CDGFr, un corpus en dépendances non-projectives pour le français

Denis Béchet[1] Ophélie Lacroix[2]
(1) LINA, 44322 Nantes Cedex 3, France
(2) LIMSI-CNRS, 91403 Orsay Cedex, France
denis.bechet@univ-nantes.fr, ophelie.lacroix@limsi.fr

Résumé. Dans le cadre de l'analyse en dépendances du français, le phénomène de la non-projectivité est peu pris en compte, en majeure partie car les données sur lesquelles sont entraînés les analyseurs représentent peu ou pas ces cas particuliers. Nous présentons, dans cet article, un nouveau corpus en dépendances pour le français, librement disponible, contenant un nombre substantiel de dépendances non-projectives. Ce corpus permettra d'étudier et de mieux prendre en compte les cas de non-projectivité dans l'analyse du français.

Abstract.
CDGFr, a Non-projective Dependency Corpus for French.
The non-projective cases, as a part of the dependency parsing of French, are often disregarded, mainly because the treebanks on which parsers are trained contain little or no non-projective dependencies. In this paper, we present a new freely available dependency treebank for French that includes a substantial number of non-projective dependencies. This corpus can be used to study and process non-projectivity more effectively within the context of French dependency parsing.

Mots-clés : Corpus français, annotation en dépendances, dépendances non-projectives.

Keywords: Treebank for French, dependency annotation, non-projective dependencies.

1 Introduction

Les développements actuels d'analyseurs syntaxiques en TALN reposent principalement sur l'utilisation de corpus arborés. Ces données se substituent à la notion plus traditionnelle de grammaire que l'on utilise maintenant principalement pour modéliser la syntaxe des langages artificiels comme les langages informatiques. De nombreuses raisons peuvent expliquer ce phénomène comme la difficulté de développer une grammaire pour une langue qui soit précise, robuste, à large couverture, qui puisse évoluer au cours du temps, en fonction du domaine, etc. De l'autre coté, l'approche *dirigée par les données* est simple à mettre en œuvre, rapide (si l'on ne tient pas compte de la difficulté de disposer des données d'apprentissage) et donne de bons résultats à la fois en termes de vitesse d'analyse et de robustesse.

Il semble important de noter que même si les analyseurs n'ont plus besoin de grammaire pour fonctionner, les corpus qu'ils utilisent lors de la phase d'apprentissage reposent souvent sur un modèle linguistique qui a permis de créer directement ce corpus ou bien de l'obtenir par des transformations depuis d'autres ressources qui elles ont un modèle linguistique. Suivant le point de vue de Dikovsky (2011), le lien entre les corpus et les modèles sur lesquels ils reposent et en tenant compte des transformations utilisées est essentiel à la compréhension des corpus. Par exemple, les corpus FTB (*French Treebank*) (Abeille *et al.*, 2003) et Sequoia (Candito & Seddah, 2012) permettent d'entraîner des analyseurs syntaxiques en dépendances après transformation en arbres de dépendances (Candito *et al.*, 2010). Le modèle en constituant de départ ne permet en général pas d'obtenir beaucoup de dépendances non-projectives (i.e. des dépendances qui peuvent croiser les autres dépendances en raison d'une discontinuité dans la langue comme la dépendance entre « en » et « directives » de la figure 1). Par conséquent, un analyseur basé sur ces corpus va peu produire ce type de structures même s'il est capable de traiter ces dépendances efficacement.

D'ailleurs, les corpus en dépendances librement accessibles pour le français et comportant un nombre conséquent de structures de dépendances non-projectives ne sont pas très nombreux. Le FTB convertit en dépendances ne contient pas de dépendances non-projectives. Le corpus Sequoia comprend 1,2 % de structures de dépendances non-projectives et, plus récent, le corpus UDT (*Universal Dependency Treebank*) (McDonald *et al.*, 2013) en comprend 12,4 % pour le français,

ce qui correspond à, respectivement, 0,2 % et 2,4 % de dépendances non-projectives [1] sur l'ensemble des dépendances de chacun de ces corpus. En conséquence, les cas de non-projectivité dans la langue française sont difficiles à traiter et sont donc peu ou pas pris en compte lors de l'analyse de ces données. Notons également qu'il existe d'autres corpus en dépendances pour le français avec lesquels nous ne nous comparerons pas directement puisque ces corpus proposent des représentations en dépendances différentes de celles que nous proposons et donc ciblent des usages différents. Par exemple, il existe une version en dépendances profondes du Sequoia (Candito *et al.*, 2014). Citons également le corpus Rhapsodie (Lacheret *et al.*, 2014) traitant du français parlé.

