *être*, *avoir* et *faire* et des copules au lemme *être*. Dans les premières versions, les auxiliaires comprenaient aussi les auxiliaires modaux et d'aspect et dans certaines d'entre elles, les copules étaient étendues à quelques verbes d'état.

– Changement de l'annotation des dates avec le chiffre du jour comme tête.

Dans la figure 6, nous donnons un exemple de règle qui modifie localement l'annotation des dates. La partie *pattern* définit la partie du graphe à modifier. Elle comprend deux nœuds `DAY` et `MONTH` qui représentent le jour et le mois. Le jour est dépendant du mois par une relation `nummod` ou `amod`. La partie *commands* donne la suite des opérations qui va permettre de transformer le motif détecté. La dépendance entre le mois et la date est remplacée par une dépendance `nmod` de `DAY` vers `MONTH`. Les commandes `shift_in` et `shift_out` permettent de modifier la façon dont les nœuds du motif sont reconnectés au contexte. Ici, comme on change la tête de `MONTH` à `DAY`, la commande `shift_in` déplace la dépendance pointant sur l'ancienne tête `MONTH` vers la nouvelle `DAY`; la commande `shift_out` déplace les dépendances issues de l'ancienne tête (sauf celles de type `nmod` et `nummod`).

```
rule day_month {
  pattern {
    DAY [upos=NUM|ADJ]; MONTH [lemma="janvier"|...|"décembre"];
    e: MONTH -[nummod|amod]-> DAY; DAY << MONTH;
  }
  commands {
    del_edge e; add_edge DAY -[nmod]-> MONTH;
    shift_in MONTH ==> DAY; shift_out MONTH =[^nmod|nummod]=> DAY;
  }
}
```

Figure 6. *Une des règles utilisées pour la conversion des dates*

4.2.2. *Corrections manuelles*

Quand les modifications à apporter ne sont pas complètement prédictibles à partir de la syntaxe existante, il faut procéder à une annotation manuelle. C'est aussi le cas pour les corrections ponctuelles : on corrige au cas par cas mais en essayant à chaque fois d'appliquer ces modifications de façon homogène à tout le corpus en recherchant des contextes similaires à celui qu'on est en train de corriger.

Dans de nombreux cas, l'outil GREW-MATCH a été utilisé car il permet justement d'afficher un ensemble de contextes similaires décrits par un motif. Les utilisations de cette méthode ont été très nombreuses (de l'ordre de la centaine) et nous en donnons quelques-unes à titre d'exemple :

– détection des incohérences entre une relation et la nature du dépendant (relation `amod` qui ne pointe pas sur un adjectif par exemple) ;

– recherche de deux relations `sujet` portant sur le même verbe ;

– recherche d'une relation `sujet` dont le gouverneur est un nom sans copule (figure 7) ;

– recherche des défauts d'accord (en genre ou en nombre) entre un verbe et son sujet, un adjectif et le nom sur lequel il porte.

```
pattern { N [upos=NOUN]; V -[nsubj]-> N }
without { V -[cop]-> * }
```

Figure 7. *Motif pour retrouver les noms communs sujets en l'absence de copule*

4.2.3. *Application de la méthode de double conversion*

La double conversion du corpus UD_FRENCH-GSD, telle qu'elle a été présentée à la sous-section 3.3 a été appliquée. Le corpus obtenu (appelé UD_FRENCH-GSD^{DC}) a été comparé avec le corpus initial UD_FRENCH-GSD.

Lors de la première application de la méthode, 3 955 des 400 440 relations (soit 1 %) étaient différentes (gouverneur différent ou étiquette différente). Une différence peut provenir d'une erreur dans l'écriture des règles de conversion qu'il faut alors corriger. Sinon, elle vient d'une erreur d'annotation de UD_FRENCH-GSD. L'erreur est alors corrigée, mais comme il est expliqué plus haut (4.2.2), on vérifie systématiquement les contextes similaires pour corriger de façon globale d'autres erreurs similaires. En itérant la mise à jour des données et de la conversion, le nombre de différences entre UD_FRENCH-GSD et UD_FRENCH-GSD^{DC} a été réduit de 3 955 à 351. Le format UD est plus précis que SUD pour représenter les coordinations imbriquées, ce qui explique 337 différences. Les 14 différences restantes sont liées à des phénomènes très spécifiques non pris en compte par les règles de conversion.

