Réflexion sur l'annotation de corpus écrits du français en syntaxe et en sémantique

Bruno Guillaume, Guy Perrier

Réflexion sur l'annotation de corpus écrits du français en syntaxe et en sémantique

Bruno Guillaume[1] Guy Perrier[2]
(1) LORIA - Inria, Campus Scientifique BP 239, 54506 Vandœuvre-lès-Nancy cedex, France
(2) LORIA - Université de Lorraine, Campus Scientifique BP 239, 54506 Vandœuvre-lès-Nancy cedex, France
bruno.guillaume@loria.fr, guy.perrier@loria.fr

RÉSUMÉ
Cet article est une réflexion sur l'annotation de corpus écrits du français en syntaxe et en sémantique, réflexion qui est le fruit de travaux menés sur les corpus SEQUOIA et UD-FRENCH.

ABSTRACT
Reflection on the annotation of written French corpora in syntax and semantics

This article is a reflection on the annotation of written French corpora in syntax and semantics, reflection which is the fruit of work carried out on the corpus SEQUOIA and UD-FRENCH.

MOTS-CLÉS : annotation de corpus, syntaxe, sémantique.

KEYWORDS: corpus annotation, syntax, semantics.

Introduction

Dans cet article, nous nous intéressons à la syntaxe et à la sémantique des langues pour l'écrit. Dans ce domaine, que ce soit pour les linguistes qui travaillent en linguistique de corpus ou pour les informaticiens qui travaillent en TAL à l'aide de méthodes d'apprentissage, il est nécessaire de disposer de corpus de taille importante annotés en syntaxe et en sémantique avec une qualité suffisante.

L'état de l'art pour le français n'est actuellement pas satisfaisant :

— Un des plus anciens et des plus connus corpus du français annoté en syntaxe est le French Treebank (FTB) (Abeillé *et al.*, 2003) [1]. Il est composé de 21 550 phrases extraites du Journal *Le Monde*. Il est annoté en constituants et cette annotation a été convertie de façon semi-automatique en syntaxe de dépendances (Candito *et al.*, 2009). Malheureusement, le FTB n'est pas librement disponible, ce qui est un handicap pour sa diffusion et sa correction. En outre, il n'existe pas de versions référencées et la conversion en dépendances n'a jamais été corrigée systématiquement manuellement.

— Le corpus SEQUOIA [2] est formé de 3 099 phrases issues de plusieurs domaines (presse, médical, Parlement Européen et wikipédia). Initialement, il a été annoté en constituants selon le schéma du FTB, puis cette annotation a été convertie en dépendances (Candito & Seddah,

1. http://ftb.linguist.univ-paris-diderot.fr/
2. https://deep-sequoia.inria.fr/

2012) (avec la même procédure que pour la conversion du FTB). Une annotation en syntaxe profonde a été ensuite ajoutée à la syntaxe de surface (Perrier *et al.*, 2014). Une limite de SEQUOIA est sa petite taille.

— Le corpus UD-FRENCH [3] est formé de 16 329 phrases couvrant aussi plusieurs types de textes (nouvelles de Google, blogs, wikipédia, commentaires d'utilisateurs). L'origine de ce corpus est le GOOGLEDATASET (McDonald *et al.*, 2013) qui a été validé manuellement par un annotateur. La première version du corpus UD-FRENCH (la version 1.1) a été produite automatiquement à partir du GOOGLEDATASET et n'a pas fait l'objet d'une validation systématique. UD-FRENCH est annoté selon le schéma du projet *Universal Dependencies (UD)* dont l'objectif est d'annoter en syntaxe de dépendances un maximum de langues selon un format commun. Le corpus UD-FRENCH est progressivement corrigé mais il subsiste encore beaucoup d'erreurs et le schéma d'annotation est en plein développement avec des flous pour certains phénomènes.