Pour ces raisons, nous avons entrepris le développement d'un corpus de structures de dépendances centré sur la notion de dépendances non-projectives comprenant un noyau ayant servi, dans un premier temps, à développer en parallèle une grammaire catégorielle pour le français et son corpus de développement puis, dans un deuxième temps, à ajouter des corpus supplémentaires. Et tandis que les corpus en dépendances existants pour le français rassemblent des phrases provenant majoritairement de textes journalistiques, nous avons choisi ici d'annoter des extraits de textes divers dont des périodiques mais également des textes littéraires variés en terme de style et de genre. Les structures de dépendances du corpus sont disponibles [2] sous la licence LGPL-LR (Lesser General Public Licence For Linguistic Resources) [3], les œuvres d'origines restant la propriété des auteurs ou de leurs ayants droit. Le corpus a été développé en suivant la méthode proposée dans (Dikovsky, 2011) basée sur la construction incrémentale d'une grammaire catégorielle de dépendance (CDG) et d'un corpus de développement. Nous avons utilisé pour cela l'environnement de développement des CDG proposé par (Béchet *et al.*, 2014). Nous pensons que ce corpus permettra de montrer l'intérêt d'étudier et de prendre en compte les structures non-projectives dans les langues naturelles en particulier pour le français.

Dans la suite de l'article, la structure générale du corpus et la provenance des textes sont abordées. Nous présentons ensuite la structure et le schéma d'annotation des unités lexicales (mots ou groupe de mots), de leurs classes grammaticales et de leurs traits puis la structure des types de dépendances présents dans les arbres de dépendances. La section révèle également la méthodologie employée lors de la création du corpus et de ses annotations et comporte une analyse statistique du contenu. Nous évoquons pour finir des travaux d'analyse en dépendances dirigés par les données présentant des résultats préalables sur le corpus CDGFr.

2 Corpus

2.1 Origines

Le corpus CDGFr comporte des phrases de trois origines différentes : un noyau de développement du modèle linguistique, des extraits de textes littéraires de la fin du XIXème siècle et du XXème siècle et des extraits de périodiques contemporains. La première partie du corpus, nommée **CDGFr-devel** est particulièrement importante. Elle est constituée de phrases souvent assez courtes provenant de multiples sources et représentant les différentes structures syntaxiques du français. Les autres sources (littérature et périodiques) ont permis de valider le modèle sur des extraits réels de textes.

Littérature

Les oeuvres littéraires ont été choisies en fonction de leur style assez différents les uns des autres.

- **Zola** : extrait du chapitre 1 de « *Germinal* » de E. Zola publié en 1885.
- **Céline** : extrait du chapitre 1 du « *Voyage au bout de la nuit* » de L.F. Céline publié en 1932.
- **Camus** : extrait du chapitre 1 de la première partie de « *L'étranger* » de A. Camus publié en 1942.
- **Le Clézio** : extrait de « *L'échappé* » de « *La ronde et autres faits divers* » de J.M.G. Le Clézio publié en 1982.

Dans Germinal, on trouve des phrases descriptives souvent longues avec des constructions apposées donnant lieu à des dépendances parfois très longues. Le texte est ponctué de dialogues avec leur syntaxe propre, enchâssés dans le texte. Le style de Céline tient plus du langage parlé parfois pas très grammatical (au sens académique du terme). Les phrases y sont plutôt courtes même s'il a été choisi de parfois les regrouper en une seule analyse comme par exemple la phrase suivante :

1. Formellement, une dépendance est non-projective s'il existe au moins un mot situé entre sa tête et son dépendant direct qui n'est pas dominé par la tête, i.e. qui ne dépend pas directement ou indirectement de la tête.
2. La version actuelle est consultable à l'adresse suivante : `http://pagesperso.lina.univ-nantes.fr/~bechet-d/CDGFr`
3. `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/lgpllr.html`

