Traitement automatique des langues

Varia

sous la direction de
Cécile Fabre
Emmanuel Morin
Sophie Rosset
Pascale Sébillot

Vol. 62 - n°1 / 2021

ATALA

Varia

Cécile Fabre, Emmanuel Morin, Sophie Rosset, Pascale Sébillot
Préface

Imen Akermi, Johannes Heinecke, Frédéric Herledan
Génération automatique de texte en langage naturel pour les systèmes de questions-réponses

Guy Perrier
Étude des dépendances syntaxiques non projectives en français

Denis Maurel
Notes de lecture

Sylvain Pogodalla
Résumés de thèses et HDR

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
62

n°1
2021

Varia

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 62 – n°1 / 2021

VARIA

Table des matières

Préface
Cécile Fabre, Emmanuel Morin, Sophie Rosset, Pascale Sébillot 7

Génération automatique de texte en langage naturel pour les systèmes de questions-réponses
Imen Akermi, Johannes Heinecke, Frédéric Herledan 13

Étude des dépendances syntaxiques non projectives en français
Guy Perrier 39

Notes de lecture
Denis Maurel 65

Résumés de thèses et HDR
Sylvain Pogodalla 75

Préface

Les préfaces des numéros non thématiques de la revue *TAL* permettent de faire chaque année le point sur la vie de la revue et de commenter les statistiques que nous présentons traditionnellement pour les numéros des trois dernières années.

Différentes actions visant l'amélioration de la visibilité de la revue ont été entreprises ou développées au cours de l'année 2021. Les pages de la revue TAL sur le site de l'ATALA ont ainsi été traduites intégralement en anglais, et l'accès aux anciens articles, mis à mal par un changement des URL lors de la refonte du site en 2019, a été consolidé. Concernant l'indexation, des métadonnées manquantes ont été complétées et un convertisseur vers des formats attendus par des sites de référencement est en cours de mise au point ; par ailleurs une indexation *via* ACL Ontology et un dépôt automatisé des articles sur HAL sont à l'étude. Nous remercions Gaël Guibon, Damien Nouvel et Philippe Muller pour leurs contributions respectives sur ces aspects.

La vie du comité a été marquée, lors du renouvellement annuel des membres, par l'accueil de quatre femmes, qui nous a permis d'atteindre pour la première fois une parité en genre. Parmi les autres sujets dont nous pouvons nous réjouir, la régularité de la publication de la revue a été maintenue, en nous fondant sur notre calendrier prévisionnel qui nous permet de caler les différents numéros d'un même volume et de tenir de façon régulière les réunions du comité de rédaction. Rappelons que l'une des caractéristiques de notre revue, à laquelle nous sommes foncièrement attachés, est la tenue des réunions du comité de rédaction – en visioconférence – au cours desquelles, entre autres, nous décidons collégialement, à l'appui des relectures reçues, de l'acceptation ou du rejet des articles soumis.

L'attractivité de la revue a été, au cours de cette année, le sujet de plusieurs discussions lors de ces réunions du comité de rédaction. Dès 2020, la diminution du nombre de soumissions aux derniers numéros *Varia* avait soulevé la question de la pertinence du maintien d'un tel numéro annuel. La décision avait été prise de se donner le temps de vérifier la persistance de cette tendance, que les chiffres du numéro 62:1 semblent confirmer (sept soumissions ; *cf.* tableau 1 *infra*). Les discussions de 2021 ont plus spécifiquement concerné les numéros thématiques, dont l'attractivité varie fortement d'un numéro à l'autre (de quatre à vingt-trois soumissions si l'on considère les trois

derniers volumes). La décision a été prise de proposer en 2021 deux numéros à thématiques assez larges, afin de tester si les appels pouvaient attirer un public plus large. Il semblerait que ce ne soit pas le cas, puisque le numéro portant sur les nouvelles applications du TAL n'a abouti qu'à deux soumissions et celui sur la diversité linguistique à trois. Le comité de rédaction de la revue doit donc continuer à se pencher sur cette question de l'attractivité, tout en gardant à l'esprit que les chiffres de ces deux dernières années peuvent être liés à la crise sanitaire, aux confinements successifs et à la tenue des conférences à distance.

Intitulé	**Vol.**	**N°**	**Année**	**Soumis**	**Acceptés**	**% acceptés**
Varia	59	1	2018	10	3	30,0 %
Apprentissage profond pour le TAL	59	2	2018	8	3	37,5 %
Trait. auto. des langues peu dotées	59	3	2018	23	4	17,4 %
Sous-total	**59**		**2018**	**41**	**10**	**24,4 %**
Varia	60	1	2019	8	1	12,5 %
Corpus annotés	60	2	2019	6	3	50,0 %
TAL et humanités numériques	60	3	2019	13	5	38,5 %
Sous-total	**60**		**2019**	**27**	**9**	**33,3 %**
Varia	61	1	2020	8	1	12,5 %
TAL et santé	61	2	2020	4	3	75,0 %
Dialogue et systèmes de dialogue	61	3	2020	5	3	60,0 %
Sous-total	**61**		**2020**	**17**	**7**	**41,2 %**
Varia	62	1	2021	7	2	28,6 %
Total	**Dix derniers n^{os}**			**92**	**28**	**30,4 %**

Tableau 1. *Taux de sélection aux appels de la revue TAL sur les dix derniers numéros de la période 2018-2021*

Les statistiques que nous présentons dans le tableau 1 considèrent les dix derniers numéros sur les trois dernières années, en l'occurrence donc du début de 2018 jusqu'à ce numéro *Varia* de 2021 inclus. Ce tableau donne les taux de sélection par numéro et par volume. La ligne du total synthétise ces chiffres sur l'ensemble des dix numéros considérés. Le taux de sélection sur l'ensemble de ces numéros s'élève à 30,4 % en moyenne, avec d'assez grandes variations selon les numéros (de 12,5 % pour les deux *Varia* de 2019 et 2020, dont les chiffres sont rigoureusement identiques, à 75 % pour le numéro dédié à la thématique TAL et santé).

Rappelons qu'un numéro ne peut pas excéder cinq articles, pour des raisons liées au coût du processus d'édition. Ce seuil a été atteint une fois seulement dans la période, pour le numéro consacré aux humanités numériques.

Intitulé	**Vol.**	**N°**	**Année**	**% 1er auteur hors France**	**% en anglais**
Varia	59	1	2018	0,0 %	0,0 %
Apprentissage profond pour le TAL	59	2	2018	33,3 %	66,6 %
Trait. auto. des langues peu dotées	59	3	2018	25,0 %	0,0 %
Pourcentages par volume	**59**		**2018**	**20,0 %**	**20,0 %**
Varia	60	1	2019	0,0 %	0,0 %
Corpus annotés	60	2	2019	0,0 %	0,0 %
TAL et humanités numériques	60	3	2019	40,0 %	40,0 %
Pourcentages par volume	**60**		**2019**	**22,2 %**	**22,2 %**
Varia	61	1	2020	0,0%	0,0%
TAL et santé	61	2	2020	33,3 %	33,3 %
Dialogue et systèmes de dialogue	61	3	2020	66,7 %	66,7 %
Pourcentages par volume	**61**		**2020**	**42,9 %**	**42,9 %**
Varia	62	1	2021	0,0%	0,0%
Pourcentages totaux	**Dix derniers nos**			**25,0 %**	**25,0 %**

Tableau 2. *Proportion des articles publiés d'un premier auteur d'un laboratoire hors de France et proportion des articles publiés rédigés en anglais sur les dix derniers numéros de la période 2018-2021. Attention, les pourcentages totaux ne sont pas de simples moyennes des chiffres donnés plus haut, car les dénominateurs changent.*

Le comité de rédaction de la revue est très attaché à sélectionner les articles sur le seul critère de leur qualité, indépendamment du nombre d'articles soumis, et n'hésite pas, comme dans le cas des trois derniers *Varia*, à préserver cette exigence malgré un nombre de soumissions limité.

Les statistiques que nous donnons dans le tableau 2 sur l'origine des articles considèrent le pays du laboratoire du premier auteur, hors de France ou pas, ainsi que la langue de la soumission, le français ou l'anglais, l'anglais n'étant possible que si l'un des coauteurs n'est pas francophone. Les chiffres sont fournis pour la même période de temps que le tableau 1. Globalement, un quart des premiers auteurs sont des chercheurs de laboratoires hors de France, et un quart des articles sont en anglais. Néanmoins, ces chiffres cachent des disparités fortes selon les numéros, avec, pour chacune des trois années considérées, un numéro qui se singularise et attire plus nettement les chercheurs étrangers (apprentissage profond, TAL et humanités numériques, dialogue et systèmes de dialogue). Contrairement à des années plus anciennes (2016 ou 2017 par exemple), les numéros *Varia* n'attirent ou ne retiennent pas d'articles ayant ce caractère international. Ce dernier point soulève deux remarques fortement liées : d'une part, il pourrait être intéressant d'introduire dans nos statistiques le décompte des soumissions à caractère international – et non juste celles retenues pour

publication –, afin de mieux refléter l'attractivité hors de nos frontières de la revue TAL, d'autre part, à l'aune de ces nouveaux nombres spécifiquement pour les numéros *Varia*, il serait possible d'établir si ces numéros non thématiques répondent ou non à une attente à l'international.

Le présent numéro contient les articles retenus suite à l'appel non thématique lancé en juillet 2020. Cet appel portait comme d'habitude sur tous les aspects du traitement automatique des langues. Sept articles ont été soumis. À l'issue du processus de sélection, deux articles ont été acceptés pour publication :

– *Génération automatique de texte en langage naturel pour les systèmes de questions-réponses*, Imen Akermi, Johannes Heinecke, Frédéric Herledan (Orange Innovation) : cet article propose une méthode non supervisée permettant de générer des réponses concises mais complètes à des questions en langue naturelle ;

– *Étude des dépendances syntaxiques non projectives en français*, Guy Perrier (LORIA, Université de Lorraine) : en se fondant sur deux corpus annotés selon deux formats distincts, cet article présente une étude des constructions du français dont la structure syntaxique fait l'objet d'une représentation en dépendances non projective (caractère local du croisement de dépendances, etc.).

On trouvera à la suite de ces articles des notes de lecture rassemblées par Denis Maurel. Nous encourageons nos lecteurs à utiliser ce moyen pour faire profiter de leurs lectures la communauté. Suit une liste de résumés de thèses ou d'habilitations à diriger les recherches en TAL préparée par Sylvain Pogodalla. Merci à Denis et Sylvain pour leur travail de veille et de collecte.

Merci aux membres du comité de rédaction de la revue qui ont participé aux différentes étapes d'élaboration de ce numéro, et en particulier à ceux qui ont pris en charge des relectures (voir la composition du comité sur le site de la revue : https://www.atala.org/content/comité-de-rédaction-0).

Merci aux relecteurs spécifiques de ce numéro :

– Delphine Battistelli (MoDyCo, Université Paris-Nanterre) ;
– Denis Béchet (LS2N, Université de Nantes) ;
– Anne-Laure Ligozat (LISN, ENSIIE) ;
– Thierry Poibeau (CNRS, ENS/PSL et Université Sorbonne Nouvelle) ;
– Jean-Philippe Prost (LPL, Aix-Marseille Université) ;
– Agata Savary (LIFAT, Université de Tours) ;
– Laure Soulier (LIP6, Sorbonne Université).

La revue TAL reçoit un soutien financier de l'Institut des sciences humaines et sociales (INSHS) du CNRS et de la délégation générale à la langue française et aux langues de France (DGLFLF). Nous adressons nos remerciements à ces organismes.

Cécile Fabre
CLLE, Université Toulouse 2
cecile.fabre@univ-tlse2.fr

Emmanuel Morin
LS2N, Université de Nantes
emmanuel.morin@univ-nantes.fr

Sophie Rosset
LISN, CNRS
sophie.rosset@lisn.fr

Pascale Sébillot
IRISA, INSA Rennes
pascale.sebillot@irisa.fr

Génération automatique de texte en langage naturel pour les systèmes de questions-réponses

Imen Akermi* — **Johannes Heinecke*** — **Frédéric Herledan***

* *Orange Innovation, 2 av. Pierre Marzin, 22307 Lannion, France*
imenakermi@yahoo.fr,
{johannes.heinecke,frederic.herledan}@orange.com

RÉSUMÉ. *Cet article traite de la génération du langage naturel dans le contexte des systèmes de questions-réponses. Les différents travaux portant sur ces systèmes se sont focalisés sur la génération d'une réponse courte ou d'un paragraphe contenant la réponse, à partir de données structurées ou de pages Web. La longueur de ces réponses n'est généralement pas appropriée du fait que les réponses peuvent être perçues comme trop brèves ou trop longues pour être lues à haute voix par un assistant intelligent. Dans ce travail, nous présentons une approche non supervisée de génération de réponses concises qui ne nécessite pas de données annotées. Testée sur des corpus de données en anglais et en français, l'approche proposée montre des résultats très prometteurs.*

MOTS-CLÉS : *systèmes de questions-réponses, génération du langage naturel, analyse en dépendances.*

TITLE. *Compact answer generation with a Transformer based approach.*

ABSTRACT. *This paper presents an unsupervised approach for natural language generation within the framework of question-answering systems. This approach addresses the issue of generating answers that are usually too short or too long without having to resort to annotated data. This approach shows promising results for English and for French.*

KEYWORDS: *question answering systems, natural language generation, dependency analysis.*

1. Introduction

Les systèmes de questions-réponses (SQR) analysent et traitent les questions des utilisateurs afin de leur fournir des réponses pertinentes (Hirschman et Gaizauskas, 2001). L'intérêt pour les SQR s'est accru avec la popularité récente des assistants intelligents. Ces derniers permettent aux utilisateurs de poser des questions en langage naturel, en utilisant leur propre terminologie, et d'avoir directement des réponses sans avoir à parcourir une longue liste de documents pour trouver les réponses appropriées. Les SQR sont ainsi devenus un élément central des échanges « humain-machine ».

La plupart des travaux de recherche existants se focalisent sur le traitement et l'interprétation de la question. Ils accordent souvent peu d'importance à la représentation de la réponse donnée en sortie. Généralement, la réponse est soit représentée par un ensemble de termes courts répondant exactement à la question, soit par un passage dans un document qui contient la réponse exacte mais qui peut aussi intégrer d'autres informations inutiles ne relevant pas du contexte de la question posée.

Prenons l'exemple de la question *Qui vivait au Costa Rica avant les Espagnols ?*, avec les SQR actuels, deux structures de réponses sont généralement renvoyées :

Amérindiens

À l'époque précolombienne, les Amérindiens de l'actuel Costa Rica faisaient partie d'un complexe culturel connu sous le nom de « zone intermédiaire », entre les régions culturelles mésoaméricaine et andine.

La première réponse pourra être perçue par les utilisateurs comme trop brève et ne rappelant pas le contexte de la question. La deuxième pourra être perçue comme trop longue et nécessitera à l'utilisateur une lecture attentive pour deviner la réponse à sa question au milieu d'informations non pertinentes.

Pour répondre aux attentes des utilisateurs, il conviendrait de produire une réponse synthétique qui soit concise, c'est-à-dire ne contenant pas d'autres informations que la stricte réponse à la question et qui soit aussi complète, c'est-à-dire rappelant le contexte de la question posée. Pour cette tâche, une approche simple pourrait consister à utiliser des règles prédéfinies pour générer les structures possibles de réponses (Reiter et Dale, 1997). Cependant, de telles approches ne sont pas généralisables et échouent à capturer les spécificités de la réponse. D'autre part, les approches par apprentissage supervisé qui se basent sur des architectures neuronales nécessitent de larges corpus de données qui feront correspondre une question à une réponse complète et concise. Elles sont très performantes pour générer des réponses pour des données du domaine sur lequel elles ont été entraînées mais elles ont souvent des limites pour généraliser sur d'autres domaines.

L'approche de génération de réponses que nous proposons est non supervisée et ne nécessite donc pas de corpus d'apprentissage. Elle peut facilement être adaptée à n'importe quelle langue. La performance de cette approche est attestée par des expérimentations et une évaluation humaine sur des données de test en anglais et en

français. Les données ont été acquises par écrit. Nous avons pu vérifier que les derniers systèmes de transcription de la parole sont capables de sortir du texte prenant en compte certaines spécificités de l'écrit comme les majuscules pour les entités. Nous avons donc estimé que ces données étaient aussi pertinentes pour l'usage oral que nous envisagions. En effet, bien que l'expression orale diffère de celle de l'écrit, le prototype (Rojas Barahona *et al.*, 2019) nous a montré que cette différence est moindre pour un usage de questions-réponses. Ces données portaient exclusivement sur la connaissance générale. Hormis pour des constructions syntaxiques très spécifiques à certains métiers, nous ne voyons pas d'obstacle à appliquer notre méthode aux données de domaines de spécialité. Elle ouvre également des pistes très prometteuses, d'une part pour générer des réponses synthétiques dans un contexte de dialogue et, d'autre part, pour créer automatiquement un corpus d'entraînement ou de test afin de creuser ultérieurement des approches supervisées. Les principales contributions de cet article peuvent être résumées comme suit :

– une approche de génération de réponses synthétiques en langage naturel ;

– une approche de construction automatique de corpus de questions-réponses pour creuser ultérieurement la capacité des systèmes supervisés à générer des réponses synthétiques

L'article est organisé comme suit. Nous présentons dans la section 2 une revue de la littérature sur les approches de génération automatique de texte de l'analyse en dépendances dans le contexte des systèmes de questions-réponses. La section 3 détaille l'approche de génération proposée. Nous décrivons dans la section 4 les expériences menées et nous introduisons dans la section 5 une approche de construction automatique de corpus de questions-réponses. Nous concluons dans la section 6 avec un résumé des approches proposées dans cet article ainsi que des réflexions pour les travaux futurs.