— Le corpus UD-FRENCH-PARTUT [4] contient des annotations en dépendances de surface sur l'anglais, le français et l'italien. Il a récemment été converti au format UD. La partie française comprend 620 phrases couvrant 6 domaines différents.

Tous ces corpus sont annotés en syntaxe. En matière d'annotation sémantique, les ressources en français sont encore plus pauvres. Le projet ASFALDA a permis d'ajouter une couche sémantique aux corpus du FTB et de SEQUOIA suivant le schéma de FrameNet (Djemaa *et al.*, 2016) mais seule la partie SEQUOIA est librement disponible.

Les réflexions que nous allons présenter maintenant sont le fruit d'une expérience accumulée sur 4 ans en annotation de corpus. Nous avons contribué à l'annotation en syntaxe profonde de SEQUOIA et nous continuons à en corriger à la fois l'annotation en dépendances de surface et celle en dépendances profondes. Nous participons au développement de l'UD-FRENCH à la fois en corrigeant ponctuellement les annotations et en contribuant à la réflexion sur le schéma d'annotation. À partir de l'annotation en syntaxe profonde de SEQUOIA, nous avons produit de façon automatique deux annotations sémantiques de ce corpus, la première selon le format de l'*Abstract Meaning Representation* (AMR) (Banarescu *et al.*, 2012) et la seconde selon le format de la *Dependency Minimal Recursion Semantics* (DMRS) (Copestake, 2009) [5]. Ces annotations n'ont pas été validées manuellement.

1 Des choix de base importants pour l'annotation d'un corpus

Quand on entame l'annotation d'un corpus, il est des choix qui peuvent être lourds de conséquences. Il est essentiel que le corpus et son annotation soient librement disponibles. C'est un facteur important pour favoriser la diffusion du corpus et inciter les utilisateurs à contribuer à sa correction. Il est alors utile, comme c'est la règle dans le projet UD, de publier à intervalles réguliers des versions révisées du corpus (deux fois par an pour UD).

Il est important de diversifier les types de corpus afin d'éviter que l'annotation ne soit biaisée par un type particulier de texte : articles de journaux, textes littéraires, notices techniques, messages ... Cela permet aussi de répondre aux besoins des utilisateurs qui sont très divers. En plus, il vaut mieux choisir des textes complets plutôt que des phrases isolées. Cela ne coûte rien et ces corpus peuvent

3. http://universaldependencies.org/

4. http://www.di.unito.it/~tutreeb/partut.html

5. La DMRS est une version compacte de la *Minimal Recursion Semantics* (MRS) dans laquelle les représentations sémantiques sont des graphes de dépendances entre sens des mots pleins.

alors être utilisés pour le traitement du discours (structures rhétoriques, résolution des anaphores…). Par ailleurs, l'annotation en structures discursives de ces corpus pourra ainsi s'appuyer la présence d'une annotation syntaxique et éventuellement sémantique.

Dans la construction d'une annotation, qu'elle soit syntaxique ou sémantique, qu'on utilise une méthode statistique ou une méthode symbolique, il est souvent utile de faire appel à des lexiques, notamment lorsque l'on vise une annotation de qualité. C'est particulièrement vrai au niveau sémantique où le lien avec les lexiques est souvent conservé dans le résultat de l'annotation. Il est alors important de veiller à ce que les lexiques aient une large couverture, que les informations qu'ils contiennent soient suffisamment fiables et riches.