> Quand il fait très froid, non plus, il n'y a personne dans les rues ; c'est lui, même que je m'en souviens, qui m'avait dit à ce propos : « Les gens de Paris ont l'air toujours d'être occupés, mais en fait, ils se promènent du matin au soir ; la preuve, c'est que lorsqu'il ne fait pas bon à se promener, trop froid ou trop chaud, on ne les voit plus ; ils sont tous dedans à prendre des cafés-crème et des bocks. C'est ainsi ! Siècle de vitesse ! qu'ils disent. Où ça ? Grands changements ! qu'ils racontent. Comment ça ? Rien n'est changé en vérité. Ils continuent à s'admirer et c'est tout. Et ça n'est pas nouveau non plus. Des mots, et encore pas beaucoup, même parmi les mots, qui sont changés ! Deux ou trois par-ci, par-là, des petits ... »

Les phrases de Camus portent aussi sur un texte à la première personne mais dans un style beaucoup plus neutre. Finalement, le texte de Le Clézio est au présent et comporte peu de dialogues. Sa forme est moderne et neutre.

Ainsi, à travers ses quatre textes, nous avons un échantillon assez large de formes : description au passé, au présent ; dialogue ordinaire, dialogue argotique, narration à la première personne. Nous pensons qu'ils représentent une partie importante des structures syntaxiques utilisées dans la littérature française.

Périodiques

Les deux extraits choisis comportent un texte de journalisme scientifique paru dans le journal Le Monde et un texte de journalisme historique paru dans la revue mensuelle de la ville de Nantes distribuée gratuitement dans l'agglomération nantaise et disponible sur Internet [4]. Ce ne sont donc pas des dépêches mais plutôt des textes déjà bien construits exprimant des faits réels actuels ou passés.

— **Nantes Passion** : extrait de l'article « Il y a 70 ans, le procès des « 42 » » de L. Abed-Denesle pour le magazine municipale mensuel Nantes Passion (numéro 230, janvier 2013)
— **Univers** : extrait de l'article « L'enfance de l'univers dévoilée » de J.L. Puget pour Le Monde (22 mars 2013)

2.2 Caractéristiques

Les caractéristiques des différents corpus présentés précédemment sont exposées dans la table 1, ainsi que les caractéristiques du corpus total correspondant à l'union de ces corpus. Il est intéressant d'observer les différences effectives entre les données provenant de ces différentes sources. Les corpus dont les phrases proviennent d'oeuvres littéraires rassemblent les phrases les plus longues. On remarquera en particulier le corpus Zola, dont la longueur des phrases varie fortement et qui comprennent en moyenne 30 mots. Par ailleurs, le corpus CGDFr-devel concentre le plus grand nombre de phrases mais celles-ci sont relativement courtes en moyenne par rapport à celles des autres sources.

Corpus	**Phrases**			**Mots (ponctuations comprises)**	
	total	**longueur**		**total**	**formes fléchies diff.**
		moyenne	**écart-type**		
CDGFr-devel	1 995	11,1	7,0	22 195	3 526
Zola	100	30,0	25,0	3 004	1 029
Céline	91	19,8	25,1	1 801	603
Camus	319	16,5	12,4	5 253	1 268
Le Clézio	528	18,7	10,2	9 894	1 730
Nantes Passion	42	22,8	13,4	957	436
Univers	64	25,3	13,5	1 619	643
Total	3 139	14,3	11,5	44 723	5 892