5. Le corpus UD_FRENCH-SEQUOIA

Le corpus UD_FRENCH-SEQUOIA est apparu plus récemment dans le projet UD. Il est disponible depuis la version 2.0 (mars 2017) et nous l'avons obtenu par conversion du corpus SEQUOIA[17] existant. Nous présentons ici le corpus tel qu'il existe au format avant la conversion ainsi que la conversion elle-même.

17. `http://deep-sequoia.inria.fr`

5.1. *Le format d'annotation* FTB

Le corpus SEQUOIA a été annoté en constituants en suivant le schéma d'annotation du *French Treebank* (FTB) (Abeillé et Barrier, 2004) et converti automatiquement en dépendances (Candito et Seddah, 2012b) (nous noterons FTB ce format en dépendances dans la suite). Les dépendances à distance ont été corrigées manuellement (Candito et Seddah, 2012a). Depuis, le corpus a notamment évolué lors de l'annotation avec des dépendances profondes (Candito *et al.*, 2014).

Le format FTB est proche du format SUD en ce sens que les mots fonctionnels sont la tête des expressions qu'ils introduisent, y compris les conjonctions de coordination. En revanche, les auxiliaires sont dépendants des verbes principaux auxquels ils s'appliquent, comme dans UD (figure 8). Alors que dans SUD, le nombre d'étiquettes principales est très réduit, il y en a beaucoup plus dans FTB. Les arguments verbaux sont représentés dans SUD avec la dépendance `comp`, alors que dans FTB, on spécifie `ato` (attribut de l'objet), `ats` (attribut du sujet), `a_obj` (objet indirect introduit par *à*), `de_obj` (objet indirect introduit par *de*), `obj` (objet direct) ou `p_obj.o` (objet indirect introduit par une autre proposition que *à* ou *de*).

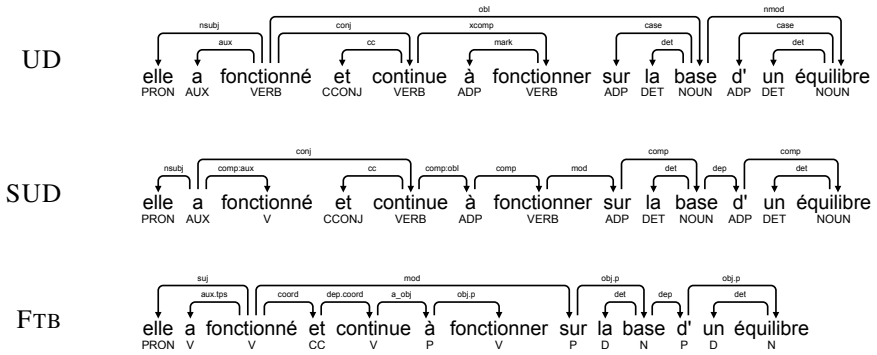

Figure 8. *Annotation de la phrase* elle a fonctionné et continue à fonctionner sur la base d'un équilibre *suivant les trois formats* UD*,* SUD *et* FTB

5.2. *La conversion du corpus* SEQUOIA *dans le format* UD

Les deux formats d'annotation FTB et UD contiennent essentiellement les mêmes informations, exprimées différemment. Il est donc possible d'envisager une conversion automatique de l'un vers l'autre. La plupart des travaux de conversion d'un corpus existant au schéma UD ont utilisé des règles de mapping automatiques, suivies de corrections manuelles (Nivre (2014) pour le corpus suédois Talbanken, Pyysalo *et al.* (2015) pour le corpus finlandais *Turku Dependency Treebank* (TDT), Attardi *et al.* (2016) pour l'*Italian Stanford Dependency Treebank* (ISDT), Taji *et al.* (2017) pour le *Penn Arabic Treebank*, Bouma et Van Noord (2017) pour le corpus néerlandais *Lassy*

Small, Cecchini *et al.* (2018) pour le corpus latin *Index Thomisticus*, Przepiórkowski et Patejuk (2019) pour le corpus polonais POLFIE). Dans certains corpus UD émanant de corpus existants, pour limiter la perte d'information, plusieurs relations secondaires ont été introduites (dans le TDT finlandais (Pyysalo *et al.*, 2015), le corpus polonais SZ (Wróblewska, 2018), le *treebank* hébreu HTB (Sade *et al.*, 2018)). Nous avons toutefois tenté d'éviter cette solution autant que possible.