Un même corpus peut être utilisé pour plusieurs annotations qui ne suivent pas forcément les mêmes choix linguistiques. Il est important que la segmentation en phrases et en tokens soit indépendante de ces choix. Un mauvais exemple est la façon dont SEQUOIA était initialement segmenté en tokens. Les expressions multi-mots étaient considérées comme formant un seul token. Il s'est avéré par la suite qu'il aurait mieux valu considérer certaines expressions multi-mots comme de simples expressions et inversement ; parfois, il aurait été aussi utile d'exhiber leur structure interne. Il a fallu revoir la segmentation et donc aussi en partie l'annotation syntaxique des phrases concernées. Il a alors été décidé de choisir des tokens minimaux sans se préoccuper de considérations linguistiques en prenant systématiquement pour séparateurs de tokens les espaces. Les signes de ponctuation sont considérés comme des tokens à l'exception de l'apostrophe et du tiret. L'apostrophe est accolée au mot qui la précède immédiatement mais la partie qui suit est dans un token séparé. Pour le tiret, la segmentation dépend du lexique : le tiret peut-être complètement intégré au token (*après-midi*), intégré à la partie gauche (*politico-*), intégré à la partie droite (*-il* dans *se souvient-il*) et séparé quand il est en début de phrase ou qu'il joue le rôle de parenthèses.

Le fait de segmenter systématiquement sur les espaces laisse ensuite toute liberté de traiter les expressions multi-mots comme on l'entend [6]. Nous pensons que cette façon de segmenter peut être adoptée comme règle générale pour n'importe quel corpus de l'écrit, peut-être en forçant la segmentation isolé du tiret pour ne pas dépendre de lexique lors de la segmentation.

Il est une question qui interfère avec la segmentation en tokens, c'est la façon de traiter les amalgames (*de le* en *du*, *de lequel* en *duquel*…). Doit-on les dissocier, comme le fait UD, dans un processus qui précède l'annotation ou doit-on les garder tels quels, comme le fait SEQUOIA ? Chaque solution a ses avantages et ses inconvénients : la première évite les cas particuliers dans l'analyse syntaxique mais nécessite un mécanisme supplémentaire pour garder la trace de la phrase telle qu'elle est présente dans le corpus d'origine ; pour la seconde c'est l'inverse avec une rupture dans le traitement uniforme de la syntaxe [7].

Il est habituel de conserver tout le texte du corpus brut initial, sans en écarter des parties a priori. Cependant cela conduit a gérer des parties de texte qui n'ont pas de structures syntaxiques en français : texte en langue étrangère, référence bibliographique, utilisation détournée de la typographie… Par homogénéité, en pratique, une pseudo-annotation par défaut est appliquée mais souvent de façon

6. Cela ne règle pas toutes les questions portant sur les expressions multi-mots. En particulier, il est souvent nécessaire d'affecter des traits linguistiques à l'expression qu'il n'est pas possible de retrouver à partir de ses composants. Ainsi *savoir-faire* est un nom alors que *savoir* et *faire* sont des verbes. Il faut pouvoir renseigner cette information, ce que le format de UD ne prévoit pas.

7. Nous avons une préférence pour la première en allant jusqu'à dissocier systématiquement *du* en *de le* même pour l'article partitif. Cela évite l'analyse qu'il est nécessaire de faire en amont de la segmentation quand on veut distinguer l'article partitif de la préposition amalgamée.

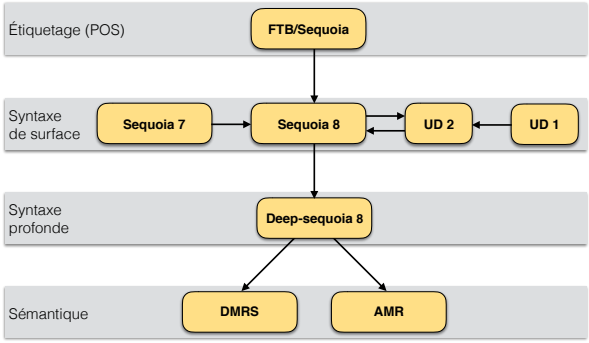

FIGURE 1 – Formats d'annotation

incohérente. On comprend bien le danger de supprimer ces occurrences dans les corpus produits, le plus raisonnable serait de notre point de vue, d'ajouter des annotations qui permettent de les supprimer ou de les garder en fonction des usages du corpus.