2. État de l'art

2.1. *Génération automatique de texte*

La génération automatique de texte (GAT) est considérée comme un sous-domaine de l'intelligence artificielle et de la linguistique computationnelle. Elle s'intéresse à la construction de systèmes capables de produire des textes en langage naturel qui soient compréhensibles, et ceci à partir d'informations extraites de textes, de données structurées ou de données visuelles telles que les images ou les vidéos (Reiter et Dale, 1997). Elle trouve une application particulière dans les SQR.

De nos jours, la grande quantité d'informations disponibles rend la recherche d'information complexe et chronophage. En renvoyant directement la réponse exacte à une question posée en langage naturel, les SQR évitent à l'utilisateur de devoir filtrer lui-même les informations renvoyées. Les SQR couvrent principalement trois tâches : l'analyse de la question, la recherche d'information et l'extraction de la ré-

ponse (Lopez *et al.*, 2011). Dans les travaux existants, ces tâches sont abordées de différentes manières, en fonction des bases de connaissances utilisées, des types de questions traitées (Iida *et al.*, 2019 ; Zayaraz *et al.*, 2015) et de la façon avec laquelle la réponse est présentée. Dans la littérature, nous distinguons généralement deux formes de représentation. La réponse peut prendre la forme d'un paragraphe sélectionné à partir d'un ensemble de passages textuels extraits du Web ou à partir de bases de connaissances (Asai *et al.*, 2018 ; Du et Cardie, 2018), comme elle peut également être uniquement une réponse courte, par exemple un groupe nominal (Wu *et al.*, 2003 ; Bhaskar *et al.*, 2013 ; Le *et al.*, 2016). Dans les systèmes qui extraient les réponses à partir de bases de connaissances, la réponse prend généralement une forme très concise se limitant à une information brièvement représentée et qui, certes, permet de répondre à la question, mais qui manque considérablement de contexte. Ces formes de réponses, trop brèves ou trop longues pourraient considérablement entraver le dialogue « homme machine » en le rendant moins naturel.

Malgré l'abondance des travaux dans le domaine des SQR, la problématique de formulation des réponses a reçu très peu d'attention. Une première approche traitant indirectement cette tâche a été proposée dans Brill *et al.* (2001) et Brill *et al.* (2002). En effet, les auteurs avaient pour but de diversifier les motifs possibles de réponses en permutant les termes de la question en vue de maximiser le nombre de documents extraits susceptibles de contenir la réponse. Une autre approche de représentation de réponse basée sur des règles de reformulation a été également proposée dans Agichtein et Gravano (2000) et Lawrence et Giles (1998) dans le contexte de l'expansion de requêtes pour la recherche de documents et non pour l'extraction de la réponse exacte.

Le peu de travaux qui se sont intéressés à cette tâche dans le cadre des SQR l'ont adressée sous l'angle de la génération de résumés de textes (Ishida *et al.*, 2018 ; Iida *et al.*, 2019 ; Rush *et al.*, 2015 ; Chopra *et al.*, 2016 ; Nallapati *et al.*, 2016 ; Miao et Blunsom, 2016 ; See *et al.*, 2017 ; Oh *et al.*, 2016 ; Sharp *et al.*, 2016 ; Tan *et al.*, 2016 ; dos Santos *et al.*, 2016). La majorité de ces travaux n'ont considéré que les questions de causalité de type « *pourquoi* » où les réponses sont des paragraphes. Pour rendre ces réponses plus concises, ils procèdent à un compactage des paragraphes extraits.

D'autres approches (Kruengkrai *et al.*, 2017 ; Girju, 2003 ; Verberne *et al.*, 2011 ; Oh *et al.*, 2013) ont exploré cette tâche comme un problème de classification où il s'agit de prédire si un passage de texte pourrait constituer une réponse à une question donnée.

Il faut noter que ces approches ont pour seul but de diversifier au maximum les formules possibles pour augmenter la probabilité d'extraire la bonne réponse et non pour générer une réponse qui soit conviviale pour l'utilisateur. Il faut également souligner que ces approches ne sont valables que pour les SQR qui génèrent les réponses sous forme d'extraits de textes et ne pourront pas être appliquées aux réponses courtes.

Les travaux présentés dans (Pal *et al.*, 2019) ont tenté d'aborder ce problème en proposant une approche supervisée dont l'apprentissage s'est fait sur un petit ensemble de données dont les paires questions-réponses ont été extraites à partir de corpus de

données axés sur la tâche de compréhension de texte et ont été également ajoutées manuellement, ce qui rend la généralisation et la capture de la variation très limitées.

Notre approche de génération de réponses concises diffère de ces travaux car elle est non supervisée, elle peut s'adapter à n'importe quel type de questions factuelles (à l'exception de celles de type *pourquoi*) et elle s'appuie sur des données facilement accessibles et non annotées. Pour cela, nous nous basons sur l'analyse en dépendances afin d'avoir une idée du rôle de la réponse dans la question posée, par exemple, pour savoir si la réponse est le sujet ou l'objet de la question.

2.2. *Analyse en dépendances*

L'analyse en dépendances est très utilisée pour obtenir la structure d'une phrase et bien d'autres tâches de TALN s'appuient sur cette analyse. C'est notamment le cas depuis l'arrivée des outils d'analyse à base de transition en apprentissage supervisé (Nivre, 2003), faciles à utiliser, comme par exemple Maltparser, Nivre *et al.* (2006)) et plus récemment des outils d'analyse à base de graphes (Kiperwasser et Goldberg, 2016 ; Dozat *et al.*, 2017). Des approches et des outils très performants et d'une haute qualité ont été présentés lors de deux campagnes d'évaluation de l'analyse en dépendances CoNLL 2017 (Zeman *et al.*, 2017) et CoNLL 2018 (Zeman *et al.*, 2018). En 2017, des équipes participantes utilisaient des parsers à base de transition ou à base de graphes (Kübler *et al.*, 2009), mais suite aux bons résultats du gagnant de CoNLL 2017 (Dozat *et al.*, 2017) (un parser à base de graphes) les outils les plus performants sont maintenant presque tous à base de graphes, comme, par exemple, les gagnants en termes de MLAS de CoNLL 2018 (Straka (2018) ; Kondratyuk et Straka (2019)).

Pour l'instant la plupart des analyseurs à base d'apprentissage supervisé utilisent les treebanks fournis par le projet Universal Dependencies (UD)[1] (Nivre *et al.*, 2016). Ce projet fournit 183 treebanks en 104 langues[2]. Certains treebanks sont néanmoins très petits, mais comme les approches de la campagne CoNLL 2018 ont pu le montrer, la plupart d'entre eux permettent d'apprendre des analyseurs pour obtenir des résultats de bonne qualité. Malgré le fait qu'une partie des treebanks a été créée antérieurement au projet UD (p. ex. Abeillé *et al.* (2003) pour le français et Marneffe et Manning (2008) pour l'anglais). Des apprentissages crosslingues sont possibles, car tous les treebanks ont été annotés en suivant le même guide d'annotation et utilisent les mêmes catégories pour désigner les parties de discours ainsi que les relations en dépendances et les traits syntactico-morphologiques.

Les treebanks du projet UD sont actuellement en train d'être enrichis par des *enhanced dependencies*, qui en plus des relations de base indiquent des relations indirectes entre mots comme le sujet d'un verbe coordonné, etc. (Nivre *et al.*, 2018 ; Oepen

1. `http://universaldependencies.org`
2. Version 2.7 du 15 novembre 2020, `http://hdl.handle.net/11234/1-3424`

et al., 2020). En revanche, pour l'instant les treebanks pour l'anglais ou le français ne sont pas encore exhaustivement annotés.

Pour évaluer l'analyse en dépendances, quatre métriques similaires sont actuellement utilisées :

– *Unlabeled Attachement Score* (UAS), qui exprime le pourcentage des mots étant attachés à la bonne tête sans prendre en compte le type de relation de dépendance ;

– *Labeled Attachement Score* (LAS), qui exprime le pourcentage des mots étant attachés à la bonne tête et ayant le bon type de relation de dépendance. Le LAS est la valeur f-mesure (Zeman *et al.*, 2018) :

$$P = \frac{\#nœudsCorrects}{\#nœudsPrédits} \qquad R = \frac{\#nœudsCorrects}{\#nœudsGold}$$

$$LAS = F_1 = \frac{2PR}{P+R}$$

Si le LAS n'est pas pondéré, P et R sont toujours identiques, donc $LAS = F_1 = P = R$. Pour le LAS pondéré *#nœudCorrects*, *#nœudsPrédits* et *#nœudGold* ne sont pas simplement les sommes des mots correctement annotés, mais pour les mots fonctionnels un facteur n (par exemple 0,1) est appliqué. Une erreur de partie de discours influence donc le LAS et précision et rappel ne sont plus forcément identiques ;

– *Content Word Labeled Attachement Score* (CLAS), une variante du LAS, qui ignore les relations de dépendance des mots fonctionnels afin de pouvoir comparer des scores de langues très différentes (Nivre et Fang, 2017). Par exemple, le finnois ayant des cas locaux au lieu de prépositions a moins de mots que le français pour une phrase comme *tu vas de Helsinki à Turku*, donc un parser ne peut pas attacher incorrectement un mot absent (par exemple le pronom et les prépositions dans la traduction finnoise : *menet Helsingistä Turkuun*) ;

– *Morphology Aware Labeled Attachement Score* (MLAS), une extension du CLAS, qui en plus prend en compte les valeurs des parties de discours et les traits morphologiques en plus de l'arbre syntaxique (Zeman *et al.*, 2018).

Pour une comparaison de qualité de deux versions de modèles pour une langue, LAS est la métrique à préférer, car il exprime la qualité globale de l'analyse. Pour la comparaison crosslingue d'un modèle CLAS et MLAS sont plus adaptés afin de ne pas « désavantager » les langues avec beaucoup des mots fonctionnels (comme le français).

3. Approche de génération d'une réponse naturelle

L'approche de génération de réponses que nous décrivons dans cet article est un composant d'un SQR qui a été développé par Rojas Barahona *et al.* (2019). Comme illustré dans la figure 1, l'architecture de ce système se compose d'un frontal de traite-

ment de la parole, d'un composant de compréhension, d'un gestionnaire de contexte, d'un composant de génération et d'un composant de synthèse.

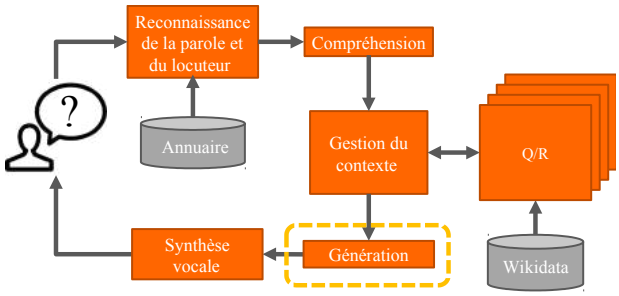

Figure 1. *L'architecture globale du système de questions-réponses conversationnel*

Il s'agit premièrement de comprendre la question posée par l'utilisateur puis de traduire cette question en langue naturelle (français ou anglais) dans une représentation formelle pour ensuite transformer cette représentation formelle en une requête Sparql[3]. Grâce à la requête Sparql nous cherchons la réponse dans une base de connaissances RDF, dans notre cas Wikidata[4]. La réponse est toujours une liste d'URI ou de valeurs. Prenons l'exemple de la question *Qui a joué le rôle de Gus Fring dans Breaking Bad ?* la requête Sparql

```
SELECT DISTINCT ?uri WHERE { wd:Q5620660 ^pq:P453/ps:P161 ?uri }
```

sera envoyée à Wikidata pour extraire la réponse. L'entité `wd:Q5620660` représente le rôle de *Gus Fring* ainsi que le prédicat `^pq:P453/ps:P161`[5] représente le chemin dans le graphe de connaissance entre `wd:Q5620660` et la réponse (`wd:Q726142`) qui réfère à l'acteur *Giancarlo Esposito* en passant par d'autres nœuds.

Bien que nous arrivons à trouver la réponse exacte à une question, sa représentation n'est pas conviviale pour l'utilisateur. De ce fait, nous proposons une approche non supervisée qui intègre l'utilisation des modèles transformers tels que BERT (Devlin *et al.*, 2019) et GPT (Radford *et al.*, 2018). Le choix d'une approche non supervisée émane du fait qu'il n'existe pas un corpus d'apprentissage associant une question à une réponse compacte, exhaustive et qui permettrait d'appliquer en mode supervisé une architecture neuronale *end-to-end* apprenant à générer une phrase répondant à une question. Cette approche part du fait que nous avons déjà extrait la réponse exacte à une question posée. Nous supposons qu'une réponse bien formulée n'est que la reformulation de la question même associée à la réponse exacte. Cette approche comprend deux étapes fondamentales. La première étape consiste à effectuer une analyse en dé-

3. `https://www.w3.org/TR/sparql11-overview/`
4. `https://www.wikidata.org/`
5. `P453` « rôle de », `P161` « acteur ».

pendances de la question en entrée et nous procédons dans une deuxième étape à la génération de la réponse.

3.1. *Analyse en dépendances*

Pour cette première étape d'analyse en dépendances, nous utilisons une version améliorée de UdpipeFuture (Straka, 2018) qui était le système gagnant en termes de la métrique MLAS de la tâche de l'analyse en dépendances (Zeman *et al.*, 2018). UdpipeFuture est un analyseur, qui fait l'étiquetage en parties de discours et la lemmatisation avec un LSTM. L'analyse en dépendances est faite avec un parser à base de graphes, inspiré de Dozat *et al.* (2017).

Notre modification consiste à intégrer les plongements contextuels à UdpipeFuture lors de l'apprentissage. Pour cela, nous nous sommes orientés vers BERT multilingue (Devlin *et al.*, 2019), XLM-R (Conneau *et al.*, 2019) (pour l'anglais et le français), RoBERTA (Liu *et al.*, 2019) (pour l'anglais), FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2019) (pour le français[6]) lors de l'apprentissage des treebanks French-GSD et English-EWT[7], issus du projet Universal Dependencies (UD). L'ajout des plongements contextuels a significativement augmenté les résultats pour les trois métriques, LAS, CLAS et MLAS (cf. tableau 1 pour le français et l'anglais, pour d'autres langues cf. Heinecke (2020))

français (UD-French-GSD)

	Straka (2018)	FlauBERT	BERT	CamemBERT	**XLM-R**
MLAS	77,29	79,53	81,64	82,17	**82,62**
CLAS	82,49	84,16	86,21	86,45	**86,94**
LAS	85,74	87,98	89,68	89,67	**89,82**

anglais (UD-English-EWT)

	Straka (2018)	BERT	RoBERTA	**XLM-R**
MLAS	74,71	81,16	82,38	**82,91**
CLAS	79,14	85,89	86,89	**87,24**
LAS	82,51	88,63	89,40	**89,54**

Tableau 1. *Analyse en dépendances du français et de l'anglais (UD 2.2), les meilleurs résultats en gras*

Néanmoins, pour l'analyse en dépendances des questions simples de type quiz, les deux treebanks UD (French-GSD et English-EWT) ne sont pas adaptés, car leurs

6. Nos expérimentations s'appuient sur `flaubert_base_cased` et `camembert-base`.
7. Comme la campagne d'évaluation CoNLL 2018, nous utilisons la version 2.2 des treebanks UD pour la comparaison.

corpus d'apprentissage ne contiennent pas ou très peu de questions[8]. Les mauvais résultats de l'analyse en dépendances des questions avec des modèles appris avec ces treebanks sont résumés en tableau 2.

français (Fr-GSD)

	BERT	**CamemBERT**	FlauBERT	XLM-R
MLAS	60,52	**61,32**	58,09	59,23
CLAS	73,04	**75,26**	70,96	73,52
LAS	79,27	**80,49**	78,40	79,27

anglais (En-EWT)

	BERT	RoBERTa	XLM-R
MLAS	**80,45**	80,68	80,68
CLAS	**88,02**	89,17	89,42
LAS	**90,58**	91,49	91,88

Tableau 2. *Analyse des questions avec des modèles appris sur UD sans modifications*

Afin d'améliorer l'analyse, nous avons enrichi les treebanks d'apprentissage French-GSD et English-EWT en annotant 309 questions anglaises des challenges QALD7 (Usbeck *et al.*, 2017) et QALD8[9] (ainsi 91 questions pour le test) en supprimant les doublons. Pour le français, nous avons traduit des questions issues de QALD7, et formulé des questions nous-mêmes (66 pour le test, 267 pour l'apprentissage). Les annotations ont été effectuées par deux linguistes avec le guide d'annotation du projet Universal Dependencies[10] et la documentation des treebanks French-GSD (pour les questions françaises) et English-EWT (pour l'anglais). Nous avons procédé à une pré-annotation automatique des deux corpus de questions avec des modèles appris sur French-GSD et English-EWT. Puis nous avons effectué deux passes de validation et corrections. Ensuite nous avons fait une évaluation 4-fold. Comme le tableau 3 le montre, la qualité de l'analyse augmente considérablement. Les *embeddings* CamemBERT (pour le français) et BERT (anglais) ont à nouveau le meilleur impact.