TABLE 1 – Caractéristiques des corpus

4. http://www.nantes.fr/nantes-passion

3 Annotations

3.1 Schéma d'annotation morpho-syntaxique (en classes grammaticales)

La grammaire catégorielle servant de modèle linguistique regroupe les unités lexicales en classes ayant des propriétés syntaxiques proches. Ces classes sont elles-mêmes regroupées en classes grammaticales générales listées dans la table 2. Une classe se distingue d'une autre classe, dans la même classe générale, par sa fonction, sa forme ou le type de ses arguments. Par exemple, pour la classe générale *N* des noms, *N(Lex=proper)* correspond aux noms propres, *N(Lex=common)* aux noms communs, *N(Lex=time)* aux noms des dates ou du temps ainsi qu'aux adverbes pouvant être une réponse à une question comme « quand venez-vous ? » : « aujourd'hui », « bientôt », « jamais », « heure », « avril », etc. Dans la classe générale *Vi* des verbes intransitifs, on trouvera *Vi(F=fin)* pour les formes finies (conjuguées), *Vi(F=inf)* pour les infinitifs, *Vi(F=pz,T=past)* pour les participes passés et *Vi(F=pz,T=pres)* pour les participes présents. Les verbes avec deux arguments de *V2t* sont regroupés suivant le type des deux compléments : *V2t(F=fin,C1=a,C2=d)* correspond aux verbes avec un complément d'objet direct (accusatif) et un complément second introduit par la préposition *à* (datif), etc. En plus de la classe grammaticale de chaque unité de la phrase, le corpus précise sa forme dans le lexique (par exemple, sans majuscule pour le premier mot d'une phrase), sa forme normalisée (infinitif pour un verbe ou forme au masculin singulier) et la liste des traits (genre, nombre, mode, temps, personne). Les deux derniers attributs indiquent la provenance de la forme dans le lexique (principalement basée sur le Lefff (Sagot, 2010)) et la manière avec laquelle la forme a été associée à la classe grammaticale (basée principalement sur les informations disponibles sur les formes et leurs arguments dans le Lefff). Quelques classes supplémentaires ont été introduites pour les termes qui ne sont pas reconnus par le lexique : les formes pouvant correspondre à des noms propres (avec des majuscules), à des nombres (composés de chiffres), des nombres complexes en toutes lettres et sinon des termes inconnus de classe *UT(Lex=V|N|Adj|Adv)* traitée comme une agrégation des propriétés syntaxiques des verbes, des noms, des adjectifs et des adverbes. Des exemples de phrases comportant cette classe se trouvent dans le corpus *CDGFr-devel* : « Adam va y xxx bientôt » où « xxx » porte cette classe. Dans ce cas, la pseudo-forme associée dans le lexique est *$UnknownTerm*. La classe des termes inconnus est utilisée dans le corpus pour signifier qu'un lexique ne pourra jamais être complet (terme ancien, trop récent, peu utilisé, utilisé localement, faute d'orthographe, etc) et qu'il peut être intéressant de les traiter de cette manière.

Adjectifs	**Adj**	Prépositions	**PP**	Ponctuations	**Dash**
Adverbes	**Adv**	Verbes auxiliaires	**Vaux**		**Parentheses**
Collocations	**Colloc**	Verbes copules	**Vcopul**		**QuestMark**
Conjonctions	**Conj**	Verbes intransitifs	**Vi**		**Quotes**
Déterminants	**Det**	Verbes substituts	**Vlight**		**SemiColon**
Interjections	**Expletives**	Verbes transitifs	**Vt**		**Chevrons**
Noms	**N**	Verbes ditransitifs	**V2t**		**Colon**
Nombres	**Num**	Unités inconnues	**UT**		**FullStop**
Partitifs	**Part**				**EmphatMark**
Pronoms	**PN**				**Comma**

TABLE 2 – Liste des classes grammaticales générales de la grammaire catégorielle de dépendances du français

3.2 Schéma d'annotation en dépendances

Les structures de dépendances du corpus sont techniquement des graphes acycliques orientés (*DAG*) superposant deux arbres dont les dépendances sont des arcs de différents types. Ceux-ci sont de 3 types : les dépendances projectives (*projective*), les dépendances discontinues (*discontinuous*) et les ancres (*anchor*). La majeure partie des dépendances sont des dépendances projectives (les arcs pleins en noir dans la figure 1) tandis que les dépendances discontinues (les arcs en pointillé) sont des dépendances qui peuvent croiser les autres dépendances dans la structure. Les ancres sont, en outre, des pseudo-dépendances de deux sortes : d'une part, la plupart des pseudo-dépendances arrivant sur une ponctuation sont des ancres (les arcs dont l'étiquette est préfixée par « @ »), d'autre part, chaque mot recevant une dépendance discontinue reçoit également une ancre (les arcs dont l'étiquette est préfixée par « # ») de même rôle syntaxique, e.g. le mot « en » dans la figure 1 possède donc deux têtes, une vraie tête discontinue « données » et une pseudo-tête projective « sont ». De la sorte, l'union des dépendances projectives et des dépendances discontinues, pour une phrase donnée, forme