2 Les formats d'annotation

C'est le guide d'annotation qui définit le format d'annotation. Il est important qu'il soit suffisamment complet pour couvrir la plupart des phénomènes, mais en même temps, il faut qu'il soit facile d'utilisation. On sait très bien que si nous voulons un guide exhaustif, celui-ci devient vite volumineux mais il faut garder un certain équilibre entre l'exhaustivité et la facilité d'utilisation. De la qualité du guide, dépend la qualité de l'annotation. Le fait que le guide d'UD reste muet sur des phénomènes importants (causatives, comparaisons, clivées…) n'est peut-être pas étranger à ce que le corpus UD-FRENCH contienne encore beaucoup d'incohérences.

La figure 1 présente les formats que nous avons utilisé dans nos différents travaux sur les corpus. La multiplicité des formats d'annotation est inévitable car d'une part les besoins sont très divers et d'autre part, les choix linguistiques qui président à ces formats peuvent être divergents. Pour ne prendre que la syntaxe en dépendances, SEQUOIA considère plutôt les mots de liaison (prépositions, conjonctions) comme gouvernant les mots sémantiquement pleins qu'ils introduisent et UD a fait le choix inverse. D'autres divergences sont plus profondes, dans UD, il n'y a pas de distinction entre les compléments requis par les verbes et les modificateurs, alors que c'est le cas dans SEQUOIA. Pour convertir une annotation du format UD dans le format SEQUOIA, on peut certes faire appel à un lexique pour pallier le manque d'information, ce qui ne résoud pas forcément tous les cas. Et encore, UD et SEQUOIA sont tous les deux des formats syntaxiques en dépendance.

Les divergences sont encore plus grandes au niveau sémantique. Comme il est extrêmement compliqué d'avoir une annotation qui soit la plus complète et la plus fine possible, chaque format privilégie un aspect au détriment des autres. Ainsi, l'AMR (Banarescu *et al.*, 2012) cherche avant tout à modéliser les relations prédicat-argument de la façon la plus précise possible alors que la MRS (Copestake *et al.*, 2005) s'attache à représenter de façon sous-spécifiée la structure logique de la phrase. Dans ces conditions, il semble difficile d'envisager des outils de conversion d'un format dans l'autre. Comme en général, une annotation sémantique s'obtient par transformation d'une annotation syntaxique,

pour chaque format sémantique, il faut concevoir un outil particulier. Un moyen de factoriser une partie du travail nécessaire consiste à utiliser un niveau intermédiaire commun. Plusieurs théories linguistiques distinguent en syntaxe deux niveaux : la *syntaxe de surface* de la *syntaxe profonde* (Sgall *et al.*, 1986; Mel'čuk, 1988). La syntaxe de surface est proche de la forme phonologique de la phrase alors que la syntaxe profonde est une abstraction de la syntaxe de surface vers la sémantique. Dans le cadre de l'annotation du corpus SEQUOIA, (Perrier *et al.*, 2014) ont conçu leur propre niveau de *syntaxe profonde*. Ce niveau intermédiaire facilite la construction d'outils de transformation de formats syntaxiques en formats sémantiques. En effet, il suffit de concevoir un outil qui permette de passer de la syntaxe de surface à la syntaxe profonde et ensuite un outil particulier à chaque format sémantique pour passer de la syntaxe profonde à la sémantique. L'avantage est que la transformation de la syntaxe profonde vers la sémantique est plus simple que la transformation directe de la syntaxe de surface vers la sémantique.

L'exemple du passage d'une annotation au niveau syntaxique à une annotation au niveau sémantique illustre le fait qu'il est souvent plus complexe de s'appuyer sur une annotation à un niveau pour produire une annotation à un niveau voisin que d'effectuer une conversion de format en restant au même niveau. De ce point de vue, le format d'entrée et le format de sortie ne sont pas indifférents.