Nous nous appuyons sur la version Udpipe-Future[11] que nous avons améliorée avec BERT et CamemBERT et qui donne les meilleurs résultats en termes d'analyse en dépendances pour procéder au découpage de la question en fragments textuels (appelés également *chunks*) : $Q = \{c_1, c_2, \ldots, c_n\}$.

8. Il existe également un treebank de questions, French-FQB (Seddah et Candito, 2016), mais la plupart des questions de ce treebank sont plutôt des questions longues, et peu similaires aux questions « typiques » du quiz.

9. `https://github.com/ag-sc/QALD`

10. `https://universaldependencies.org/guidelines.html`

11. `https://github.com/Orange-Opensource/udparse`

français (French-GSD)

	BERT	**CamemBERT**	FlauBERT	XLM-R
MLAS	91,20	**92,12**	90,53	91,23
CLAS	96,10	**97,37**	94,74	96,14
LAS	97,55	**98,26**	96,86	97,56

anglais (English-EWT)

	BERT	RoBERTa	XLM-R
MLAS	**84,85**	83,08	83,08
CLAS	**91,92**	91,67	90,66
LAS	**94,24**	93,85	93,59

Tableau 3. *Analyse des questions avec des corpus d'apprentissage UD de base enrichis de questions*

Si on reprend l'exemple précédent de la question *Qui a joué le rôle de Gus Fring dans Breaking Bad ?* l'ensemble des fragments textuels serait *Q* = {*Qui, a joué, le rôle de Gus Fring, dans Breaking Bad*}.

3.2. *Processus de génération de réponses*

Dans cette deuxième étape, nous procédons d'abord à un premier test de l'ensemble Q pour vérifier si le fragment textuel qui contient un marqueur de question (*quel*, *quand*, *qui*, etc.) représente le sujet *nsubj* ou l'objet *obj* dans l'arbre en dépendances de la question analysée. Si c'est le cas, nous remplaçons tout simplement ce fragment textuel par la réponse que nous avons identifiée précédemment. Reprenons l'exemple précédent de la question *Qui a joué le rôle de Gus Fring dans Breaking Bad ?*. Le système détecte automatiquement que le fragment textuel contenant le marqueur de question *Qui* représente bien le sujet. Ce sujet sera donc remplacé directement par la réponse exacte *Giancarlo Esposito*. Par la suite, la réponse générée sera *Giancarlo Esposito a joué le rôle de Gus Fring dans Breaking Bad*. Autrement, nous procédons à la suppression du fragment textuel contenant le marqueur de question que nous avons détecté et nous rajoutons la réponse *R* à l'ensemble *Q* :

$$Q = \{c_1, c_2, \ldots, c_{n-1}, R\}$$

À partir de l'ensemble de fragments textuels *Q*, nous générons par permutation toutes les structures de réponses possibles qui peuvent former la phrase répondant à la question traitée :

$$S = \{s_1(R, c_1, c_2, \ldots, c_{n-1}), s_2(c_1, R, c_2, \ldots, c_{n-1}), \ldots, s_m(c_1, c_2, \ldots, c_{n-1}, R)\}$$

Nous nous référons à l'utilisation d'un modèle de langue (ML) qui permet d'assigner une probabilité d'occurrences pour les séquences de mots générées. Dans notre approche, l'ensemble des structures S sera évalué par un modèle de langue basé sur des modèles transformer qui permettra d'extraire la séquence de fragments textuels la plus probable qui servira de réponse :

$$structure^* = s \in S; p(s) = \underset{s_i \in S}{\operatorname{argmax}}\, p(s_i)$$

Une fois que nous avons identifié la structure qui représentera la réponse à la question traitée, nous passons à la génération des termes manquants. En effet, nous supposons qu'il pourrait y avoir un ou plusieurs termes qui ne figurent pas nécessairement dans la question ou dans la réponse mais qui sont en revanche nécessaires à la génération d'une bonne structure grammaticale de la réponse. Ce processus nécessite que nous définissions deux paramètres, le nombre de termes manquants possible et leurs positions dans la structure sélectionnée. Dans cet article, pour fixer ces deux paramètres, nous faisons l'hypothèse qu'un seul terme pourrait être manquant et qu'il est situé avant la réponse courte dans la structure identifiée, comme cela pourrait être le cas pour un article défini manquant (*la, les*, etc.) ou encore une préposition (*dans, à*, etc.) par exemple. Par conséquent, pour prédire ce terme manquant, nous utilisons des modèles de génération (MG) basés sur le modèle transformer BERT pour sa capacité à capturer de manière bidirectionnelle le contexte d'un mot donné dans une phrase. Dans le cas où le modèle de génération renvoie une séquence de caractères non alphabétiques, nous supposons que la structure optimale, telle que prédite par le ML, n'a pas besoin d'être complétée par un terme supplémentaire. Dans ce qui suit, nous illustrons le déroulement des différentes étapes de l'approche proposée avec un exemple en anglais :

Question : *how far is Ponte Vedra beach from Jacksonville FL ?*

1) Analyse de la question et extraction de la réponse moyennant notre SQR (Rojas Barahona *et al.*, 2019) :
   Réponse_courte = {eighteen miles southeast}
2) Découpage de la question en fragments textuels à partir de l'analyse en dépendances que nous avons définie :
   Q = {*How far*, *is*, *Ponte Vedra beach*, *from Jacksonville FL*}
3) Suppression des marqueurs de question (*How far*) :
   Q = {*is*, *Ponte Vedra beach*, *from Jacksonville FL*}
4) Ajout de la réponse courte extraite :
   Q = {*is*, *Ponte Vedra beach*, *from Jacksonville FL*, *eighteen miles southeast*}
5) Génération des structures de réponses possibles *S* :
   S = {*Ponte Vedra beach*, *is*, *from Jacksonville FL*, *eighteen miles southeast* ;
   from Jacksonville FL, *Ponte Vedra beach*, *is*, *eighteen miles southeast* ; ...}
6) Évaluation des structures par un modèle de langue :
   p(structure)* $= \operatorname{argmax}_{s_i \in S}\, p(s_i)$:
   structure* = *Ponte Vedra beach*, *is*, *eighteen miles southeast*, *from Jacksonville FL*

7) Génération des termes possiblement manquants à *structure** par un modèle de génération (mot manquant = *about*) :
Ponte Vedra beach is [mot manquant] *eighteen miles southeast from Jacksonville FL*

Réponse : *Ponte Vedra beach is about eighteen miles southeast from Jacksonville FL.*

Comme nous pouvons le remarquer, la réponse finale générée avec l'ajout du terme manquant s'apparente considérablement à une réponse naturelle qui pourrait être émise par un humain.

4. Expérimentation et évaluation

Les corpus de tests existants pour l'évaluation des SQR sont soit adaptés aux systèmes qui génèrent la réponse exacte à la question et donc une réponse courte, soit plus axés vers la tâche de *Machine Reading Comprehension* où la réponse est un passage de texte contenant la réponse exacte. Par la suite, nous avons créé un jeu de données qui consiste à associer des questions extraites du corpus QALD-7 challenge (Usbeck *et al.*, 2017) avec des réponses en langage naturel qui ont été définies manuellement par un linguiste et que nous avons revues individuellement. Ce corpus appelé Quereo consiste en 150 questions avec leurs réponses exactes. On note en moyenne trois réponses possibles en langage naturel pour chaque question. Ce corpus existe en versions française et anglaise.

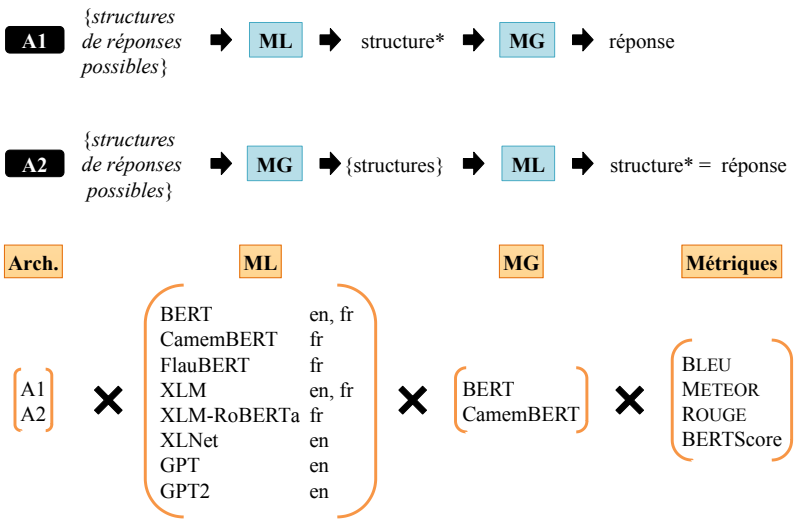

Figure 2. *Cadre d'expérimentation pour identifier la meilleure configuration*

Comme illustré dans la figure 2, nous avons évalué deux architectures possibles de notre approche pour la génération de réponses. La première architecture *A1* consiste

d'abord à lister toutes les structures de réponses possibles, puis à les faire évaluer par un ML qui permet la sélection de la structure optimale, enfin de générer avec le MG le terme manquant dans la structure sélectionnée. La deuxième architecture *A2* consiste d'abord à lister toutes les structures de réponses possibles, puis à générer avec le MG les termes manquants dans chaque structure, enfin à faire évaluer l'ensemble de ces structures par le ML pour sélectionner la structure optimale. Pour cet article, nous supposons qu'il y a seulement un terme manquant par structure.

Pour évaluer l'approche proposée, nous avons opté pour trois métriques *n*-gram (BLEU, METEOR et ROUGE) utilisées dans la littérature pour évaluer ce type de tâche et la métrique BERTScore qui exploite les plongements lexicaux pré-entraînés de BERT pour calculer la similarité entre la réponse générée et la réponse de référence. Pour pouvoir comparer les différentes variantes de l'approche, nous nous sommes référés au test de Friedman (Milton, 1939) qui permet de détecter les écarts de performances entre plusieurs modèles évalués par plusieurs métriques en se basant sur les rangs moyens.

Nous avons également mené une évaluation humaine pour les versions française et anglaise du corpus de données, dans laquelle nous avons demandé à 20 locuteurs natifs des deux langues d'évaluer la pertinence d'une réponse générée (*correcte* ou *pas correcte*) pour une question donnée en indiquant le ou les types d'erreurs détectées (*accord grammatical*, *préposition incorrecte*, *ordre des mots*, etc.). La figure 3 présente le cadre d'évaluation que nous avons mis en œuvre et fourni aux participants. Les résultats de chaque participant sont enregistrés dans un fichier *json*. Le taux d'interaccords entre les participants qui a été mesuré par le coefficient Kappa de Fleiss (Fleiss, 1971) a atteint 70 %, ce qui indique un accord substantiel d'après le tableau d'interprétation de Landis et Koch (1977). À travers l'étude d'évaluation humaine, nous voulions explorer dans quelle mesure les métriques standard sont fiables pour évaluer les approches GAT dans le contexte des systèmes de questions-réponses.

Le tableau 4 (corpus français) et le tableau 6 (corpus anglais) ne présentent que les résultats obtenus pour les trois meilleurs modèles selon le classement du test de Friedman. Les modèles de langue utilisés sont adaptés selon la langue du corpus mis en test. Vanté par ses mérites en tant que modèle génératif très puissant entraîné sur un très large corpus de données constitué de 8 millions de pages Web associant entre autres des termes en anglais et en français, le modèle GPT a été également testé avec le corpus français pour voir s'il arrivait à détecter la meilleure structure à choisir pour une question. En effet, notre corpus d'évaluation peut, dans certains, cas inclure des questions qui associent des termes en anglais et en français, tels que le nom d'un film. Les valeurs mises entre crochets représentent le rang d'un modèle selon la métrique utilisée.

Nous notons que le score de précision le plus élevé pour le français d'environ 85 % a été obtenu avec la première architecture avec BERT comme modèle de génération (MG) et CamemBERT comme modèle de langage (ML). On remarque également que l'architecture *A1*, qui considère l'évaluation de la structure par un ML avant de générer les termes manquants, fonctionne mieux. Étonnamment, en tant que modèle génératif,

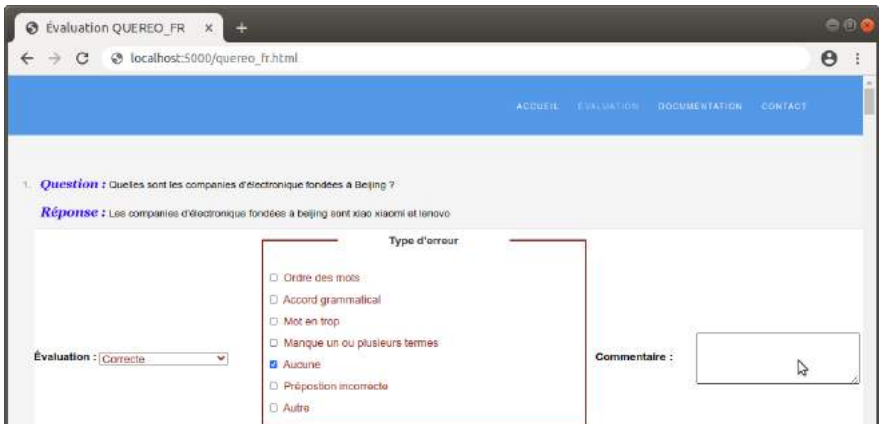

Figure 3. *La plate-forme d'évaluation*

le modèle multilingue BERT prédit mieux les termes manquants que CamemBERT pour les réponses du corpus français. Ces résultats sont également confirmés par le test de Friedman où nous pouvons clairement voir que la première configuration classée correspond à la meilleure configuration classée selon l'évaluation humaine, avec une légère différence par rapport aux autres configurations. Voyons si cela signifie que les quatre métriques sont corrélées à la précision humaine.

D'après le tableau 5 qui présente la corrélation Pearson (Benesty *et al.*, 2009) entre la précision humaine avec les quatre métriques et la figure 4 qui illustre le classement donné par chaque métrique d'évaluation avec le jugement humain pour chaque configuration (c.-à-d. *configuration* = *MG* × *architecture* × *ML*) testée, nous pouvons clairement voir que les résultats de l'évaluation humaine sont positivement et fortement corrélés aux scores BLEU, METEOR et BERT. Ces métriques correspondent pratiquement au classement humain et sont donc évidemment capables d'identifier quelle configuration donne les meilleurs résultats. En revanche, la métrique ROUGE, utilisée pour l'évaluation des questions-réponses en français, est modérément corrélée à l'évaluation humaine, ce qui signifie que cette métrique ne doit pas être considérée comme la seule métrique d'évaluation pour évaluer une telle tâche. En revanche, lorsque la métrique ROUGE est considérée avec les autres métriques, elle permet de se rapprocher du jugement humain.

Le tableau 6 présente les résultats pour le corpus de données anglais et indique que la meilleure précision est d'environ 72 % avec *A1*, BERT comme modèle de génération et le transformer GPT (*Generative Pretrained Model*) comme modèle de langue. Selon les trois premières configurations, c'est l'architecture *A2* qui se démarque et le transformer GPT qui prend le dessus sur les autres modèles de langue.

rang hum.	rang Friedman	Arch.	MG	ML	*accuracy* humaine
1	**1**	**A1**	**BERT**	**CamemBERT**	**84,85**
2	2	A2	BERT	FlauBERT-small-cased	84,09
2	3	A1	BERT	XLM-RoBERTa-base	84,09
2	9	A1	BERT	BERT-base-multilingual-cased	84,09
5	4	A1	BERT	FlauBERT-base-uncased	83,33
5	5	A1	BERT	mlm-1024	83,33
5	6	A2	BERT	GPT2	83,33
5	10	A2	BERT	XLM-clm-enfr-1024	83,33
5	11	A1	BERT	XLM-clm-enfr-1024	83,33
5	12	A2	BERT	FlauBERT-large-cased	83,33
5	13	A2	BERT	mlm-1024	83,33
5	14	A2	BERT	XLM-RoBERTa-base	83,33

rang	rang	BLEU		METEOR		ROUGE		BERTS	
hum.	Friedman	score	rang	score	rang	score	rang	score	rang
1	**1**	**86,28**	**[1]**	**96,76**	**[1]**	**93,69**	**[6]**	**97,89**	**[2]**
2	2	85,87	[7]	96,75	[2]	94,22	[1]	97,96	[1]
2	3	85,93	[4]	96,63	[6]	93,79	[5]	97,88	[3]
2	9	85,01	[19]	96,52	[22]	93,81	[4]	97,79	[7]
5	4	86,17	[2]	96,72	[3]	93,56	[14]	97,81	[6]
5	5	85,39	[10]	96,60	[8]	93,61	[10]	97,83	[4]
5	6	85,46	[9]	96,67	[4]	93,48	[17]	97,76	[10]
5	10	85,89	[6]	96,55	[18]	93,57	[13]	97,71	[19]
5	11	84,99	[20]	96,52	[23]	93,87	[3]	97,76	[12]
5	12	86,15	[3]	96,57	[13]	93,14	[37]	97,79	[8]
5	13	85,90	[5]	96,54	[20]	93,30	[27]	97,76	[11]
5	14	85,32	[13]	96,46	[28]	93,63	[9]	97,71	[17]

Tableau 4. *Classement des modèles selon l'évaluation humaine (meilleur en gras) et le test Friedman (meilleur en jaune), corpus français*

	BLEU	METEOR	ROUGE	BERT-score
Quereo_fr	98 %	99 %	46 %	97%
Quereo_en	85 %	80 %	83 %	88 %

Tableau 5. *Coefficient de la corrélation de Pearson entre les quatre métriques et l'évaluation humaine*

Ces résultats sont confirmés par le test de Friedman avec une légère différence de classement et également appuyés par les scores de corrélation entre l'évaluation humaine et chacune des quatre métriques comme le montre le tableau 5.