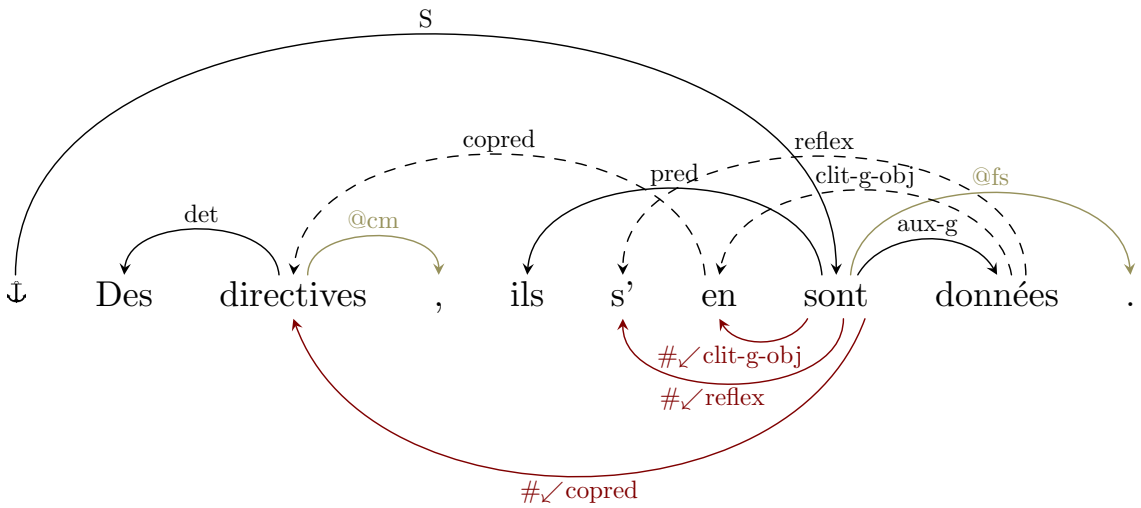

FIGURE 1 – Structure de dépendances du corpus CDGFr-devel pour la phrase « Des directives, ils s'en sont données »

un arbre éventuellement non-projectif [5] tandis que l'union des dépendances projectives et des ancres forme toujours un arbre projectif.

De plus, les dépendances se distinguent également par le rôle syntaxique (nom de la dépendance) qui leur est associé. La terminologie des noms de dépendances provient de la théorie sens-texte de Mel'čuk & Pertsov (1986); Mel'čuk (1988). Une présentation succincte des dépendances utilisées ici se trouve dans (Dikovsky, 2011). La présentation exhaustive est disponible dans la documentation associée au corpus. On y trouve notamment les informations sur le site d'ancrage des mots sur lesquels une dépendance discontinue arrive. Les noms des dépendances (figurants en minuscules dans le corpus) sont regroupés en groupes (figurants en majuscules). Les dépendances d'un même groupe se distinguent par le cas de l'argument (nominatif, accusatif, datif, etc), son type (nom, adjectif, complément verbal), ou des propriétés du contexte. Par exemple, dans le groupe *OBJ* des arguments des verbes (autre que le sujet), on trouve *a-obj* pour le complément d'objet direct, *d-obj* pour un complément indirect au datif (introduit par « à »), etc. Le groupe *COPUL* des copules comporte la dépendance *a_copul* pour les adjectifs attributs (comme « il est jeune »), *n_copul* pour les noms attributs (comme « il s'appelle Pierre » ou *c_copul* pour des compléments introduits par une préposition (comme « il était au commencement »). Dans le groupe *OBJ*, la dépendance *a-obj-g* lie un verbe et son complément d'objet direct mais indique aussi que le complément doit posséder une dépendance vers un clitique placé avant le verbe comme la dépendance entre « avait » et « besoin » dans la phrase « il en avait besoin ».

AGENT	(0,21 % / 0,02 %)	COPUL	(2,67 % / 0,04 %)	PRED	(9,94 % / 0,00 %)
AGGR	(1,46 % / 0,05 %)	CORREL	(0,04 % / 0,00 %)	PREFIXA	(0,02 % / 0,00 %)
APPOS	(0,92 % / 0,15 %)	DEICT	(0,03 % / 0,00 %)	PREPOS	(8,29 % / 0,00 %)
APPROX	(0,06 % / 0,00 %)	DET	(10,64 % / 0,00 %)	PUNCT	(12,74 % / 0,00 %)
ATTR	(2,95 % / 0,02 %)	EMPHAT	(0,71 % / 0,00 %)	QUANT	(0,92 % / 0,08 %)
AUX	(2,35 % / 0,00 %)	EXPLET	(0,11 % / 0,02 %)	QUANTIF	(0,29 % / 0,00 %)
CIRC	(6,38 % / 0,00 %)	GER	(0,15 % / 0,00 %)	REFLEX	(0,97 % / 0,50 %)
CLAUS	(2,83 % / 0,00 %)	INF	(2,96 % / 0,00 %)	REL	(1,03 % / 0,26 %)
CLIT	(1,82 % / 0,85 %)	INTERROG	(0,06 % / 0,00 %)	RESTRICT	(1,14 % / 0,02 %)
COMPAR	(0,44 % / 0,00 %)	JUNC	(3,49 % / 0,00 %)	SELECT	(0,12 % / 0,02 %)
CONJ	(0,47 % / 0,00 %)	MODIF	(3,14 % / 0,08 %)	SENT	(7,53 % / 0,00 %)
COORDV	(1,36 % / 0,00 %)	NEG	(1,51 % / 1,15 %)	VOCATIVE	(0,19 % / 0,00 %)
COPRED	(0,08 % / 0,33 %)	OBJ	(6,26 % / 0,12 %)		