3 Quels outils pour annoter ?

Pour créer un corpus de qualité, l'intervention manuelle est indispensable et les outils pour produire ou corriger manuellement des annotations existantes sont nombreux (par exemple, BRAT [8] ou ARBORATOR (Gerdes, 2013) pour l'annotation en dépendances syntaxiques). Plus généralement, lors d'une campagne d'annotation, on met généralement en place un processus en plusieurs étapes avec plusieurs annotateurs sur les mêmes données dans un premier temps puis le travail d'un expert pour trancher les cas pour lesquels les premiers annotateurs n'ont pas donné des avis identiques. Il existe des outils pour aider à la gestion de ce type de campagne d'annotation. WEBANNO (Yimam *et al.*, 2013) par exemple est construit au-dessus de BRAT, il permet à un administrateur de gérer les utilisateurs, de rendre les données à annoter (ou corriger) disponibles à ces utilisateurs, de suivre la progression des annotateurs et enfin, dans une interface dédiée de comparer les annotations et de trancher dans les cas de divergences. Ces outils sont indispensables pour faire une annotation ou une vérification manuelle systématique, phrase par phrase mais ils ne sont pas toujours suffisants d'autres types d'interventions sur les corpus.

Nous présentons ici quelques outils complémentaires pour faire de la pré-annotation, de la conversion entre formats, des corrections automatiques ou pour repérer des cas à trancher manuellement. Ces outils sont également utilisés pour détecter des incohérences dans les corpus ou pour étudier des phénomènes linguistiques particuliers. Notre proposition est d'utiliser la réécriture de graphes pour gérer les conversions entre différents niveaux de représentation linguistique ou entre différents formats d'un même niveau. La réécriture de graphes permet de décrire ces conversions en les décomposant en transformations élémentaires et locales. L'outil GREW [9] (Guillaume *et al.*, 2012) est une implantation de ce modèle qui permet d'appliquer des ensembles de règles de réécriture pour transformer ou mettre à jour une structure. Les ensembles de règles sont organisés en modules qui permettent de gérer les stratégies d'application et notamment d'imposer des contraintes sur l'ordre d'application des règles.

8. http://brat.nlplab.org/
9. http://grew.loria.fr

Dans la figure 1, chaque flèche représente un système de règles de réécriture que nous avons écrit pour GREW afin de passer d'un format à un autre.

Les types de changements à prendre en compte lors de ces conversions peuvent être très simples comme des renommages d'étiquette (`mwe` en UD1 est devenu `fixed` en UD2). Mais ils peuvent également être plus complexes comme des changements de tête évoqués plus haut : pour le format SEQUOIA, la tête d'un groupe prépositionnel est la préposition alors que pour UD, c'est le nom. La conversion d'un format à l'autre demande donc une modification plus profonde de la structure syntaxique. Ces modifications de structures peuvent être décrites dans le formalisme de la réécriture de graphes. En effet, GREW dispose d'une commande `shift` dont l'effet est de déplacer toutes les arêtes incidentes à un nœud vers un autre nœud, ce qui correspond à la modification de structure induite par un changement de tête syntaxique dans un constituant.

La gestion des expressions multi-mots a changé dans la transformation SEQUOIA 7 vers SEQUOIA 8. Ainsi, le token *en_particulier* de SEQUOIA 7 qui a la catégorie `ADV` doit être remplacé par les deux tokens *en* et *particulier* dans SEQUOIA 8 avec les catégories `P` et `ADJ` respectivement. De l'information lexicale est donc nécessaire pour annoter les bonnes catégories pour les nouveaux tokens. L'outil GREW permet de contrôler l'application des règles par des informations lexicales et donc de prendre en compte ces modifications.