Sur certains points, le format FTB est plus riche que celui de UD ; il fait par exemple systématiquement la distinction entre modificateur et argument pour les dépendants obliques des verbes, des adjectifs et des adverbes[18]. Cela a permis d'avoir systématiquement le typage `obl:mod` vs `obl:arg` dans UD_FRENCH-SEQUOIA.

Inversement, nous avons été confrontés à quelques cas dans lesquels SEQUOIA manquait d'information pour être correctement converti en UD. Par exemple, les deux exemples *faire installer des poubelles* et *faire jouer les enfants* étaient annotés de façon identique (*poubelles* objet de *installer* et *enfants* objet de *jouer*) alors que dans UD, on a la relation `obj` dans le premier cas et la relation `obj:agent` dans le second. Nous avons décidé d'enrichir l'annotation de départ du corpus SEQUOIA en ajoutant un trait *agent=y* uniquement dans le second cas pour le différencier du premier. Même si cette information est disponible dans l'annotation en syntaxe profonde (Candito *et al.*, 2014), nous avons décidé de ne pas faire la conversion depuis cette version enrichie, notamment pour permettre l'application de cette conversion à d'autres corpus pour lesquels l'annotation en syntaxe profonde ne serait pas disponible.

Comme souvent, lors de la conversion d'un corpus, l'étude des cas qui posent problème amène soit à modifier la conversion elle-même, soit à revoir l'annotation originale.

La conversion est codée sous la forme d'un système de réécriture de graphes qui contient environ 200 règles. Une grande partie de ces règles sont de simples changements d'étiquettes de partie du discours (par exemple, `N` devient `NOUN`) ou de noms des traits morphologiques (par exemple, *g=m* devient *Gender=Masc*). Les autres règles qui modifient réellement la structure en dépendances sont dans l'ordre d'application : le traitement des énumérations et des coordinations (15 règles), les changements de tête (11 règles), le traitement de la copule (1 règle) et les EP (9 règles). La règle de traitement de la copule (nommée `verb_ats`) est donnée dans la figure 9.

```
rule verb_ats {
  pattern {V[upos=VERB, lemma="être"]; e: V-[ats]->A}
  without {A[upos=VERB, VerbForm=Inf|Fin]}
  commands {del_edge e; shift V==>A; add_edge A-[cop]->V; V.upos=AUX}
}
```

Figure 9. *Règle* `verb_ats` *de traitement de la copule*

18. Pour les groupes prépositionnels dépendant de noms, le choix a été fait comme dans UD de ne pas trancher la distinction entre modificateur et argument car elle est trop difficile à annoter.

La clause `pattern` décrit les conditions dans lesquelles cette règle doit s'appliquer : le verbe *être* est la source d'une dépendance `ats`. La clause `without` est une condition négative qui permet de décrire une exception et bloque l'application de la règle si la cible de la dépendance `ats` est un verbe (la justification de cette condition négative est donnée à la sous-section 2.1.1). Si ces conditions sont vérifiées, on applique dans l'ordre les quatre commandes : suppression de l'arc repéré, changement de tête (tous les liens attachés au verbe *être* sont reportés sur l'attribut), ajout de l'arc `cop` en sens inverse et changement de la catégorie de *être* de `VERB` à `AUX`. La figure 10 donne un exemple d'application de la règle.

Figure 10. *Annotation d'une phrase avant et après l'application de la règle* `verb_ats`

Le fait de traiter la copule après les remplacements d'étiquettes permet d'avoir des règles plus simples pour les changements d'étiquettes qui doivent tenir compte de la catégorie du dépendant. Dans la phrase *Je dois être honnête* la relation entre *doit* et *être honnête* est changée en `xcomp` car la tête de *être honnête* est *être* avant l'application de la règle `verb_ats`. Si cette règle avait été appliquée avant, la tête de *être honnête* serait *honnête* et la prédiction de la relation `xcomp` nécessiterait un motif plus complexe.

6. Évaluation

6.1. *Comparaison avec une annotation de référence*

Parmi les méthodes d'évaluation, la comparaison avec une annotation de référence est essentielle car elle permet de confronter les annotations à l'avis d'expert. Malheureusement, cette évaluation est coûteuse en temps et pas toujours facile à réaliser.