TABLE 3 – Liste des groupes de dépendances de la grammaire catégorielle de dépendances du français et pourcentage de dépendances (projectives / discontinues) parmi les dépendances associées aux groupes dans l'ensemble du corpus

5. Une structure de dépendances ne peut être non-projective que si elle contient au moins une dépendance discontinue.

3.3 Méthodologie d'annotation

Le corpus comporte deux parties bien distinctes. Le corpus *CDGFr-devel* a été développé en même temps que le modèle grammatical associé qui se présente sous la forme d'une grammaire catégorielle de dépendances. Le développement a consisté, à partir d'une grammaire catégorielle noyau du français, en l'ajout progressif au corpus de phrases introduisant de nouvelles structures syntaxiques et de la modification correspondante de la grammaire catégorielle.

La seconde partie comportant les corpus sur la littérature et les périodiques a été créée en utilisant les outils d'analyse syntaxiques dérivés du modèle défini par la première phase, la documentation dérivée portant sur les dépendances syntaxiques et les exemples d'analyses fournis par le corpus de développement. Ce processus d'annotation semi-automatique pouvait également être renforcé, suivant le choix de l'annotateur, par une étape de pré-annotation manuelle (combinant segmentation et étiquetage grammatical) permettant de réduire le temps de la phase d'analyse avec la grammaire. Après analyse, le processus a requis une étape de validation manuelle des structures de dépendances et des annotations morphosyntaxiques. En dernier ressort, les analyses sélectionnées ont été vérifiées par l'équipe qui a créé le modèle. Les problèmes éventuels ont été détectés pour que la grammaire initiale et les corpus déjà créés puissent être modifiés (par exemple si un mot n'appartenait pas à une classe grammaticale ou plus rarement, si une structure syntaxique a été oubliée ou n'était pas utilisée de manière cohérente).

En outre, le corpus a d'ores et déjà été exploité pour l'élaboration d'outils de pré-annotation automatique supervisé (Lacroix *et al.*, 2014) dans le but d'améliorer la rapidité et le confort d'annotation pour le développement futur du corpus.

3.4 Statistiques

La table 4 rassemble les statistiques sur les structures de dépendances et les dépendances elles-mêmes pour les différents corpus annotés. En particulier, sont présentés les taux de dépendances discontinues et de structures de dépendances non-projectives. Les statistiques sur les dépendances ne prennent pas en compte les ancres (i.e. les ancres liées aux ponctuations et les ancres qui vont de pair avec les dépendances discontinues).

Corpus	**Structures de dépendances**			**Dépendances**		
	total	**non-projectives** (%)		**total**	**discontinues** (%)	
CDGFr-devel	1 995	864	(43,3 %)	19 340	1 095	(5,7 %)
Zola	100	41	(41,0 %)	2 646	65	(2,7 %)
Céline	91	36	(39,6 %)	1 560	68	(4,4 %)
Camus	319	158	(49,5 %)	4 707	216	(4,6 %)
Le Clézio	528	151	(28,6 %)	8 791	187	(2,1 %)
Nantes Passion	42	9	(21,4 %)	868	9	(1,0 %)
Univers	64	21	(32,8 %)	1 460	21	(1,4 %)
Total	3 139	1 280	(40,8 %)	39 372	1 661	(4,2 %)

TABLE 4 – Statistiques sur les dépendances et les structures

4 Analyse en dépendances

À partir du schéma d'annotation en dépendances présenté en section 3.2 il est possible d'extraire, depuis le corpus CDGFr, des arbres standards possiblement non-projectifs ou uniquement projectifs (en conservant les ancres à la place des dépendances discontinues). Ces arbres sont donc adaptables aux systèmes standards d'analyse en dépendances dirigés par les données tels que les analyseurs par transition ou les analyseurs basés sur les graphes. Inversement, tout corpus standard non-projectif peut également être adapté à la représentation en dépendances des CDG par l'emploi d'une méthode automatique de projectivisation des dépendances non-projectives.