Malheureusement, certains changements dans les formats d'annotation induisent des modifications qui ne peuvent pas être automatisées et qui demandent donc une annotation manuelle. Là encore, avec nos outils de réécriture de graphe, nous pouvons repérer les cas à trancher manuellement et ainsi faciliter le travail de l'expert. Dans la conversion de UD1 vers UD2, la relation `nmod` qui était utilisée pour décrire tous les modificateurs nominaux a été raffinée en deux relations, notés `nmod` si le gouverneur de la relation est nominal et `obl` si le gouverneur de la relation est une clause. En cas de construction avec copule, il n'est pas possible de déterminer la bonne étiquette par une règle. En effet, dans la phrase *"Max est président de la république depuis 5 ans"*, les deux compléments *"de la république"* et *"depuis 5 ans"* sont tous deux `nmod` dans le format UD1 et leur traitement doit être différent : *"de la république"* se rapporte à *président* et reste `nmod` alors que *"depuis 5 ans"* modifie toute la proposition *"Max est président"*, il est donc `obl` bien que son gouverneur soit la tête de la proposition *"Max est président"*, c'est-à-dire le nom *président*.

Quelle que soit la façon dont un corpus a été conçu et annoté, il est toujours susceptible de contenir des erreurs ou des incohérences. Pour les corriger, il est utile de pouvoir chercher et visualiser de façon systématique sur l'ensemble d'un corpus les occurrences comparables d'un token, d'une relation ou d'une combinaison plus complexe d'un ensemble de tokens et de relations entre eux. Avec la réécriture de graphes définie dans l'outil GREW, la partie *reconnaissance de motifs* (qui est à la base de la réécriture) peut également être utilisée de façon autonome pour ce type de recherche. En pratique, c'est souvent à partir d'une annotation problématique que l'on peut vérifier la cohérence des occurrences similaires. Nous proposons une interface en ligne GREW-WEB [10] (voir par exemple figure 2 : la recherche de verbes sans sujet dans la version 2.0 de UD) pour faciliter la recherche de motifs sur une série de corpus librement disponibles.

À noter que cela est aussi utile pour faire de la recherche linguistique, pour tester des hypothèses ou pour trouver des exemples suivant un motif syntaxique précis. GREW-WEB est utile également en complément d'un guide d'annotation pour comprendre au travers d'exemples comment les parties implicites du guide sont réellement mis en œuvre dans l'annotation.

10. `grew.loria.fr/demo`

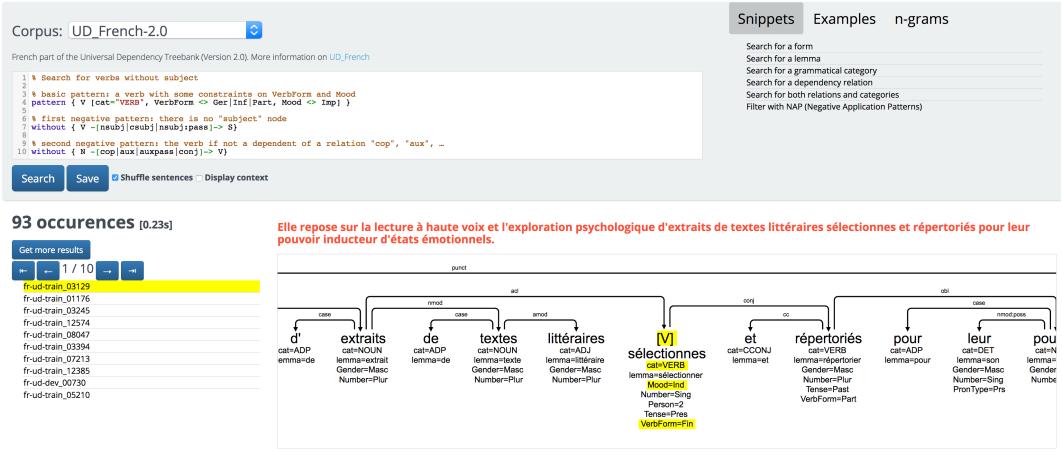

FIGURE 2 – Interface Web de recherche de motifs

Conclusion

Une limite de la réflexion présentée ici est qu'elle ne s'appuie que sur les corpus de l'écrit alors que nous sommes bien conscients qu'il serait intéressant de partager les leçons tirées de l'écrit avec celles tirées de l'oral, mais ses limites sont aussi celles de notre expérience. Néanmoins, nous espérons que les enseignements que nous en avons tirés seront utiles à la communauté pour combler le retard en matière de corpus du français écrit annotés en syntaxe et en sémantique.