Comme nous voulons observer l'évolution de la qualité de l'annotation au fil des versions, nous avons dû construire une référence avec des phrases issues de ce corpus. Nous avons donc extrait aléatoirement 108 phrases (soit 2 502 tokens) du corpus UD_FRENCH-GSD (partie `train`) pour lesquelles une annotation manuelle a été faite par trois annotateurs. Chaque phrase a été annotée par deux des trois auteurs et les différences ont fait l'objet de discussions communes pour construire l'annotation de référence utilisée ensuite pour l'évaluation. Le corpus ainsi obtenu est noté UD_FRENCH-GSD^{GOLD}.

Pour éviter que l'annotation actuelle des 108 phrases dans UD_FRENCH-GSD ne biaise la construction de la référence, les phrases ont été préannotées en utilisant un

	A1/A2	A1/A3	A2/A3	Moyenne
Résultats bruts	89,11	85,97	93,42	89,50
Après normalisation des dates	90,33	88,08	93,42	90,61

Tableau 4. *Accords entre annotateurs pour le corpus de référence (108 phrases)*

Version	2.0	2.1	2.2	2.3	2.4
Exactitude (%)	85,13	87,86	88,35	91,12	92.00

Tableau 5. *Exactitude pour les différentes versions du corpus* UD_FRENCH-GSD

analyseur à base de règles (Guillaume et Perrier, 2015) qui est indépendant du corpus. Inversement, les annotations de ce corpus UD_FRENCH-GSD^{GOLD}n'ont pas été utilisées dans d'autres parties du travail et notamment pour la détection d'erreurs. En revanche, un biais évident de notre évaluation est que les trois annotateurs impliqués dans la création de la référence le sont également dans la mise à jour du corpus.

Les valeurs d'exactitude calculées dans les tableaux 4 et 5 ont été obtenues avec le script `conll17_ud_eval.py` fourni avec la version 2.0[19]. La mesure utilisée est le *Weighted LAS* pour laquelle seul le poids de la ponctuation est mis à zéro.

Le tableau 4 donne les accords entre les annotateurs (%) pour les trois paires d'annotateurs. Une partie significative des différences avec l'annotateur A1 porte sur l'annotation des dates. La dernière ligne du tableau donne les scores d'accord si on ne tient pas compte de ces différences.

Dans le tableau 5 figurent les évaluations des différentes versions du corpus par rapport aux données de UD_FRENCH-GSD^{GOLD}. Les conventions d'annotation ayant changé lors de la version 2.0, il n'est pas pertinent d'évaluer les versions 1.x. La progression de l'exactitude des versions 2.0 à 2.4 par rapport au corpus UD_FRENCH-GSD^{GOLD} est significative et nous pensons que, malgré les biais évoqués plus haut, cela reflète une évolution de la qualité du corpus UD_FRENCH-GSD.

19. `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2184`

6.2. *Estimation de la cohérence de l'annotation à l'aide d'un analyseur syntaxique*

Une autre méthode permettant d'avoir des indications sur la qualité d'un corpus ou tout du moins sur la cohérence des annotations qui y figurent est d'entraîner un analyseur syntaxique et d'évaluer ses performances. Dans cette expérience, nous avons utilisé l'analyseur UDPIPE (Straka et Straková, 2017) que nous avons entraîné avec le sous-corpus `train`, puis utilisé pour analyser les deux autres sous-corpus disponibles (`dev` et `test`). Nous avons utilisé les paramètres par défaut de UDPIPE. En effet, notre but n'est pas d'optimiser les performances de l'analyseur mais d'évaluer l'évolution du corpus et donc d'utiliser les mêmes paramètres pour chacun de nos tests.

La figure 11 montre l'exactitude obtenue par l'analyseur UDPIPE sur les sous-corpus `dev` à gauche et `test` à droite. De nouveau malgré les biais inhérents à ce type d'évaluation, nous observons une progression significative qui montre que la cohérence interne du corpus a nettement progressé.

Ces évaluations montrent également une différence très nette entre les niveaux d'exactitude obtenus sur le sous-corpus `dev` d'une part et sur le sous-corpus `test` d'autre part. Cette différence était déjà manifeste dans les données de la version 1.1 (5,79 % de différence pour le LAS) et elle est moindre mais toujours présente dans la version 2.4 (2,13 %). Nous ne savons pas expliquer cette différence car nous n'avons pas d'information sur la façon dont les différents sous-corpus ont été construits dans les données originales. En revanche, dans nos modifications, nous avons toujours traité l'ensemble du corpus sans tenir compte de ce découpage.