Des résultats d'analyses en dépendances, effectuées sur le corpus CDGFr à l'aide de différents algorithmes d'analyse par transition, peuvent être trouvés dans (Lacroix & Béchet, 2014), ainsi que la présentation d'un algorithme spécialement adapté à la représentation en dépendances exploitées dans nos corpus.

5 Conclusion

Nous avons présenté un nouveau corpus arboré en dépendances pour le français contenant un nombre substantiel de dépendances non-projectives. Les différents corpus formant le corpus intégral sont visualisables à l'adresse `http://pagesperso.lina.univ-nantes.fr/~bechet-d/CDGFr` et téléchargeable au format XML. Cet ensemble de phrases annotées en dépendances servira à l'étude, notamment, des phénomènes non-projectifs (discontinuités) dans la langue française et ainsi amènera à prendre en compte et à mieux traiter ces cas particuliers dans le cadre de l'analyse en dépendances globale de données sur le français. Certaines dénominations comme le nom de la dépendance *pred* (predicative) proviennent de (Mel'čuk & Pertsov, 1986; Mel'čuk, 1988). Par respect pour le travail d'Alexandre Dikovsky présenté dans (Dikovsky, 2011), nous n'avons pas cherché à les renommer dans cette version du corpus (version 3.4).

Remerciements

La grammaire de CDG sur laquelle repose le CDGFr a été développée par Alexandre Dikovsky qui est aussi l'auteur du rapport technique sur les dépendances que l'on peut consulter sur le site de téléchargement du corpus. Le lexique associé a été développé par Denis Béchet, Alexandre Dikovsky et Ramadan Alfared. Nous voudrions remercier particulièrement Danièle Bauquier qui a analysé une partie conséquente du corpus. Nous dédions cette article à Alexandre Dikovsky qui nous a malheureusement quitté en 2014. Sans lui ce projet n'aurait jamais existé.

Références

ABEILLE A., CLEMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In A. ABEILLÉ, Ed., *Treebanks*, volume 20 of *Text, Speech and Language Technology*, p. 165–187. Springer Netherlands.

BÉCHET D., DIKOVSKY A. & LACROIX O. (2014). "CDG Lab" : an Integrated Environment for Categorial Dependency Grammar and Dependency Treebank Development. In *Computational Dependency Theory*, volume 258 of *Frontiers in Artificial Intelligence and Applications*, p. 153–169. IOS Press.

CANDITO M., CRABBÉ B. & DENIS P. (2010). Statistical French Dependency Parsing : Treebank Conversion and First Results. In *Proceedings of the Language Resources and Evaluation Conference*, LREC 2010, Valletta, Malta.

CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & DE LA CLERGERIE É. (2014). Deep Syntax Annotation of the Sequoia French Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland.

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

DIKOVSKY A. (2011). Categorial Dependency Grammars : from Theory to Large Scale Grammars. In *Proceedings of the International Conference on Dependency Linguistics*, DEPLING 2011, Barcelona, Spain.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P., TCHOBANOV A. *et al.* (2014). Rhapsodie : a Prosodic-Syntactic Treebank for Spoken French. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland.

LACROIX O. & BÉCHET D. (2014). A three-step transition-based system for non-projective dependency parsing. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING 2014, Dublin, Irlande.

LACROIX O., BÉCHET D. & BOUDIN F. (2014). Label pre-annotation for building non-projective dependency treebanks for french. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing 2014, Kathmandu, Népal.

MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TÄCKSTRÖM O., BEDINI C., BERTOMEU CASTELLÓ N. & LEE J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, Sofia, Bulgaria.

MEL'ČUK I. (1988). *Dependency syntax : Theory and Practice*. State University of New York Press.

MEL'ČUK I. A. & PERTSOV N. V. (1986). *Surface syntax of English : A formal model within the Meaning-Text framework*. John Benjamins Publishing Company.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, LREC'10, Valletta, Malte.