Pour donner un coup de fouet à cette tâche et pour éviter la dispersion des forces, il serait utile de se mettre d'accord sur un programme minimum. Il semble raisonnable de chercher à disposer d'une plateforme commune de corpus de textes écrits, librement disponibles, chaque corpus pouvant être annotés selon différents formats. Même si on a plusieurs annotations pour un même corpus, il est important que la segmentation soit la même pour toutes les annotations.

Ensuite, il semble plus difficile de converger sur les formats d'annotation. En syntaxe, il y a un format qui a tendance à prendre une place prédominante, même s'il est encore en devenir, c'est celui d'UD. Pour le français, il nous semble néanmoins important de conserver le format de SEQUOIA qui a fait d'autres choix qui sont tout aussi justifiés.

Le format d'UD évolue vers la prise en compte d'un niveau de syntaxe profonde, c'est ce qu'on appelle *Enhanced* UD [11]. Il sera intéressant de comparer la syntaxe profonde d'UD avec la syntaxe profonde de SEQUOIA.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In *Treebanks. Building and Using Parsed Corpora*, chapter 10. Kluwer Academic Publishers.

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

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2012). Abstract meaning representation (amr) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs." In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle : ACL*, p. 1533–1544.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012*, Grenoble, France.

CANDITO M.-H., CRABBÉ B., DENIS P. & GUÉRIN F. (2009). Analyse syntaxique statistique du français : des constituants aux dépendances. In *TALN 2009*, Senlis, France.

COPESTAKE A. (2009). Slacker semantics : why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1–9, Athens, Greece : Association for Computational Linguistics.

COPESTAKE A., FLICKINGER D., POLLARD C. & SAG I. A. (2005). Minimal recursion semantics : An introduction. *Research on Language and Computation*, **3**(2-3), 281–332.

DJEMAA M., CANDITO M., MULLER P. & VIEU L. (2016). Corpus annotation within the french framenet : a domain-by-domain methodology. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

GERDES K. (2013). Collaborative dependency annotation. In *DepLing 2013*, volume 88, p. 88–97.

GUILLAUME B., BONFANTE G., MASSON P., MOREY M. & PERRIER G. (2012). Grew : un outil de réécriture de graphes pour le TAL. In *Proc. of TALN*, Grenoble, France.

MCDONALD R. T., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K. B., PETROV S., ZHANG H., TÄCKSTRÖM O. *et al.* (2013). Universal dependency annotation for multilingual parsing. In *ACL 2013, Sofia, Bulgaria*, p. 92–97 : Citeseer.

MEL'ČUK I. (1988). *Dependency Syntax : Theory and Practice*. Albany, N.Y. : The SUNY Press.

PERRIER G., CANDITO M., GUILLAUME B., RIBEYRE C., FORT K. & SEDDAH D. (2014). Un schéma d'annotation en dépendances syntaxiques profondes pour le français. In *TALN - Traitement Automatique des Langues Naturelles*, p. 574–579, Marseille, France.

SGALL P., HAJICOVÁ E. & PANEVOVÁ J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

YIMAM S. M., GUREVYCH I., ECKART DE CASTILHO R. & BIEMANN C. (2013). Webanno : A flexible, web-based and visually supported system for distributed annotations. In *Actes de Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 1–6, Sofia, Bulgarie : Association for Computational Linguistics.