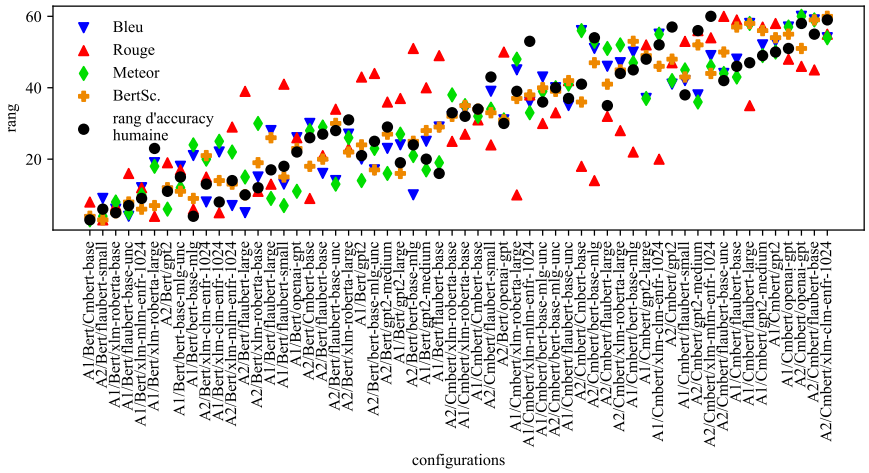

Figure 4. *Corrélation entre les évaluations humaines et les métriques* BLEU, METEOR, ROUGE *et* BERT *score pour le Q/R en français (« CmBERT » veut dire CamemBERT)*

L'objectif de ces résultats est d'identifier parmi les métriques standard celles qui sont fiables pour évaluer la tâche de génération de réponses dans le contexte des SQR.

Nous avons également procédé à l'analyse des erreurs signalées par les participants. Comme nous pouvons le constater à partir de la figure 5, l'erreur la plus courante signalée pour les deux corpus de données anglais et français est l'ordre des mots. Ceci sous-entend un problème lié à la phase d'évaluation des structures de réponses possibles par le modèle de langue. La deuxième erreur la plus signalée est l'indication d'un ou plusieurs termes manquants dans la réponse (corpus français) ou la présence de termes intrus (anglais). Ceci concerne le processus de génération (MG).

Figure 5. *Distribution d'erreurs de la génération*

rang hum.	rang Fr.	arch.	MG	ML	*accuracy* humaine
1	**1**	**A2**	**BERT–base-multiling-cased**	**GPT**	**72,36**
1	2	A1	BERT-base-multiling-cased	GPT	72,36
3	2	A1	BERT-large-cased	GPT	71,55
3	4	A2	BERT-large-cased	GPT2	71,55
3	5	A2	BERT-base-multiling-cased	GPT2	71,55
3	6	A2	BERT-base-multiling-cased	GPT2-large	71,55
3	7	A2	BERT-base-multiling-cased	GPT2-medium	71,55
3	7	A2	BERT-large-cased	GPT2-medium	71,55
3	10	A2	BERT-large-cased	GPT	71,55
10	7	A2	BERT-large-cased	GPT2-large	70,73
10	30	A1	BERT-base-multiling-cased	BERT-base-unc.	70,73

rang	rang	BLEU		METEOR		ROUGE		BERTS	
hum.	Friedman	score	rang	score	rang	score	rang	score	rang
1	**1**	**78.25**	**[2]**	**94,63**	**[1]**	**92,83**	**[2]**	**97,21**	**[3]**
1	2	78,25	[1]	94,51	[2]	92,53	[10]	97,23	[2]
3	2	77,12	[6]	94,45	[3]	92,80	[5]	97,32	[1]
3	4	76,98	[7]	94,40	[7]	92,86	[1]	97,17	[5]
3	5	77,53	[4]	94,39	[8]	92,82	[4]	97,14	[6]
3	6	77,85	[3]	94,41	[5]	92,65	[7]	97,07	[10]
3	7	77,42	[5]	94,40	[6]	92,58	[9]	97,10	[9]
3	7	75,96	[13]	94,41	[4]	92,60	[8]	97,18	[4]
3	10	76,28	[11]	94,30	[9]	92,76	[6]	97,14	[7]
10	7	76,74	[8]	94,26	[10]	92,83	[3]	97,14	[8]
10	30	74,85	[30]	93,94	[31]	90,86	[26]	96,61	[28]

Tableau 6. *Classement des modèles selon l'évaluation humaine (meilleur en gras) et le test Friedman (meilleur en jaune), corpus anglais*

5. Une approche de génération automatique de corpus de questions-réponses

Les résultats prometteurs que nous avons obtenus suite aux deux évaluations que nous avons conduites, nous ont amenés à envisager d'utiliser cette même approche pour construire des corpus de type questions-réponses qui associent une réponse en langage naturel à une question en langage naturel. Notre idée est de générer ainsi des corpus d'apprentissage de grande taille pour entraîner des approches neuronales de type *end-to-end*. Notre approche de génération de réponses concises conduit évidemment à des réponses synthétiques un peu stéréotypées. Pour introduire un peu de variété dans l'expression des réponses, il était nécessaire de rajouter dans ce corpus d'apprentissage des réponses plus naturelles.

Nous avons eu l'idée d'exploiter les corpus MRQA (*Machine Reading for Question Answering*) existants pour alimenter nos corpus d'apprentissage. Dans un corpus MRQA, chaque entrée comporte une question en langage naturel, une réponse courte et une réponse longue, représentée sous forme d'un paragraphe de plusieurs phrases dont au moins une contient la réponse courte. La figure 6 montre un exemple de ce type de représentation de réponse[12].

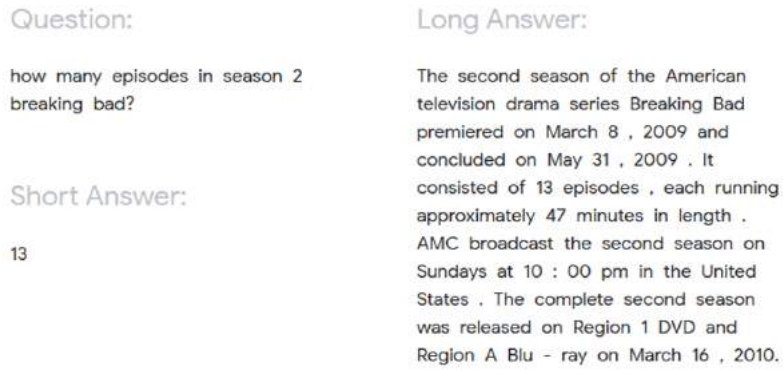

Figure 6. *Un exemple de questions-réponses extrait du corpus GNQ*

L'approche que nous proposons consiste à explorer l'utilisation des corpus MRQA pour d'une part, générer une réponse synthétique à partir de la question et de la réponse courte (avec l'approche de génération décrite précédemment) et d'autre part, extraire une réponse concise et naturelle à partir de la réponse longue. Ce procédé d'augmentation des données (Shorten et Khoshgoftaar, 2019) permet de régulariser et de réduire le surajustement lors de l'apprentissage.

Pour extraire une réponse concise et naturelle à partir de chaque réponse longue, nous avons d'abord supprimé de ces réponses les références externes vers d'autres pages ou les notes de bas de page, choses que l'on retrouve souvent dans les pages Wikipédia ou dans des articles. Puis, nous avons découpé chaque réponse longue en phrases, pour ne retenir que celles contenant la réponse courte. Enfin, nous avons calculé la similarité sémantique entre la question et chaque phrase candidate pour ne retenir comme réponse naturelle que la phrase ayant le score de similarité le plus élevé.

On dénote :

$$C = \{(Q_1, Sa_1, La_1), (Q_2, Sa_2, La_2), \ldots, (Q_i, Sa_i, La_i)\}; i \in [1, m]$$

un corpus MRQA qui contient un ensemble de triplets *(question, réponse courte, réponse longue)*, m étant le nombre de triplets dans le corpus. La_i =

12. `https://ai.google.com/research/NaturalQuestions`

$\{S_1, S_2, \ldots, S_j\}; j \in [1, n]$ est la réponse longue pour la question Q_i segmentée en un ensemble de phrases S_j, *n* étant le nombre de phrases que peut contenir une réponse longue. Nous avons procédé à un prétraitement de ces phrases afin de supprimer les références externes vers d'autres pages ou les notes de bas de pages que l'on retrouve souvent dans les pages Wikipédia ou dans des articles.

Notre approche consiste à extraire de l'ensemble La_i un sous-ensemble de phrases S_j candidates pour la réponse naturelle à la question posée Q_i :

$$S_{candidates} = \{S_1, S_2, \ldots, S_k\} \sqsubseteq La_i; k \in [1, n] \forall S_k \in S_{candidates} : Sa_i \in S_k$$

Ce sous-ensemble ne rassemble que les phrases qui contiennent la réponse courte Sa_i identifiée pour la question Q_i. Sachant que le nombre de phrases candidates varie dans un intervalle $[1, n]$, deux cas se présentent. Le premier cas survient quand nous avons plus qu'une phrase candidate $k > 1$, nous procédons alors au calcul de ce que l'on appelle *un score de confiance* ($confidence_{\text{score}} \in [0, 1]$) pour chaque phrase candidate. En fait, ce score de confiance représente la similarité sémantique entre une phrase candidate et la question. Ceci permet de ne sélectionner que la phrase qui relève du contexte de la question.

Prenons l'exemple de la question *Qui est le maire de paris ?* illustrée dans la figure 7. On remarque qu'il y a plusieurs phrases candidates comportant la réponse courte (Anne Hidalgo) à la question. Pourtant, toutes ces phrases candidates ne sont pas des réponses acceptables. Nous choisissons donc la phrase dont le sens est le plus proche de celui de la question. Pour cela, nous calculons la similarité sémantique de la question et celle de chaque phrase candidate puis, sélectionnons celle ayant la représentation la plus proche de celle de la question.

Question	Réponse longue
Qui est le maire de Paris ?	Ana María Hidalgo Aleu, dite Anne Hidalgo, née le 19 juin 1959 à San Fernando (Espagne), est une femme politique française possédant également la nationalité espagnole. Anne Hidalgo est titulaire d'une maîtrise de sciences sociales du travail, obtenue à l'université Jean-Moulin-Lyon-III et d'un DEA de droit social et syndical. Membre du Parti socialiste, elle est première adjointe au maire de Paris de 2001 à 2014 et conseillère régionale d'Île-de-France de 2004 à 2014. À l'issue des élections municipales de 2014, Anne Hidalgo devient la première femme maire de Paris et est réélue à la suite des élections municipales de 2020.
Réponse courte	
Anne Hidalgo	

Figure 7. *Un exemple de questions-réponses*

Pour la mesure de similarité sémantique, nous utilisons l'approche Simbow (Charlet et Damnati, 2017) qui a prouvé sa performance dans la tâche de similarité entre les questions dans le challenge SemEval-2017. Cette métrique s'apparente à la métrique *soft-cosinus* mais considère en plus les relations entre mots qui peuvent être d'ordre lexical ou sémantique en faisant introduire dans la formule une matrice de relations :

$$Simbow_{\cos}(X, Y) = \frac{X^t \cdot M \cdot Y}{\sqrt{X^t \cdot M \cdot X}\sqrt{Y^t \cdot M \cdot Y}}$$

$$X^t \cdot M \cdot Y = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i m_{i,j} y_j$$

où M est une matrice d'élément $m_{i,j}$ qui représente la relation entre les termes i et j et qui est calculée par la métrique *cosinus* entre les vecteurs de plongement `word2vec` représentatifs des termes i et j. Ceci introduit la notion de similarité sémantique qui permet d'éviter de se retrouver avec une similarité nulle dans le cas où deux textes n'ont aucun terme en commun.

$$S^* = S \in S_{candidates}$$

$$confidence_{\text{score}}(S^*) = \underset{S_i \in S_{candidates}}{\text{argmax}} \; Simbow_{\cos}(S_i, Q)$$

Dans le cas où il n'existe qu'une seule phrase candidate $S_{candidates} = \{S^*\}$, cette dernière sera considérée comme la meilleure réponse à la question.

Afin d'évaluer cette approche de génération automatique de corpus de questions-réponses naturelles à partir des corpus MRQA, nous nous sommes basés sur le corpus Google's Natural Questions (GNQ, Kwiatkowski *et al.* (2019)), dont un exemple est présenté dans la figure 6. Ce corpus contient des questions posées par de vrais utilisateurs qui, suite à la lecture et la compréhension d'un ensemble d'articles Wikipédia, identifient les sections qui intègrent les réponses. Comme il n'existe pas des réponses de référence pour ce type de corpus, nous avons procédé avec une évaluation humaine. Pour cela, nous avons développé une plate-forme d'évaluation qui expose pour chaque question sa réponse courte, une réponse concise et naturelle, appelée *R1*, extraite de la réponse longue par le procédé que nous venons de décrire dans la présente section et la réponse concise et synthétique, appelée *R2*, générée par l'approche que nous avons exposée dans la section 3. D'un ensemble de questions-réponses qui est de l'ordre de 307 000 questions-réponses, nous avons extrait pour le test un échantillon de 1 000 questions-réponses choisies aléatoirement.

Comme affiché dans la figure, chaque participant est tenu d'indiquer, pour chaque réponse, si elle est *correcte* ou *pas correcte*. En option, il peut aussi indiquer si la réponse est également naturelle. Une réponse est considérée *correcte* si elle est grammaticalement correcte et répond bien à la question posée. Une autre option permet de signaler d'éventuelles erreurs liées, par exemple, à une syntaxe grammaticale incorrecte de la question posée ou à une réponse courte non pertinente. L'ajout de cette option s'est avéré nécessaire car nous avons relevé quelques incohérences de ce type d'erreurs dans le corpus de départ GNQ. Cette évaluation étant toujours en cours, nous exposons dans cet article les résultats que nous avons obtenus à ce jour pour un seul participant. Le tableau 9 expose les résultats préliminaires obtenus.

Ces résultats, même préliminaires, sont toutefois très encourageants et montrent un réel potentiel à l'approche hybride de génération automatique de corpus que nous proposons.

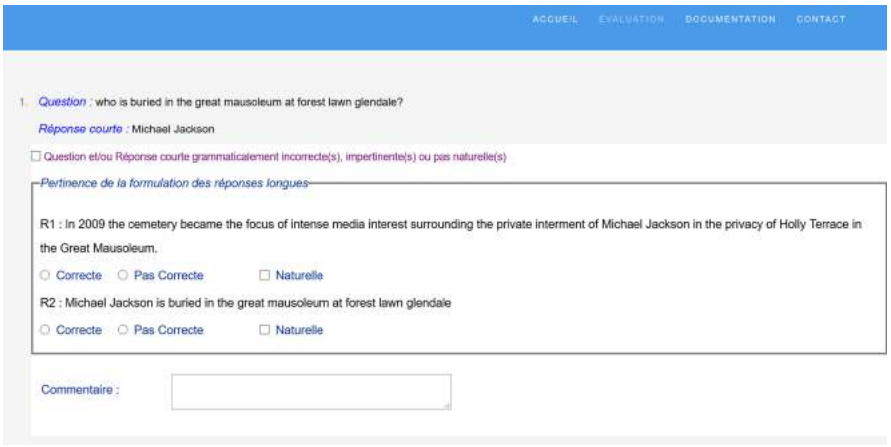

Figure 8. *Plate-forme d'évaluation pour l'approche de génération automatique de corpus de questions-réponses*

Figure 9. *Résultats préliminaires obtenus*

6. Conclusion

Dans cet article, nous avons proposé une approche pour la génération d'une réponse concise, en langage naturel, pour les systèmes de questions-réponses (SQR). Elle s'appuie sur l'analyse en dépendances de la question pour déterminer le rôle grammatical de chaque terme. Elle exploite ensuite la distribution de probabilité des séquences de mots ainsi que des modèles génératifs pour générer une réponse correcte. Les résultats obtenus avec des métriques standard sur des questions de test en français et en anglais sont très prometteurs. De plus, une expérimentation a montré une bonne corrélation entre ces métriques et le jugement humain.