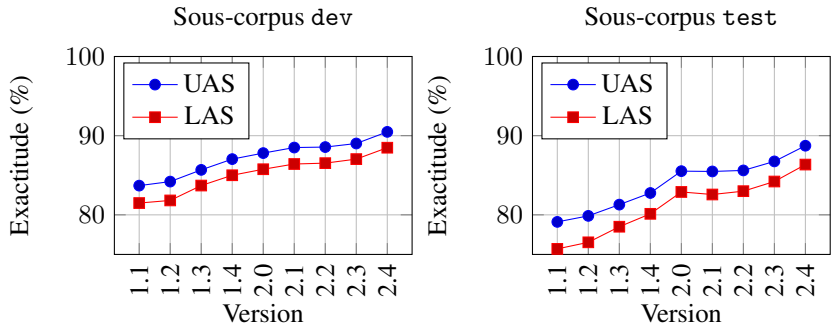

Figure 11. *Performance de* UDPIPE *entraîné puis évalué sur les différentes versions de UD_French-GSD*

Les mêmes expériences menées sur le corpus UD_FRENCH-SEQUOIA ne montrent pas d'évolution significative (les annotations ont très peu changé entre les versions 2.0 à 2.4). Toutes les valeurs de UAS sont entre 87,27 et 88,05 et celles de LAS entre 85,61 et 86,59. Ces valeurs sont inférieures de 1 à 3 points par rapport à celles obtenues pour les versions récentes de UD_FRENCH-GSD, sans doute car le

corpus d'apprentissage est nettement plus petit (50 536 tokens pour UD_FRENCH-SEQUOIA contre 354 655 tokens pour UD_FRENCH-GSD).

7. Conclusion

Dans cet article, nous avons décrit le travail effectué pour faire évoluer des corpus existants. Les deux corpus ont fait l'objet de démarches assez différentes. Pour UD_FRENCH-GSD, nous avons travaillé uniquement sur les données de la version 1.1 de UD et nous ne disposions pas des données annotées initialement avant conversion ni des métadonnées. Pour UD_FRENCH-SEQUOIA, en revanche, nous avons essentiellement travaillé sur la question de la conversion du format initial du corpus vers le format UD. Pour évaluer la qualité d'une annotation, il est courant d'avoir recours à une annotation par un expert sur une petite partie des données qui servent alors de référence. Nous avons appliqué cette méthode au corpus UD_FRENCH-GSD. Sur le même corpus, nous avons également observé la cohérence des annotations en entraînant un analyseur syntaxique sur une partie de données et en l'appliquant au reste. Ces deux évaluations montrent une progression de la cohérence et de la qualité du corpus.

Le contexte du projet UD permet aujourd'hui un travail beaucoup plus systématique pour harmoniser les données de différents corpus et les rendre plus faciles à comparer. Même si l'on reste dans un cadre monolingue, comme c'est le cas dans cet article, l'harmonisation entre les corpus développés par différentes équipes reste une question compliquée et qui nécessite beaucoup de concertation. Nous avons largement entamé cette harmonisation en collaboration avec les collègues travaillant sur d'autres corpus du français, mais il reste encore beaucoup à faire pour améliorer la situation pour les prochaines versions. Une suite naturelle à ce travail serait également d'élargir cette harmonisation vers d'autres langues et au moins, dans un premier temps, aux autres corpus en langue romane. Une autre direction pour enrichir les données décrites ici sera d'annoter dans les corpus du français les *Enhanced Universal Dependencies*[20] qui facilitent l'utilisation des corpus pour l'analyse sémantique de la langue.

Remerciements

Les auteurs remercient Alane Shur, Matias Grioni et Carly Dickerson qui ont participé à l'annotation de UD_FRENCH-GSD ; ils remercient également l'un des relecteurs et le comité de la revue dont les commentaires détaillés et pertinents ont permis d'améliorer le document. Ce travail a bénéficié de l'infrastructure du CPER LCHN (Contrat Plan État Région « Langues, Connaissances & Humanités Numériques »), ainsi que d'un *Google Faculty Research Award* attribué à Marie-Catherine de Marneffe.

20. `http://universaldependencies.org/u/overview/enhanced-syntax.html`

8. Bibliographie

Abeillé A., Barrier N., « Enriching a French Treebank. », *Proceedings of LREC 2004*, 2004.