Par ailleurs, nous avons aussi proposé une méthode associant notre approche à un procédé d'extraction de questions-réponses à partir de corpus MRQA dans l'objectif de construire un gros corpus synthétique qui permettrait de tester des approches supervisées sur la tâche de génération de réponses en langage naturel.

L'intégration de cette approche dans un prototype conversationnel (Rojas Barahona *et al.*, 2019) nous a permis d'observer son efficacité dans un contexte d'usage

réel. C'est ainsi que nous avons constaté que cette approche imaginée pour traiter des questions en dehors de tout contexte de dialogue était aussi efficace pour générer des réponses concises dans le cas de certaines questions en contexte. Il serait intéressant de poursuivre nos travaux d'analyse et de mesurer ce phénomène. Par ailleurs, dans le cas d'une utilisation intensive du prototype, la construction de la réponse à partir de la question peut être perçue comme trop stéréotypée. Il serait intéressant d'étudier comment l'utilisation de techniques de reformulation et/ou l'introduction de coréférences sur le sujet de la question permettent de limiter ce phénomène.

Remerciements

Nous tenons à remercier les collègues qui ont participé aux deux campagnes de l'évaluation humaine. Nous présentons également nos sincères remerciements aux relecteurs de cet article pour leurs remarques et suggestions constructives.

7. Bibliographie

Abeillé A., Clément L., Toussenel F., « Building a Treebank for French », *in* A. Abeillé (ed.), *Treebanks*, Springer, Heidelberg, 2003.

Agichtein E., Gravano L., « Snowball : Extracting relations from large plain-text collections », *Proceedings of the fifth ACM conference on Digital libraries*, p. 85-94, 2000.

Asai A., Eriguchi A., Hashimoto K., Tsuruoka Y., « Multilingual extractive reading comprehension by runtime machine translation », `https://arxiv.org/abs/1809.03275`, 2018.

Benesty J., Chen J., Huang Y., Cohen I., *Pearson Correlation Coefficient*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 1-4, 2009.

Bhaskar P., Banerjee S., Pakray P., Banerjee S., Bandyopadhyay S., Gelbukh A., « A hybrid question answering system for Multiple Choice Question (MCQ) », *Question Answering for Machine Reading Evaluation (QA4MRE) at CLEF*, 2013.

Brill E., Dumais S., Banko M., « An analysis of the AskMSR question-answering system », *EMNLP 2002*, ACL, p. 257-264, 2002.

Brill E., Lin J., Banko M., Dumais S., Ng A., « Data-intensive question answering », *TREC 2001*, p. 393-400, 2001.

Charlet D., Damnati G., « SimBow at SemEval-2017 Task 3 : Soft-Cosine Semantic Similarity between Questions for Community Question Answering », *SemEval-2017*, ACL, Vancouver, Canada, p. 315-319, August, 2017.

Chopra S., Auli M., Rush A. M., « Abstractive sentence summarization with attentive recurrent neural networks », *NAACL*, p. 93-98, 2016.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grace E., Ott M., Zettlemoyer L., Stoyanov V., « Unsupervised Cross-lingual Representation Learning at Scale », `https://arxiv.org/abs/1911.02116`, 2019.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of deep bidirectional transformers for language under-standing », *NAACL*, Minneapolis, p. 4171-4186, 2019.

dos Santos C., Tan M., Xiang B., Zhou B., « Attentive pooling networks », `https://arxiv.org/abs/1602.03609`, 2016.

Dozat T., Qi P., Manning C. D., « Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task », *CoNLL 2017 Shared Task. Multilingual Parsing from Raw Text to Universal Dependencies*, ACL, Vancouver, Canada, p. 20-30, 2017.

Du X., Cardie C., « Harvesting Paragraph-level Question-Answer Pairs from Wikipedia », *ACL*, ACL, Melbourne, Australia, p. 1907-1917, July, 2018.

Fleiss J. L., « Measuring nominal scale agreement among many raters. », *Psychological bulletin*, vol. 76, n° 5, p. 378, 1971.

Girju R., « Automatic detection of causal relations for question answering », *ACL workshop on Multilingual summarization and question answering*, ACL, p. 76-83, 2003.

Heinecke J., « Hybrid Enhanced Universal Dependencies Parsing », *IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, ACL, Online, p. 174-180, July, 2020.

Hirschman L., Gaizauskas R., « Natural language question answering : the view from here », *natural language engineering*, vol. 7, n° 4, p. 275-300, 2001.

Iida R., Kruengkrai C., Ishida R., Torisawa K., Oh J.-H., Kloetzer J., « Exploiting Background Knowledge in Compact Answer Generation for Why-Questions », *AAI Conference on Artificial Intelligence*, vol. 33, p. 142-151, 2019.

Ishida R., Torisawa K., Oh J.-H., Iida R., Kruengkrai C., Kloetzer J., « Semi-distantly supervised neural model for generating compact answers to open-domain why questions », *AAAI*, 2018.

Kiperwasser E., Goldberg Y., « Simple and accurate dependency parsing using bidirectional LSTM feature representations », *Transactions of the Association for Computational Linguistics*, vol. 4, p. 313-327, 2016.

Kondratyuk D., Straka M., « 75 Languages, 1 Model : Parsing Universal Dependencies Universally », `http://arxiv.org/abs/1904.02099`, 2019.

Kruengkrai C., Torisawa K., Hashimoto C., Kloetzer J., Oh J.-H., Tanaka M., « Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks », *31st AAAI Conference on Artificial Intelligence*, 2017.

Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Kelcey M., Devlin J., Lee K., Toutanova K. N., Jones L., Chang M.-W., Dai A., Uszkoreit J., Le Q., Petrov S., « Natural Questions : a Benchmark for Question Answering Research », *Transactions of the Association of Computational Linguistics*, 2019.

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

Landis J. R., Koch G. G., « The measurement of observer agreement for categorical data », *Biometrics*, vol. 33, n° 1, p. 159-174, 1977.

Lawrence S., Giles C. L., « Context and page analysis for improved web search », *IEEE Internet computing*, vol. 2, n° 4, p. 38-46, 1998.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *LREC*, 2020.

Le J., Zhang C., Niu Z., « Answer Extraction Based on Merging Score Strategy of Hot Terms », *Chinese Journal of Electronics*, vol. 25, n° 4, p. 614-620, 2016.

Liu Y., Ott M., Goyal N., Du Jingfei adn Joshi M., Chen D., Lewis M., Zettlemoyer L., Stoyanov V., « RoBERTa : A Robustly Optimized BERT Pretraining Approach », https://arxiv.org/abs/1907.11692, 2019.

Lopez V., Uren V., Sabou M., Motta E., « Is question answering fit for the semantic web ? : a survey », *Semantic Web*, vol. 2, n° 2, p. 125-155, 2011.

Marneffe M.-C. d., Manning C. D., « The Stanford typed dependencies representation », *CoLing, Workshop on Cross-framework and Cross-domain Parser Evaluation*, p. 1-8, 2008.

Martin L., Muller B., Suárez P. J. O., Dupont Y., Romary L., Villemonte de la Clergerie É., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », https://arxiv.org/abs/1911.03894, 2019.

Miao Y., Blunsom P., « Language as a latent variable : Discrete generative models for sentence compression », *EMNLP*, 2016.

Milton F., « A correction : The use of ranks to avoid the assumption of normality implicit in the analysis of variance », *JASA*, vol. 34, n° 205, p. 109, 1939.

Nallapati R., Zhou B., dos Santos C., Gùlçehre Ç., Xiang B. *et al.*, « Abstractive text summarization using sequence-to-sequence RNNs and beyond », *CoNLL*, p. 280-290, 2016.

Nivre J., « An Efficient Algorithm for Projective Dependency Parsing », *IWPT*, Dublin, p. 149-160, 2003.

Nivre J., Fang C.-T., « Universal Dependency Evaluation », *in* M.-C. d. Marneffe, J. Nivre, S. Schuster (eds), *NoDaLiDa Workshop on Universal Dependencies*, Göteborg, p. 86-95, 2017.

Nivre J., Hall J., Nilsson J., « MaltParser : A Data-Driven Parser-Generator for Dependency Parsing », *LREC*, ELRA, Genoa, Italy, May, 2006.

Nivre J., Marneffe M.-C. d., Ginter F., Goldberg Y., Goldberg Y., Hajič J., D. M. C., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D., « Universal Dependencies v1 : A Multilingual Treebank Collection », *10th LREC*, Portorož, Slovenia, p. 23-38, 2016.

Nivre J., Marongiu P., Ginter F., Kanerva J., Montemagni S., Schuster S., Simi M., « Enhancing Universal Dependency Treebanks : A Case Study », *Workshop on Universal Dependencies (UDW 2018)*, ACL, Brussels, Belgium, p. 102-107, November, 2018.

Oepen S., Abend O., Abzianidze L., Bos J., Hajič J., Hershcovich D., Li B., O'Gorman T., Xue N., Zeman D. (eds), *CoNLL 2020 Shared Task : Cross-Framework Meaning Representation Parsing*, ACL, Online, November, 2020.

Oh J.-H., Torisawa K., Hashimoto C., Iida R., Tanaka M., Kloetzer J., « A semi-supervised learning approach to why-question answering », *AAAI*, 2016.

Oh J.-H., Torisawa K., Hashimoto C., Sano M., De Saeger S., Ohtake K., « Why-question answering using intra-and inter-sentential causal relations », *ACL 2013*, p. 1733-1743, 2013.

Pal V., Shrivastava M., Bhat I., « Answering Naturally : Factoid to Full length Answer Generation », *2nd Workshop on New Frontiers in Summarization*, Association for Computational Linguistics, Hong Kong, China, p. 1-9, November, 2019.

Radford A., Narasimhan K., Salimans T., Sutskever I., « Improving language understanding by generative pre-training », https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

Reiter E., Dale R., « Building applied natural-language generation systems », *Nat. Lang. Eng.*, vol. 3, n° 1, p. 57–87, 1997.

Rojas Barahona L. M., Bellec P., Besset B., Dos Santos M., Heinecke J., Asadullah M., Leblouch O., Lancien J.-Y., Damnati G., Mory E., Herlédan F., « Spoken Conversational Search for General Knowledge », *SIGdial*, ACL, Stockholm, p. 110-113, 2019.

Rush A. M., Chopra S., Weston J., « A neural attention model for abstractive sentence summarization », `https://arxiv.org/abs/1509.00685`, 2015.

Seddah D., Candito M., « Hard Time Parsing Questions : Building a QuestionBank for French », *10th LREC*, ELRA, Portorož, Slovenia, 2016.

See A., Liu P. J., Manning C. D., « Get to the point : Summarization with pointer-generator networks », `https://arxiv.org/abs/1704.04368`, 2017.

Sharp R., Surdeanu M., Jansen P., Clark P., Hammond M., « Creating causal embeddings for question answering with minimal supervision », *EMNLP*, 2016.

Shorten C., Khoshgoftaar T., « A survey on Image Data Augmentation for Deep Learning », *Journal of Big Data*, vol. 6, p. 1-48, 2019.

Straka M., « UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task », *CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, ACL, Brussels, p. 197-207, 2018.

Tan M., dos Santos C., Xiang B., Zhou B., « Improved representation learning for question answer matching », *ACL 2016*, p. 464-473, 2016.

Usbeck R., Ngomo A.-C. N., Haarmann B., Krithara A., Röder M., Napolitano G., « 7th Open Challenge on Question Answering over Linked Data (QALD-7) », *in* M. Dragoni, M. Solanki, E. Blomqvist (eds), *Semantic Web Challenges*, Springer International Publishing, Cham, p. 59-69, 2017.

Verberne S., van Halteren H., Theijssen D., Raaijmakers S., Boves L., « Learning to rank for why-question answering », *Information Retrieval*, vol. 14, n° 2, p. 107-132, 2011.

Wu M., Zheng X., Duan M., Liu T., Strzalkowski T., Albany S., « Question answering by pattern matching, web-proofing, semantic form proofing », *TREC*, p. 500-255, 2003.

Zayaraz G. *et al.*, « Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems », *Journal of King Saud University-Computer and Information Sciences*, vol. 27, n° 1, p. 13-24, 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 », *in* D. Zeman, J. Hajič (eds), *CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, ACL, Brussels, p. 1-21, 2018.

Zeman D., Popel M., *et al.*, « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », *CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, ACL, Vancouver, p. 1-19, 2017.

Étude des dépendances syntaxiques non projectives en français

Guy Perrier*

* *Université de Lorraine – LORIA – Campus Scientifique – BP 239 - 54506 Vandœuvre-lès-Nancy cedex - France*

guy.perrier@loria.fr

RÉSUMÉ. *Cet article présente les résultats d'une étude mathématique et linguistique des dépendances syntaxiques non projectives dans des corpus du français annotés selon deux formats de syntaxe en dépendance :* Universal Dependencies (UD) *et* Surface Syntactic Universal Dependencies (SUD). *Cette étude met en évidence le caractère très local des configurations de croisement de deux dépendances qui est une des façons de caractériser la non-projectivité. Elle met aussi en évidence quatre sources linguistiques principales de la non-projectivité : la montée de clitiques, l'extraction profonde, la minimisation de la longueur des dépendances et les couples de mots dépendants distants.*

MOTS-CLÉS : *dépendances non projectives, syntaxe en dépendances, détection de motifs dans un graphe.*

TITLE. *Study of non-projective dependencies in French*

ABSTRACT. *This paper presents the results of a mathematical and linguistic study of non-projective syntactic dependencies in French corpora annotated according to two dependency syntax formats:* Universal Dependencies (UD) *and* Surface Syntactic Universal Dependencies (SUD). *This study highlights the very local character of the configurations of two crossing dependencies, which is one of the ways to characterize non-projectivity. It also highlights four main linguistic sources of non-projectivity : clitic climbing, deep extraction, dependency length minimization and pairs of distant dependent words .*

KEYWORDS: *non projective dependencies, dependency syntax, graph pattern matching.*

1. Introduction

Les dépendances non projectives posent problème tant aux linguistes qu'aux informaticiens. Elles compliquent pour les premiers la tâche de conversion des structures en dépendances en structures syntagmatiques (Gaifman, 1965), car la non-projectivité s'oppose à la continuité des syntagmes. La discontinuité dans les structures linguistiques non projectives est aussi un obstacle à leur compréhension, si bien que les dépendances non projectives sont aussi intéressantes à étudier pour les psycholinguistes. Pour les informaticiens, elle complique beaucoup les algorithmes d'analyse syntaxique, même si certains ont développé des méthodes qui permettent de réduire cette complexité (Tapanainen et Jarvinen, 1997 ; McDonald *et al.*, 2005 ; Nivre, 2009 ; Gómez-Rodríguez *et al.*, 2014 ; Straka *et al.*, 2015).

Pour traiter les dépendances non projectives sous ces différents aspects, il est utile de bien les connaître et donc de les étudier plus précisément. Or, s'il existe beaucoup de travaux sur l'analyse syntaxique des arbres de dépendances non projectifs, il existe peu d'études de ces arbres dans les corpus réels tant d'un point de vue mathématique que linguistique. Quelques études ont été menées pour des langues particulières, comme le tchèque (Hajicová *et al.*, 2004), le grec ancien (Mambrini et Passarotti, 2013) et le serbe (Miletic et Urieli, 2017) notamment. La principale étude mathématique générale, à notre connaissance, est la thèse de Havelka (2007) qui exhibe différentes caractérisations des arbres de dépendances non projectifs, pour en déduire des algorithmes de recherche de ceux-ci dans un corpus arboré, ce qui lui permet de comparer la fréquence d'apparition des dépendances non projectives entre 19 langues. Parmi ces langues, il n'y a pas le français.

Pour le français, il est à mentionner une étude de Botalla (2014), qui aborde la non-projectivité à travers l'analyse du flux des dépendances dans un treebank, et une autre de Béchet et Lacroix (2015) sur le corpus, *CDGFr*. Ce dernier ayant été annoté de façon non standard, il est difficile d'établir une comparaison avec d'autres travaux.

Dans cet article, nous présentons une étude que nous avons menée sur les dépendances non projectives en français à partir de corpus annotés. Nous avons effectué cette étude en deux temps. Dans un premier temps, nous avons considéré uniquement l'aspect topologique des annotations, c'est-à-dire la structure formée par les dépendances non projectives, indépendamment de leurs étiquettes linguistiques. Le but était de mesurer le plus précisément possible le caractère local et complexe de la non-projectivité. Parmi toutes les façons de caractériser la non-projectivité, nous avons privilégié le croisement de dépendances. Nous avons pu exhiber un nombre limité de motifs finis formés par les couples de dépendances qui se croisent.

Ces motifs ont constitué le point de départ du second temps, l'étude linguistique, où nous avons alors pris en compte les étiquettes linguistiques des dépendances. Nous avons exploré systématiquement les corpus à la recherche de toutes les occurrences des motifs mis en évidence par l'étude topologique. Nous avons utilisé pour cela l'ou-

til GREW-MATCH[1] qui procède par appariement de graphes. C'est une composante d'un outil plus vaste, GREW[2] utilisé pour la transformation d'annotations et fondé sur la réécriture de graphes à l'aide de règles (Bonfante *et al.*, 2018). Partant de motifs purement structurels, GREW-MATCH permet par l'exploration des corpus de les enrichir linguistiquement pas à pas pour faire apparaître les principaux phénomènes linguistiques responsables sur les corpus donnés de non-projectivité.