Alzetta C., Dell'Orletta F., Montemagni S., Simi M., Venturi G., « Assessing the Impact of Incremental Error Detection and Correction. A Case Study on the Italian Universal Dependency Treebank », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, p. 1-7, 2018a.

Alzetta C., Dell'Orletta F., Montemagni S., Venturi G., « Dangerous Relations in Dependency Treebanks », *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, p. 201-210, 2018b.

Attardi G., Saletti S., Simi M., « Evolution of Italian Treebank and Dependency Parsing towards Universal Dependencies », *Proceedings of the Second Italian Conference on Computational Linguistics*, p. 25-30, 2016.

Bedaride P., Gardent C., « Semantic Normalisation : a Framework and an Experiment », *8th International Conference on Computational Semantics - IWCS 2009*, Tilburg, Netherlands, January, 2009.

Blanche-Benveniste C., Deulofeu J., Stéfanini J., Van Den Eynde K., *Pronom et syntaxe : l'approche pronominale et son application au français*, vol. 1, Peeters Publishers, 1987.

Bohnet B., Wanner L., « On using a parallel graph rewriting formalism in generation », *EWNLG '01 : Proceedings of the 8th European workshop on Natural Language Generation*, Association for Computational Linguistics, p. 1-11, 2001.

Bonfante G., Guillaume B., Perrier G., *Application of Graph Rewriting to Natural Language Processing*, John Wiley & Sons, 2018.

Bouma G., Hajic J., Haug D., Nivre J., Solberg P. E., Øvrelid L., « Expletives in Universal Dependency Treebanks », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, p. 18-26, 2018.

Bouma G., Van Noord G., « Increasing Return on Annotation Investment : The Automatic Construction of a Universal Dependency Treebank for Dutch », *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, Association for Computational Linguistics, p. 19-26, 2017.

Boyd A., Dickinson M., Meurers W. D., « On detecting errors in dependency treebanks », *Research on Language & Computation*, vol. 6, n° 2, p. 113-137, 2008.

Candito M., Constant M., « Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Baltimore, Maryland, p. 743-753, June, 2014.

Candito M., Perrier G., Guillaume B., Ribeyre C., Fort K., Seddah D., Villemonte De La Clergerie É., « Deep Syntax Annotation of the Sequoia French Treebank », *Proceedings of LREC 2014*, Reykjavik, Iceland, may, 2014.

Candito M., Seddah D., « Effectively long-distance dependencies in French : annotation and parsing evaluation », *TLT 11-The 11th International Workshop on Treebanks and Linguistic Theories*, 2012a.

Candito M., Seddah D., « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. », *TALN 2012*, Grenoble, France, 2012b.

Cecchini F. M., Passarotti M., Marongiu P., Zeman D., « Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, ACL, p. 27-36, 2018.

Chaumartin F.-R., Kahane S., « Une approche paresseuse de l'analyse sémantique ou comment construire une interface syntaxe-sémantique à partir d'exemples », *TALN 2010, Montreal, Canada*, 2010.

de Marneffe M.-C., Grioni M., Kanerva J., Ginter F., « Assessing the annotation consistency of the universal dependencies corpora », *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, p. 108-115, 2017.

de Marneffe M.-C., MacCartney B., Manning C. D., « Generating Typed Dependency Parses from Phrase Structure Parses », *Proceedings of LREC 2006*, 2006.

De Smedt K., Rosén V., Meurer P., « Studying Consistency in UD Treebanks with INESS-Search », *Fourteenth Workshop on Treebanks and Linguistic Theories (TLT14)*, 2016.

Dell'Orletta F., Giulia V., Montemagni S., « Linguistically-driven Selection of Correct Arcs for Dependency Parsing », *Computacion y Sistemas*, vol. 2, p. 125-136, 2013.

Dickinson M., Meurers W. D., « Detecting Inconsistencies in Treebanks », *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT-03)*, 2003.

Dickinson M., Meurers W. D., « Detecting Errors in Discontinuous Structural Annotation », *Proceedings of the 43rd Annual Meeting of the ACL*, p. 322-329, 2005.

Gerdes K., Guillaume B., Kahane S., Perrier G., « SUD or Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD », *Universal Dependencies Workshop 2018 (UDW 2018)*, Bruxelles, Belgique, November, 2018.