Cette étude présente un intérêt tant pour l'analyse syntaxique en dépendances que pour l'annotation de corpus et pour l'analyse linguistique de corpus. Tout d'abord, on sait que la non-projectivité complique beaucoup la tâche d'analyse syntaxique. Même si des progrès importants ont été réalisés ces dernières années, les meilleurs analyseurs sont encore loin de la perfection : Kuhlmann et Nivre (2010) n'analysent correctement que la moitié des dépendances non projectives présentes dans des corpus de l'anglais et de l'allemand et la proportion monte aux deux tiers pour le tchèque.

L'étude topologique vise à obtenir des résultats quantitatifs en termes de fréquence, de localité et de complexité des dépendances non projectives qui permettront d'adapter les algorithmes d'analyse afin de concilier efficacité et expressivité. Cela peut se faire tant dans l'approche fondée sur la recherche d'arbres à portée maximale dans un graphe en suivant les travaux de Corro *et al.* (2016) que dans l'approche fondée sur des systèmes de transition comme le montrent Kuhlmann et Nivre (2006). Pour ce qui est de cette dernière approche, la mise en évidence des phénomènes linguistiques sources de non-projectivité, qui est visée dans la partie linguistique de notre étude, permettra de les confronter avec les règles de transition des systèmes utilisés pour étudier dans quelle mesure ces règles peuvent les prendre en compte. D'ailleurs, Kuhlmann et Nivre (2010) insistent sur cet aspect dans la conclusion de leur article : *« Although the experiments presented in this article have already revealed significant differences both between languages and between techniques, it would be interesting to look in more detail at the different linguistic constructions that give rise to non-projective dependencies.(Bien que les expériences présentées dans cet article aient déjà révélé des différences significatives entre les langues et entre les techniques, il serait intéressant d'examiner plus en détail les différentes constructions linguistiques qui donnent lieu à des dépendances non projectives) ».*

Maintenant, plutôt que de chercher à améliorer l'analyse syntaxique, il peut être plus facile et plus efficace de chercher après coup à corriger les dépendances non projectives mal annotées. Depuis plusieurs années, nous avons développé une expertise dans ce domaine en utilisant l'outil GREW-MATCH pour repérer les constructions erronées dans un treebank et l'outil GREW pour corriger les erreurs quand elles sont systématiques (Guillaume *et al.*, 2019). L'étude que nous présentons ici va permettre de dégager des motifs associés aux phénomènes linguistiques responsables de non-projectivité. En appliquant ces motifs à des corpus du français, il est possible tout d'abord de détecter les constructions faussement non projectives. Si le format d'anno-

1. `http://match.grew.fr/`
2. `http://match.grew.fr/`

tation n'est pas celui utilisé dans notre étude, cela exigera quelques adaptations. Pour les constructions faussement projectives, les choses sont un peu plus compliquées car il faudra recenser les erreurs possibles pour les traduire sous forme de motifs.

Une autre application concerne la transformation d'un corpus annoté syntaxiquement en constituants en un corpus annoté en dépendances. Il existe un algorithme classique qui permet de le faire mais le résultat est nécessairement un ensemble d'arbres projectifs. Pour faire apparaître la non-projectivité, il est nécessaire de compléter l'application de cet algorithme par une transformation de certaines dépendances projectives en dépendances non projectives. Candito *et al.* (2009) l'ont fait manuellement pour le FRENCH TREEBANK. Notre étude serait un point de départ à la détermination de règles de réécriture qui permettraient de le faire automatiquement.

Enfin, notre travail peut aider à une étude linguistique plus poussée de la non-projectivité. Il serait tout d'abord intéressant de confronter les phénomènes linguistiques sources de la non-projectivité à différentes théories linguistiques pour étudier dans quelle mesure ces dernières sont capables de fournir une explication. Et puis la méthode que nous présentons est suffisamment simple pour être utilisée par des linguistes qui voudraient faire une étude approfondie de la non-projectivité sur d'autres corpus que ceux que nous avons choisis.

Pour notre étude, nous avons choisi trois treebanks :

– UD_FRENCH-GSD[3] (Guillaume *et al.*, 2019) dont les données proviennent de l'Universal Dependency Treebank v2.0 de Google (McDonald *et al.*, 2013) ; il comprend 16 341 phrases d'origines très diverses (dépêches de presse, blogs, avis de consommateurs. . .) ; à partir de 2015, l'annotation du corpus a été convertie dans le format UD, sur lequel est fondé le projet Universal Dependencies[4] ; ce projet a pour but de créer un schéma d'annotation syntaxique unique qui puisse être utilisé pour un maximum de langues différentes (Nivre *et al.*, 2016) ; c'est pour cette raison que le format UD considère les relations syntaxiques comme des relations directes entre mots lexicaux, et les mots fonctionnels comme des marqueurs des mots lexicaux ;

– SUD_FRENCH-GSD [5] qui résulte d'une conversion du corpus précédent dans un nouveau format, le format SUD (Gerdes *et al.*, 2018 ; Gerdes *et al.*, 2019). SUD est une alternative à UD qui utilise de façon plus classique des critères distributionnels pour définir les relations syntaxiques, si bien que les têtes des relations sont plutôt les mots fonctionnels que les mots lexicaux ;

– UD_FRENCH-SEQUOIA[6] est issu du corpus SEQUOIA qui a d'abord été annoté en syntagmes selon le schéma du FRENCH TREEBANK (Abeillé *et al.*, 2019), puis converti en dépendances (Candito et Seddah, 2012b) ; cette annotation en dépendances a été enfin convertie dans le format UD (Guillaume *et al.*, 2019) ; le corpus comprend

3. https://github.com/UniversalDependencies/UD_French-GSD
4. http://universaldependencies.org
5. https://github.com/surfacesyntacticud/SUD_French-GSD
6. https://github.com/UniversalDependencies/UD_French-Sequoia

3 099 phrases de quatre origines différentes : l'agence européenne du médicament, Europarl, le journal régional l'*Est Républicain* et Wikipedia Fr.

Notre étude a porté sur la version 2.7 de ces trois treebanks. L'intérêt d'avoir un même corpus annoté selon deux formats différents, UD_FRENCH-GSD selon UD et SUD_FRENCH-GSD annoté selon SUD, est de pouvoir étudier dans quelle mesure le format d'annotation rend compte de cette propriété de non-projectivité.

L'intérêt d'avoir deux corpus différents, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, annotés dans un même format, UD, est lui de pouvoir étudier dans quelle mesure les dépendances non projectives dépendent du choix du corpus.

Le plan de l'article est le suivant :

– dans la section 2, nous présenterons les formats d'annotation syntaxique UD et SUD en mettant en évidence leurs différences ;

– dans la section 3, nous nous intéresserons à la topologie des configurations formées par les couples de dépendances qui se croisent, mettant en évidence que ces configurations ont un caractère très local ;

– enfin dans la section 4, nous montrerons que principalement quatre phénomènes linguistiques sont principalement source de dépendances non projectives dans les corpus étudiés : la montée de clitiques, l'extraction profonde, la minimisation de la longueur des dépendances et les couples de mots dépendants distants.

2. Les formats d'annotation syntaxique UD et SUD

Comme nous le verrons par la suite, le format d'annotation syntaxique joue un rôle important dans l'existence ou non de dépendances non projectives. C'est pourquoi il est nécessaire d'avoir un aperçu des deux formats utilisés dans les trois corpus étudiés.

2.1. *Le format* UD

La définition du format UD (Nivre *et al.*, 2016)[7] a été guidée par le souci de son universalité, c'est-à-dire qu'il puisse être utilisé pour toutes les langues. C'est pourquoi il est guidé par la sémantique. Les têtes des syntagmes sont les mots lexicaux, les mots fonctionnels étant rattachés à ceux-ci comme marqueurs. Les relations de dépendances syntaxiques sont donc des relations entre mots lexicaux.

Une caractéristique de UD, qui n'était pas requise par le souci d'universalité, est que les types des relations renvoient non seulement aux fonctions syntaxiques des mots, mais aussi à leurs parties du discours. Ainsi, la relation entre un modificateur d'un nom et ce nom peut être étiquetée `acl`, `advmod`, `amod`, `nmod`, selon que le modificateur est une proposition, un adverbe, un adjectif ou un nom.

7. `https://universaldependencies.org/guidelines.html`

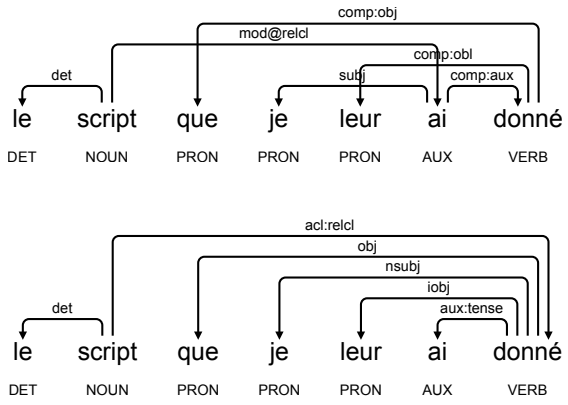

Figure 1. *Annotation de la même expression dans* SUD *au-dessus et* UD *au-dessous*

Le schéma du bas de la figure 1 présente un exemple significatif d'une expression annotée dans UD. L'annotation de la phrase fait apparaître un arbre de dépendances très plat avec parfois un nombre important de dépendances se rattachant à un même mot lexical. Ainsi, le mot *donné* a quatre dépendants qui ne sont pas du tout reliés les uns aux autres ; ils sont simplement ordonnés.

2.2. *Le format* SUD

Le format SUD (Gerdes *et al.*, 2018 ; Gerdes *et al.*, 2019)[8] se présente comme une alternative au format UD fondée sur l'approche distributionnelle plus classique des relations syntaxiques (Bloomfield, 1933 ; Mel'cuk *et al.*, 1988 ; Kahane et Gerdes, 2020). Dans cette approche, les mots fonctionnels constituent les têtes des syntagmes dans la mesure où ils déterminent leur distribution. Dans SUD, il s'agit des prépositions, des conjonctions de subordination et des auxiliaires. Les conjonctions de coordination et les déterminants, qui jouent un rôle plus discutable dans la distribution des syntagmes qu'ils introduisent, ne sont pas leur tête dans SUD.

Par ailleurs, comme la partie du discours des dépendants n'est pas déterminante dans la distribution des relations, les étiquettes de ces dernières ne la prennent pas en compte et elles ne considèrent que les fonctions syntaxiques.

Enfin, les relations sont organisées en une stricte taxonomie. Un moyen de créer une sous-relation d'une autre est d'ajouter une extension à son nom précédée d'un

8. `https://surfacesyntacticud.github.io/`

deux-points. Ainsi, `comp` représente la fonction argument syntaxique complément en général et `comp:aux` représente la fonction argument d'un auxiliaire.

Le schéma du haut de la figure 1 reprend la même phrase que pour illustrer UD et montre son annotation dans SUD. La différence avec UD est flagrante : les mots fonctionnels sont structurés les uns par rapport aux autres, ce qui augmente la profondeur des arbres de dépendances. Comme l'exemple le montre aussi, cela augmente aussi la possibilité d'avoir des dépendances non projectives.

3. Étude topologique des dépendances non projectives en français

Dans cette section, nous nous intéressons aux configurations structurelles formées par les dépendances non projectives indépendamment de leur typage linguistique, effectué en général par l'association des dépendances à des étiquettes représentant des fonctions syntaxiques (sujet, objet...). Dans toute la suite de l'étude, nous ignorerons les dépendances qui ciblent des signes de ponctuation, dans la mesure où les guides d'annotation de UD et de SUD sont trop imprécis sur comment choisir leurs gouverneurs, d'où beaucoup d'incohérences dans les corpus existants.

3.1. *Caractérisation mathématique de la non-projectivité*

Les structures représentant la syntaxe des phrases considérées dans cet article sont des arbres de dépendances totalement ordonnés. Un *arbre de dépendances totalement ordonné* est un arbre enraciné[9] dont les nœuds sont totalement ordonnés. Par la suite, les arbres de dépendances que nous utiliserons sont toujours totalement ordonnés et nous ne le mentionnerons pas à chaque fois. Les nœuds représentent les mots de la phrase [10]. La relation père-fils dans l'arbre représente la dépendance syntaxique entre les mots ; elle est notée $\rightarrow$. Sa clôture transitive est notée $\rightarrow^+$et sa clôture transitive et réflexive $\rightarrow^*$. Une instance $\texttt{H} \rightarrow \texttt{D}$ de la relation $\rightarrow$ est appelée *une dépendance*.

La relation d'ordre total entre les nœuds est notée $\leq$ quand elle est entendue au sens large et $<$ quand elle est entendue au sens strict. Elle représente l'ordre des mots dans la phrase.

Historiquement, la distinction entre projectivité et non-projectivité d'un arbre de dépendance totalement ordonné a été mise en évidence par Harper et Hays (1959) et

9. Un arbre enraciné est un graphe connexe acyclique dont on a choisi un nœud particulier comme racine.

10. Dire que les nœuds représentent les mots de la phrase est une simplification car en général, les tokens résultant du découpage de la phrase ne coïncident pas toujours avec les mots de la phrase au sens linguistique. Certains mots peuvent être formés de plusieurs tokens, ce qui est indiqué dans l'arbre syntaxique par des relations de dépendances spécifiques. Ces relations n'étant jamais responsables de non-projectivité, on peut considérer sans perte de généralité que la notion de token coïncide avec celle de mot.

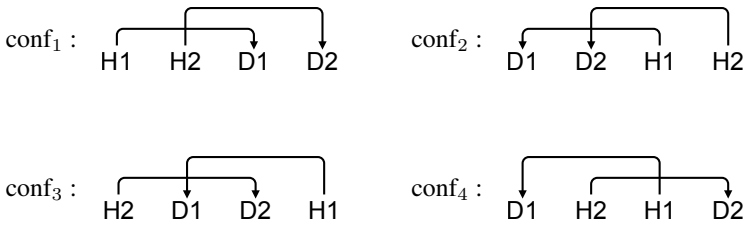

Figure 2. *Les quatre configurations possibles de croisement de deux dépendances*

Lecerf et Ihm (1960). Elle a été étudiée par Marcus (1965) qui a démontré l'équivalence entre trois caractérisations d'un arbre de dépendances projectif :

– chacun de ses nœuds `N` est projectif, c'est-à-dire que l'ensemble des nœuds `M` tels que `N` $\rightarrow^*$`M` forme un segment continu selon l'ordre des mots de la phrase (Fitialov, 1962) ; le nœud *donné* de l'arbre SUD de la figure 1 est non projectif car sa projection { *que, leur, donné* } est discontinue ;

– chacune de leurs dépendances `H` $\rightarrow$ `D` est projective, c'est-à-dire que l'ensemble des nœuds situés entre `H` et `D` est inclus dans la projection de `H` (Harper et Hays, 1959) ; l'arbre SUD de la figure 1 comporte deux dépendances non projectives : (`donné -[comp:obj]`$\rightarrow$ `que`) et (`donné -[comp:obl]`$\rightarrow$ `leur`) ;

– les dépendances de l'arbre complété ne se croisent jamais selon l'ordre de la phrase ; l'arbre complété est l'arbre de racine `R` auquel a été ajoutée une dépendance `R'` $\rightarrow$ `R` issue d'une racine fictive `R'` ajoutée à gauche de la phrase [11] ; l'arbre SUD de la figure 1 comporte trois croisements de dépendances.

L'inconvénient de la première caractérisation est qu'elle nécessite de calculer la projection complète des nœuds. La seconde est plus locale puisqu'elle demande seulement à vérifier que tous les nœuds situés entre chaque nœud et chacun de ses dépendants sont dans la projection de ce nœud. La dernière est encore plus facilement calculable car il s'agit de vérifier que chaque paire de dépendances ne donne pas lieu à un croisement de celles-ci, et c'est cette caractérisation que nous avons privilégiée dans notre étude. La figure 2 traduit graphiquement les quatre configurations possibles de croisement de deux dépendances $\mathtt{H}_1 \rightarrow \mathtt{D}_1$ et $\mathtt{H}_2 \rightarrow \mathtt{D}_2$. Les trois croisements de l'arbre SUD de la figure 1 illustrent les configurations 2 et 3.

11. On pourrait tout aussi bien ajouter `R'` à droite de la phrase et Kahane et Gerdes (2020) évitent cet arbitraire en plaçant les mots de la phrase sur un cercle et en ajoutant le nœud `R'` entre le premier et le dernier mot de la phrase.

3.2. *Mesure du degré de non-projectivité d'un corpus*

Dans un souci de concevoir des algorithmes d'analyse syntaxique qui respectent un équilibre entre efficacité et pouvoir d'expression, il est utile de connaître la proportion de nœuds non projectifs ou de dépendances non projectives dans un corpus.

On peut aller plus loin en introduisant des paramètres qui mesurent le degré de complexité des nœuds non projectifs et des dépendances non projectives, de telle façon que si ce degré ne dépasse pas une certaine borne, on puisse concevoir des algorithmes d'analyse relativement efficaces. C'est ce qu'ont cherché à faire Kuhlmann et Nivre (2006) et Corro *et al.* (2016). Nous allons reprendre certains de ces paramètres en les appliquant à nos corpus.