Gerdes K., Kahane S., « Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe », *Atelier ACor4French, Actes de la 24e conférence sur le traitement automatique des langues (TALN)*, Orléans, p. 1-9, 2017.

Guillaume B., Perrier G., « Dependency Parsing with Graph Rewriting », *IWPT 2015, 14th International Conference on Parsing Technologies*, 14th International Conference on Parsing Technologies - Proceedings of the Conference, Bilbao, Spain, p. 30-39, 2015.

Hajič J., Hajičová E., Mikulová M., Mírovskỳ J., « Prague Dependency Treebank », *Handbook of Linguistic Annotation*, Springer, p. 555-594, 2017.

Hyvönen E., « Semantic Parsing as Graph Language Transformation - a Multidimensional Approach to Parsing Highly Inflectional Languages », *COLING*, p. 517-520, 1984.

Jijkoun V., de Rijke M., « Learning to Transform Linguistic Graphs », *Second Workshop on TextGraphs : Graph-Based Algorithms for Natural Language Processing, Rochester, NY, USA*, 2007.

McDonald R. T., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K. B., Petrov S., Zhang H., Täckström O., Bedini C., Castelló N. B., Lee J., « Universal Dependency Annotation for Multilingual Parsing. », *ACL (2)*, ACL, p. 92-97, 2013.

Mel'čuk I., *Dependency Syntax : Theory and Practice*, Albany, N.Y. : The SUNY Press, 1988.

MoDyCo, LaTTiCe, CLLE-ERSS, LPL, IRCAM, « TREEBANK RHAPSODIE », , `https://hdl.handle.net/11403/rhapsodie/v1`, 2018.

Nivre J., « Universal Dependencies for Swedish », *Proceedings of the Swedish Language Technology Conference (SLTC)*, 2014.

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDonald R., Petrov S., Pyysalo S., Silveira N. *et al.*, « Universal dependencies v1 : A multilingual treebank collection », *Proceedings of LREC 2016*, p. 1659-1666, 2016.

Osborne T., Gerdes K., « The status of function words in dependency grammar : A critique of Universal Dependencies (UD) », *Glossa*, vol. 4, nº 1, p. 1-28, 2019.

Przepiórkowski A., Patejuk A., « From Lexical Functional Grammar to enhanced Universal Dependencies », *Language Resources and Evaluation*, Feb, 2019.

Pyysalo S., Kanerva J., Missilä A., Laippala V., Ginter F., « Universal Dependencies for Finnish », *Proceedings of NODALIDA 2015*, p. 163-172, 2015.

Ribeyre C., Méthodes d'analyse supervisée pour l'interface syntaxe-sémantique, PhD thesis, Université Paris Diderot, 2016.

Sade S., Seker A., Tsarfaty R., « The Hebrew Universal Dependency Treebank : Past Present and Future », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, p. 133-143, 2018.

Sagot B., « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French », *Proceedings of LREC 2010*, La Valette, Malte, mai, 2010.

Sanguinetti M., Bosco C., *PartTUT : The Turin University Parallel Treebank*, Springer International Publishing, Cham, p. 51-69, 2015.

Seddah D., Candito M., « Hard time parsing questions : Building a QuestionBank for French », *Proceedings of LREC 2016*, 2016.

Seddah D., Villemonte De La Clergerie É., Sagot B., Martinez Alonso H., Candito M., « Cheating a Parser to Death : Data-driven Cross-Treebank Annotation Transfer », *Proceedings of LREC 2018*, Miyazaki, Japan, mai, 2018.

Straka M., Straková J., « Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe », *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Vancouver, Canada, p. 88-99, August, 2017.

Taji D., Habash N., Zeman D., « Universal Dependencies for Arabic », *Proceedings of the Third Arabic Natural Language Processing Workshop*, ACL, p. 166-176, 2017.

Wisniewski G., « Errator : a Tool to Help Detect Annotation Errors in the Universal Dependencies Project », *Proceedings of LREC 2018*, p. 4489-4493, 2018.

Wróblewska A., « Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format », *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, p. 173-182, 2018.

Zeman D., « Slavic Languages in Universal Dependencies », *Proceedings of Slovko 2015 : Natural Language Processing, Corpus Linguistics, E-learning*, 2015.

Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S., « CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, p. 1-21, 2018.

Zeman D., Popel M., Straka M., Hajic J., Nivre J., Ginter F., *et al.*, « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Vancouver, Canada, p. 1-19, August, 2017.