Un premier paramètre est le nombre de trous maximal des projections des nœuds d'un arbre de dépendances.

3.2.1. *Nombre maximal de trous dans les projections des nœuds d'un arbre de dépendances*

La notion de trou a été introduite par Holan *et al.* (2000) pour caractériser la complexité des arbres de dépendances non projectifs. Donnons-en une définition formelle.

Définition 1. *Un trou dans la projection d'un nœud* `N` *est un ensemble de mots consécutifs qui n'appartiennent pas à la projection de* `N` *et qui est borné à droite et à gauche par deux éléments de cette projection.*

De cette définition, découle immédiatement que tout nœud non projectif est caractérisé par la présence d'un ou plusieurs trous dans sa projection. Ainsi, dans l'annotation SUD de la figure 1, le nœud non projectif *donné* a sa projection [*que, leur, donné*] qui comporte deux trous {*je*} et {*ai*}. Dans l'annotation UD, la projection [*que, je, leur, ai, donné*] du nœud projectif *donné* ne comporte pas de trou.

Nous avons conçu un programme Python qui, pour un corpus donné, calcule le nombre de trous pour la projection de chaque nœud des arbres syntaxiques et, pour chaque arbre, détermine le nombre de trous maximal. Les nœuds projectifs sont ceux dont la projection ne comporte pas de trou et les arbres projectifs sont ceux pour lesquels le maximum du nombre de trous est de 0.

Sur nos trois corpus, nous obtenons les résultats du tableau 1. La première constatation est que selon le format d'annotation, les résultats sont très différents. Les textes des corpus SUD_FRENCH-GSD et UD_FRENCH-GSD sont les mêmes, découpés de la même façon en tokens. Seul diffère le format d'annotation, SUD pour le premier et UD pour le second. Dans SUD_FRENCH-GSD, il y a deux fois et demie plus de nœuds non projectifs que dans UD_FRENCH-GSD. L'explication tient au fait que les arbres de dépendances UD sont beaucoup plus plats que les arbres SUD. Les têtes des syntagmes sont toujours des mots lexicaux et les mots grammaticaux se rattachent directement à ces têtes. Dans SUD au contraire, les têtes des syntagmes sont les mots fonctionnels quand ils sont présents (auxiliaires, prépositions, conjonctions de subor-

corpus	SUD-GSD	UD-GSD	UD-SEQUOIA	total
nb. de nœuds	356 393	356 393	62 706	775 492
nb. de nœuds non projectifs	1729 (0,49 %)	694 (0,19 %)	74 (0,12 %)	2 497 (0,32 %)
nœuds à 0 trou	354 664	355 699	62 632	772 995
nœuds à 1 trou	1 708	692	73	2 473
nœuds à 2 trous	21	2	1	24
nb. d'arbres syntaxiques	16 341	16 341	3 099	35 781
nb. d'arbres non projectifs	1 392 (8,52 %)	653 (2,62 %)	66 (2,13 %)	2 111 (5,90 %)
arbres à 0 trou au maximum	14 949	15 688	3 033	33 670
arbres à 1 trou au maximum	1 373	651	65	2 089
arbres à 2 trous au maximum	19	1	1	22

Tableau 1. *Statistiques sur le nombre de trous dans les projections des nœuds*

dination), et ces mots peuvent être dépendants les uns des autres d'où une structure plus en profondeur des arbres, comme le montre l'exemple de la figure 1.

La seconde constatation est que le maximum du nombre de trous par projection est très bas puisqu'il est de 2 pour les trois corpus considérés. Encore plus intéressant, considérons la répartition des nœuds des arbres d'un corpus selon le nombre de trous par projection. Pour les trois corpus considérés ensemble, on obtient la répartition [772 995, 2 473,24] correspondant à 0 trou, 1 trou, 2 trous. Si on considère les arbres et non plus les nœuds, la répartition est [33 670, 2 089, 22]. La conséquence est que pour avoir des algorithmes d'analyse efficaces, on peut limiter le nombre de trous maximal à 1 avec une perte négligeable de pouvoir expressif.

3.2.2. *Nombre maximal de composantes connexes de trous*

Ce paramètre a été introduit par Nivre (2006), toujours dans un but d'augmenter la performance des algorithmes d'analyse sans réduire trop le pouvoir d'expression.

Définition 2. *Une composante connexe de trous d'une dépendance* `H` $\rightarrow$ `D` *est un ensemble maximal de nœuds situés entre* `H` *et* `D` *qui ne sont pas dans la projection de* `H` *et qui sont connectés les uns aux autres.*

Pour l'annotation SUD de la figure 1, la dépendance (`donné -[comp=obj]`$\rightarrow$ `que`) comporte deux trous {*je*} et {*ai*} mais une seule composante connexe de trous car *je* et *ai* sont liés par une dépendance.

Nous avons conçu un programme Python qui, pour chaque dépendance d'un corpus, détermine la liste de ses composantes connexes. Ces composantes sont identifiées

corpus	SUD-GSD	UD-GSD	UD-SEQUOIA	total
nb. de dépendances	356 393	356 393	62 706	775 492
nb. de dépendances non projectives	1 547 (0,47 %)	679 (0,19 %)	73 (0,12 %)	1 924 (0,25 %)
dépendances à 0 composante connexe de trous	354 846	355 714	62 633	773 193
dépendances à 1 composante connexe de trous	1 531	667	71	2 269
dépendances à 2 composantes connexes de trous	15	11	2	28
dépendances à 3 composantes connexes de trous	1	1	0	2
nb. d'arbres syntaxiques	16 341	16 341	3 099	35 781
nb. d'arbres non projectifs	1 392 (8,52 %)	653 (4,00 %)	66 (2,13 %)	2 111 (5,90 %)
arbres à 0 composante connexe de trous max.	14 949	15 688	3 033	33 670
arbres à 1 composante connexe de trous max.	1 376	641	64	2 081
arbres à 2 composantes connexes de trous max.	15	11	2	28
arbres à 3 composantes connexes de trous max.	1	1	0	2

Tableau 2. *Statistiques sur le nombre de composantes connexes de trous par dépendances*

par leurs racines. Ces racines sont faciles à déterminer car elles se caractérisent comme un nœud d'un trou dont le gouverneur est extérieur à la dépendance considérée.

Le tableau 2 récapitule les résultats trouvés sur nos trois corpus. Il montre qu'on peut limiter le nombre maximal de composantes connexes à 1. Nivre (2006) a établi expérimentalement sur les treebanks DANISH DEPENDENCY TREEBANK et PRAGUE DEPENDENCY TREEBANK qu'avec cette limite, on peut obtenir des algorithmes d'analyse linéaires en temps, qui excluent moins de 2 % des solutions. Il reste à faire le même type de mesure sur nos corpus du français.

3.3. *La non-projectivité, un phénomène local*

Considérons maintenant la caractérisation de la non-projectivité comme croisement de deux dépendances. Cette configuration n'est pas à proprement parler locale au sens où elle constituerait un graphe fini connexe, quand on ne considère que les dépendances syntaxiques : elle ne fait apparaître aucun lien entre les deux dépendances qui se croisent. Or, il faut se rappeler que ces deux dépendances font partie d'un même arbre donc les chemins qui mènent de $\mathtt{D}_1$ et de $\mathtt{D}_2$ à la racine de l'arbre se rencontrent

Chaîne reliant les deux dépendances	**DIST(H, D_1)**	**DIST(H, D_2)**	**dépendance non projective**
H_1 $^{+}\leftarrow$ H $\rightarrow^{+}H_2$	> 1	> 1	$H_1 \rightarrow D_1$, $H_2 \rightarrow D_2$
$H_1 \rightarrow^{+} H_2$	1	> 1	$H_2 \rightarrow D_2$
$D_1 \rightarrow^{+} H_2$	0	> 1	$H_2 \rightarrow D_2$
$H_2 \rightarrow^{+} H_1$	> 1	1	$H_1 \rightarrow D_1$
$D_2 \rightarrow^{+} H_1$	> 1	0	$H_1 \rightarrow D_1$

Tableau 3. *Les cinq configurations formées par deux dépendances* $H_1 \rightarrow D_1$ *et* $H_2 \rightarrow D_2$ *qui se croisent*

nécessairement en un nœud que nous noterons H. Pour déterminer dans quelle mesure la non-projectivité est un phénomène local, il est intéressant d'étudier les distances de D_1 et de D_2 à H et de voir à quel point elles sont bornées. Dans un arbre, la distance d'un nœud N à un de ces ancêtres A, notée DIST(A, N), est le nombre de dépendances formant le chemin de A à N.

Pour mesurer le caractère local d'une dépendance non projective, Havelka (2007) définit une distance qui a un rapport direct avec les distances DIST(H, D_1) et DIST(H, D_2). Sa distance n'est pas attachée à une paire de dépendances qui se croisent mais à une dépendance non projective. Considérons-en une quelconque $H_1 \rightarrow D_1$. Havelka définit le *type de niveau* de cette dépendance comme étant le maximum de DIST(H, D_1) − DIST(H, D_2), pour $H_2 \rightarrow D_2$ étant une dépendance qui croise la première, et H étant le premier ancêtre commun à D_1 et D_2. L'objectif de Havelka est de concevoir des algorithmes efficaces de détermination de dépendances non projectives alors que le nôtre est d'exhiber des configurations locales de dépendances qui se croisent.

Selon les positions de D_1 et de D_2 par rapport à H, il y a cinq configurations possibles. Pour chacune d'elles, nous indiquons la forme de la chaîne reliant les deux dépendances qui se croisent et les valeurs possibles de DIST(H, D_1) et de DIST(H, D_2). On peut même déterminer la ou les dépendances responsables de la non-projectivité. Le tableau 3 décrit ces cinq configurations, en indiquant laquelle des deux dépendances concernées est nécessairement non projective[12]. Théoriquement, la distance DIST(H, D_1) ou DIST(H, D_2) est non bornée. Nous nous proposons de voir ce qu'il en est sur corpus. Pour cela, nous avons conçu un programme qui prend en entrée un corpus annoté en dépendances syntaxiques dans un format *conll* et qui retourne un fichier contenant toutes les dépendances qui se croisent en indiquant pour chaque croisement la valeur de DIST(H, D_1) et celle de DIST(H, D_2). En plus, sont affichées certaines statistiques sur le corpus relatives à la non-projectivité.

On a appliqué le programme aux trois corpus du français SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA. Le tableau 4 en récapitule les résultats. Le rôle déterminant du format est confirmé par ces statistiques : il y a 2,6 fois plus de croisements dans SUD_FRENCH-GSD que dans UD_FRENCH-GSD.

12. Nous rappelons que la relation $\rightarrow^{+}$ est la clôture transitive de la relation de dépendance.

corpus	**SUD-GSD**	**UD-GSD**	**UD-SEQUOIA**	**total**
nb. de tokens	416 740	416 740	73 666	907 146
nb. de croisements	3 283 (0,79 %)	1 242 (0,30 %)	107 (0,15 %)	4 632 (0,51 %)
croisements avec $\mathtt{H}_1 \ ^+\!\leftarrow \mathtt{H} \rightarrow^+\!\mathtt{H}_2$	5	3	0	8
croisements avec $\mathtt{H}_1 \rightarrow^+\!\mathtt{H}_2$	1 205	609	42	1 856
croisements avec $\mathtt{D}_1 \rightarrow^+\!\mathtt{H}_2$	282	283	5	570
croisements avec $\mathtt{H}_2 \rightarrow^+\!\mathtt{H}_1$	716	114	13	743
croisements avec $\mathtt{D}_2 \rightarrow^+\!\mathtt{H}_1$	1 075	233	47	1 355
maximum de $\textsc{dist}(\mathtt{H}, \mathtt{D}_1)$	5	3	3	5
maximum de $\textsc{dist}(\mathtt{H}, \mathtt{D}_2)$	5	4	3	5

Tableau 4. *Récapitulation des croisements pour trois corpus du français*

distance à H	0	1	2	3	4	5
distribution de $\textsc{dist}(\mathtt{H}, \mathtt{D}_1)$	570	1 856	2 028	164	13	1
distribution de $\textsc{dist}(\mathtt{H}, \mathtt{D}_2)$	1 355	843	2 264	150	13	5

Tableau 5. *Répartition des dépendances qui se croisent par distance à l'ancêtre commun* H

Une seconde constatation est que la configuration $\mathtt{H}_1 \ ^+\!\leftarrow \mathtt{H} \rightarrow^+\!\mathtt{H}_2$ est très rare. Parmi les 5 configurations possibles, c'est la seule qui donne lieu à une imbrication de sous-arbres disjoints. On peut donc avec une perte négligeable d'expressivité utiliser des algorithmes d'analyse qui prennent en compte la non-imbrication de sous-arbres disjoints pour plus d'efficacité (Kuhlmann et Nivre, 2006 ; Corro *et al.*, 2016).

Venons-en maintenant à la question principale qui nous préoccupe : dans quelle mesure la configuration de deux dépendances qui se croisent est-elle locale ? Pour y répondre, il faut examiner les limites supérieures de $\textsc{dist}(\mathtt{H}, \mathtt{D}_1)$ et de $\textsc{dist}(\mathtt{H}, \mathtt{D}_2)$. Sur les trois corpus étudiés, ces limites sont toutes les deux 5, mais en plus, elles sont rarement atteintes. La tableau 5 est révélateur à ce sujet.

Il existe seulement six phrases pour lesquelles la limite 5 de $\textsc{dist}(\mathtt{H}, \mathtt{D}_1)$ ou de $\textsc{dist}(\mathtt{H}, \mathtt{D}_2)$ est atteinte. Elles sont dans le corpus SUD_FRENCH-GSD.

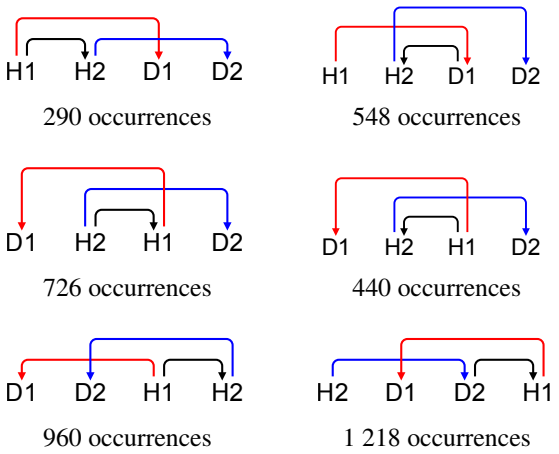

Figure 3. *Les six motifs de croisement de dépendances les plus fréquents*

Détaillons ce qu'il en est sur l'une d'elles, la phrase *fr-ud-train_02164*.

> *Ce sont là, avec d'autres, des actes clairement hostiles à lesquels (auxquels) il n'a pas été jugé utile de répondre pour l'instant.*

Les deux dépendances qui se croisent sont *actes* $\rightarrow$ *a* et *à* $\leftarrow$ *répondre*. Elles sont liées par le chemin *a* $\rightarrow$ *été* $\rightarrow$ *jugé* $\rightarrow$ *de* $\rightarrow$ *répondre* $\rightarrow$ *à*. Ce chemin est bien de longueur 5. On pourrait penser que sa longueur exceptionnelle rend la compréhension de la phrase difficile mais on peut remarquer que ce chemin suit l'ordre de lecture de la phrase, ce qui explique peut-être que la phrase est aisément compréhensible.

En nous limitant à des motifs où la chaîne qui va de $\mathtt{D}_1$ à $\mathtt{D}_2$ a une longueur au maximum de 4, on couvre 4 611 des 4 632 croisements de dépendances. Ces motifs sont au nombre de 11 et si on les combine avec les quatre façons d'ordonner les deux dépendances qui se croisent, on obtient 44 motifs. Sur ces 44 motifs, il y en a seulement 21 dont on rencontre des occurrences dans au moins un des trois corpus et parmi ces 21, il n' y en a que 11 qui ont au moins 49 occurrences, tous les autres ont moins de 15 occurrences. Enfin sur ces 11, 6 se dégagent nettement, couvrant 4 144 croisements sur 4 632. La figure 3 présente ces six motifs[13].

Même si l'étude que nous avons menée est dépendante des trois corpus sur lesquels elle a porté, la conclusion est que la non-projectivité en français est un phénomène très local. Nous sélectionnons les 11 motifs les plus fréquents qui vont nous permettre de pousser plus à fond l'étude en considérant maintenant la dimension linguistique.

13. Ces six motifs seront illustrés par des exemples dans la partie 4 et dans chaque exemple, pour le croisement concerné, la dépendance nœuds $\mathtt{H}_1 \rightarrow \mathtt{D}_1$ sera en rouge et la dépendance $\mathtt{H}_2 \rightarrow \mathtt{D}_2$ en bleu.

4. Étude linguistique des dépendances non projectives en français

Il s'agit maintenant de réintégrer les étiquettes des fonctions syntaxiques dans les dépendances formant les 11 motifs qui ont été exhibés dans la section précédente, afin de dégager les sources linguistiques de la non-projectivité en français. Nous commencerons par présenter l'outil informatique GREW-MATCH que nous avons utilisé pour notre étude puis nous donnerons les résultats de l'étude elle-même.

4.1. *L'outil de recherche automatique de motifs dans un graphe* GREW-MATCH

GREW-MATCH[14] est un outil qui permet de retrouver dans un graphe toutes les occurrences d'un motif donné.

```
pattern{ H1 -> D1; H2 -> D2; D2 -> H1;
         H2 << D1; D1 << D2; D2 << H1 }
without{ D1[upos=PUNCT] }
without{ D2-[1= comp]-> H1; D2[upos=AUX|VERB]; D1[upos=PRON] }
without{ H2 -[mod@relcl]-> D2; D1[PronType=Rel] }
without{ H2 -[mod@relcl]-> D2; D1 -> P; P[PronType=Rel] }
without{ H2 -[mod@relcl]-> D2; D1 -> D; D -> P; P[PronType=Rel] }
```

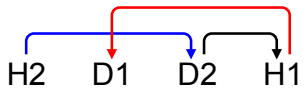

Figure 4. *Exemple de motif* GREW

La figure 4 présente un motif exprimant une configuration de deux dépendances $H_1 \rightarrow D_1$ et $H_2 \rightarrow D_2$ qui se croisent. Au-dessus, vous avez la définition du motif dans la syntaxe de GREW et au-dessous la traduction graphique de sa partie *pattern*. Le mot-clé *pattern* permet de décrire le motif sous forme d'une suite de déclarations et contraintes élémentaires séparées par un point-virgule. La première ligne contient les déclarations de trois dépendances et la suivante exprime des contraintes d'ordre entre les nœuds déclarés avec les dépendances. Chaque mot-clé *without* introduit une contrainte négative. Par exemple, le premier *without* exclut que D_1 soit un signe de ponctuation [15]. On peut mettre plusieurs *without* dans un motif.

On utilise GREW-MATCH de façon itérative : pour un motif de départ, on observe les occurrences retournées et on identifie un sous-motif linguistiquement pertinent qui revient régulièrement. On itère ensuite les recherches avec le motif de départ et en

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

15. Nous avons écarté de notre étude les dépendances impliquant des signes de ponctuation et les autres nœuds déclarés dans *pattern* ne peuvent pas l'être car un signe de ponctuation n'est jamais gouverneur d'une dépendance.

excluant (*without*) tous les sous-motifs identifiés aux étapes précédentes. Le motif de la figure 4 illustre le type de requête que l'on obtient après quelques itérations.

Dans la première étape, on applique le motif formé seulement du champ *pattern* et du premier *without* au corpus SUD_FRENCH-GSD ; on trouve 955 occurrences du motif dans le corpus, correspondant à 955 croisements. Un sous-motif revenant souvent caractérise la montée de clitiques. Pour l'exclure, on ajoute le second *without* de la figure 4, qui signifie que nous ne voulons pas qu'un complément $\texttt{H}_1$ d'un verbe $\texttt{D}_2$ ait lui-même un dépendant $\texttt{D}_1$ qui soit un pronom personnel, ce pronom se situant avant le verbe $\texttt{D}_2$. En appliquant le motif enrichi de cette contrainte, on ne trouve plus que 284 occurrences dans le corpus. Donc 671 étaient dues à la montée de clitiques.

En observant un échantillon de ces 284 occurrences, on s'aperçoit que les croisements sont souvent dus à l'extraction profonde de propositions relatives. Nous employons le terme *extraction profonde* pour indiquer que le syntagme extrait n'est pas directement dépendant de la tête de la proposition relative. Pour exclure ce phénomène et en rechercher d'autres nous ajoutons les trois derniers *without* de la figure 4. Ils expriment que $\texttt{H}_2$ est l'antécédent d'un pronom relatif, que $\texttt{D}_2$ est la tête de la relative et que le pronom relatif repéré par le trait `PronType=Rel` dépend plus ou moins directement de $\texttt{H}_1$, qui dépend, lui, directement de $\texttt{D}_2$. La dépendance plus ou moins directe du pronom relatif explique la nécessité de trois *without*, qui, ensemble, expriment le rejet de l'extraction profonde d'une relative. Lorsque l'on applique le motif complet de la figure 4, on ne trouve plus que 44 occurrences de ce motif dans le corpus. Cela signifie que 240 croisements provenaient d'une extraction profonde d'une relative.

On poursuit ce processus jusqu'à ce qu'il ne reste plus qu'une dizaine d'occurrences qui peuvent correspondre à des phénomènes extrêmement rares, mais le plus souvent mettent en exergue des erreurs d'annotation.

C'est en appliquant cette méthode qu'a été menée l'étude linguistique des dépendances non projectives dont nous allons maintenant présenter les résultats.

4.2. *Les sources de la non-projectivité en français*

Dans la section précédente, nous avions montré l'existence de 11 motifs principaux de manifestation de la non-projectivité dans nos trois corpus étudiés. À l'aide de GREW-MATCH, nous avons appliqué la méthode qui vient d'être présentée à chacun des trois corpus en partant de chacun des 11 motifs. Ainsi, nous avons pu exhiber quatre phénomènes principaux sources de non-projectivité, qui couvrent 97 % des croisements. Le tableau 6 récapitule ces résultats en indiquant pour chaque phénomène le nombre de croisements auquel il donne lieu par corpus[16]. Nous allons maintenant détailler ces quatre phénomènes les uns après les autres.

16. Un motif correspond généralement à plusieurs phénomènes et un phénomène correspond à plusieurs motifs. Le lecteur trouvera un tableau récapitulant cette correspondance en ligne (`https://nakala.fr/10.34847/nkl.ce3cmi1q`).

phénomène linguistique	**SUD-GSD**	**UD-GSD**	**UD-SEQUOIA**	**total**
montée de clitiques	2 154 (67 %)	192 (16 %)	23	2 369 (53 %)
extraction profonde	496 (15 %)	196 (17 %)	48	740 (16 %)
minimisation de la longueur des dépendances	290 (9 %)	255 (22 %)	30	575 (13 %)
couples de mots dépendants distants	274 (9 %)	532 (45 %)	0	806 (18 %)
total	3 214	1 175	101	4 490

Tableau 6. *Sources linguistiques de la non-projectivité dans les corpus du français étudiés*

4.2.1. *La montée de clitiques*

Les pronoms clitiques sont accolés aux verbes dont ils sont les compléments mais quand ces verbes sont précédés d'auxiliaires, les clitiques montent devant les auxiliaires (Abeillé et Godard, 2001). La figure 5 illustre cette montée.

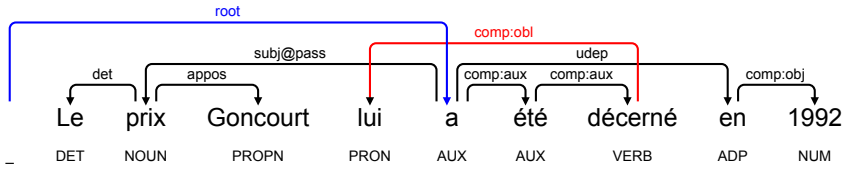

Figure 5. SUD_FRENCH-GSD *fr-ud-train_06154 : Le prix Goncourt lui a été décerné en 1992*

Dans cette phrase, la montée du clitique *lui* devant l'auxiliaire *a* entraîne nécessairement un croisement dans l'annotation SUD entre la dépendance (`décerné -[comp:obl]→ lui)` et la dépendance (`_ -[root]→ a)` dans l'arbre complété. Dans le corpus SUD_FRENCH-GSD, on trouve 612 occurrences de tels croisements.

Ces 612 occurrences représentent 28 % seulement des croisements qui, dans le tableau 6, sont considérés comme résultant d'une montée de clitique, car cette montée entraîne aussi des croisements secondaires. Sur la figure 5, la montée du clitique *lui* provoque un croisement avec les dépendants à gauche de l'auxiliaire. Dans l'exemple, il s'agit du croisement avec la dépendance (`a -[subj@pass]→ prix)`. Dans le corpus SUD_FRENCH-GSD il y a 819 croisements de ce type, soit 38 % des croisements provoqués par une montée de clitique. Ils sont plus nombreux que les croisements principaux car il peut y avoir plusieurs dépendants à gauche d'un même auxiliaire. Ces dépendants à gauche sont essentiellement des sujets (569), des modificateurs de phrases (169) et des conjonctions de coordination (79).

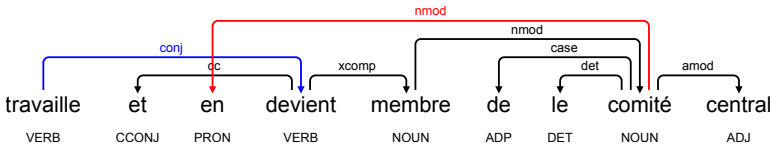

Figure 6. UD_FRENCH-GSD *fr-ud-train_07141 : [Smrkovský] travaille [pour la résistance communiste allemande] et en devient [finalement] membre du comité central*

La montée d'un clitique peut provoquer aussi des croisements secondaires avec les dépendants à droite de l'auxiliaire. Dans l'exemple, il s'agit du croisement de (`décerné -[udep]→ lui`) avec (`a -[udep]→ en`). Dans le corpus SUD_FRENCH-GSD, il y a 549 croisements de ce type, soit 25 % des croisements provoqués par une montée de clitique. Les dépendants à droite correspondant à ces croisements sont principalement des modificateurs de phrases (426) et des têtes de propositions coordonnées (78).

Dans le format UD, les auxiliaires ne sont pas la tête du complexe qu'ils forment avec le verbe principal donc la montée des clitiques devant les auxiliaires ne provoque aucun croisement. Néanmoins, certains clitiques ne dépendent pas directement du verbe auquel ils sont accolés mais d'un argument de ce verbe. Ce phénomène peut provoquer des croisements. Cela concerne essentiellement le clitique *en* lorsqu'il dépend de l'objet du verbe ou de l'attribut du sujet ou de l'objet. La figure 6[17] montre que le chemin de dépendances du verbe vers le clitique *en* qui lui est accolé peut être plus ou moins long, ici : *devient* → *membre* → *comité* → *en*. Ce phénomène pour le clitique *en* est responsable de respectivement 161, 185 et 21 croisements dans SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, soit respectivement 7 %, 96 % et 91 % des croisements résultant d'une montée de clitique.

4.2.2. *L'extraction profonde*

Les propositions relatives ou interrogatives donnent lieu dans beaucoup de cas à l'extraction d'un syntagme, mais quand ce syntagme est directement dépendant de la tête de la proposition où a lieu l'extraction, celle-ci n'entraîne pas de non-projectivité dans les dépendances. Pour que cette non-projectivité se produise, il faut que le syntagme extrait dépende d'un élément qui n'est pas la tête de la proposition relative ou interrogative. C'est alors que nous parlons d'*extraction profonde*.

Toute extraction profonde n'entraîne pas de non-projectivité. Il faut pour cela que le gouverneur du syntagme extrait se situe après la tête de la proposition rela-

17. Pour simplifier la présentation de l'annotation, nous ignorons certains passages de la phrase non essentiels pour notre propos ; ceux-ci sont marqués entre crochets dans l'énoncé de la phrase.

tive. La figure 7 montre un exemple d'extraction profonde qui entraîne de la non-projectivité. C'est *dont* qui est extrait de la relative comme complément du nom

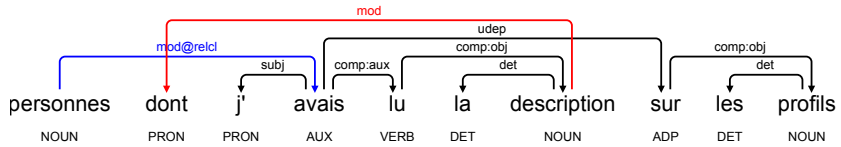

Figure 7. SUD_FRENCH-GSD *fr-ud-train_11296 : [j'ai envoyé des messages à plusieurs] personnes dont j'avais lu la description sur les profils [donnés par unicis]*

description. Ce gouverneur est bien situé après à la tête *avais* de la relative. La dépendance correspondant à l'extraction (`description -[mod]→ dont`) croise donc nécessairement celle de l'antécédent vers la tête de la relative, (`personnes -[mod@relcl]→ avais`). Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, nous trouvons respectivement 333, 153 et 35 occurrences de ce type de croisement, soit 11 %, 5 % et 6 % du nombre total de relatives et interrogatives.

Souvent, l'extraction profonde entraîne des croisements secondaires avec les dépendants à droite de la tête de la relative ou de l'interrogative. C'est ce que nous constatons sur l'exemple de la figure 7. La dépendance correspondant à l'extraction (`description -[mod]→ dont`) croise la dépendance (`avais -[mod]→ sur`). Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, nous rencontrons respectivement 149, 34 et 10 occurrences de tels croisements, soit 30 %, 17 % et 21 % du nombre total de croisements résultant d'extractions profondes.

Ce phénomène d'extraction profonde pour le français a été étudié par Candito et Seddah (2012a) sous le nom de *dépendances à longue distance effectives*. Ils y incluent aussi les clitiques qui dépendent d'arguments du verbe auquel ils sont accolés et dont nous avons parlé précédemment. Ils ont décrit relativement précisément ce phénomène sur deux treebanks, FRENCH TREEBANK et SEQUOIA, en le quantifiant. Ils se fondent sur une annotation LFG des corpus en utilisant les chemins fonctionnels propres à ce formalisme pour retrouver les dépendances à longue distance. Même si notre méthode est différente, nous retrouvons sur le corpus SEQUOIA les mêmes résultats en termes de croisements.

4.2.3. *La minimisation de la longueur des dépendances*

Ferrer Cancho (2006) a mis en évidence le fait que la meilleure façon d'ordonner les nœuds d'un arbre de dépendances pour minimiser la longueur des dépendances est de le faire de façon projective. Et la minimisation de la longueur des dépendances, MLD par la suite, aide à la compréhension comme l'a étudié Liu (2008). Cette minimisation peut être interprétée aussi comme une minimisation du flux des dépendances

dans un arbre de dépendances totalement ordonné, ce qui rend mieux compte de l'aspect cognitif de la question, comme l'ont montré Kahane et Yan (2019).

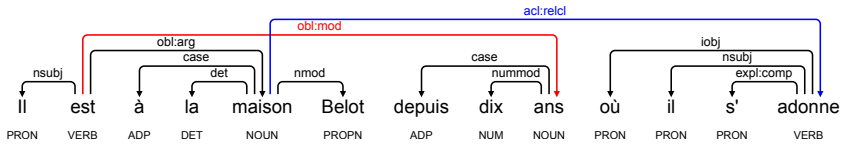

Figure 8. UD_FRENCH-SEQUOIA *annodis.er_00519 : Il est à la maison Belot depuis dix ans où il s'adonne [à la belote avec son copain Pierre Brungard]*

Si on met de côté les phénomènes de montée des clitiques et d'extraction profonde, qui viennent d'être présentés, on peut alors se demander pourquoi d'autres dépendances non projectives sont présentes dans les corpus. C'est que l'ordre linéaire des mots est soumis à d'autres contraintes qui peuvent aller à l'encontre du principe de la MLD. En particulier, la structure communicative avec la division entre thème et rhème et la focalisation impose ses propres contraintes sur l'ordre des mots. Il s'agit alors d'appliquer la MLD à l'intérieur de ces contraintes, ce qui entraîne parfois des croisements.

C'est en particulier le cas pour la phrase présentée sur la figure 8. Le verbe *est* a deux compléments *à la maison Belot* et *depuis dix ans* mais le premier complément est modifié par la relative *où il s'adonne à la belote avec son copain Pierre Brungard*. Or, cette relative n'est pas accolée à son antécédent comme c'est le cas habituellement. Elle en est séparée par *depuis dix ans*. Cet enchevêtrement entraîne un croisement de dépendances. On peut se demander pourquoi la phrase n'est pas plutôt structurée selon l'une des deux alternatives projectives suivantes : *Il est à la maison Belot où il s'adonne à la belote avec son copain Pierre Brungard depuis dix ans* et *Il est depuis dix ans à la maison Belot où il s'adonne à la belote avec son copain Pierre Brungard*. La première alternative, même si elle raccourcit la dépendance (`maison -[acl:relcl]→ adonne`), allonge considérablement la dépendance (`est -[obl:mod]→ ans`), si bien que cela entraîne une ambiguïté : on peut comprendre que c'est depuis dix ans qu'il s'adonne à la belote. La seconde n'est pas satisfaisante du point de vue de la structure communicative car l'information nouvelle qui veut être mise en avant c'est que cela fait dix ans qu'il est à la maison Belot et pas que c'est à la maison Belot qu'il est. La seule façon de concilier la structure communicative voulue avec la minimisation de la longueur des dépendances est la phrase de la figure 8. On retrouve ce type d'enchevêtrement de compléments ou modificateurs dans respectivement 182, 143 et 23 occurrences des treebanks SUD_FRENCH-GSD, UD_FRENCH-GSD et UD_FRENCH-SEQUOIA, soit 63 %, 56 % et 77 % du nombre total de croisements provenant de la minimisation de la longueur des dépendances.

Il est une autre configuration qui se rapporte au même phénomène et qui est illustrée par la figure 9 : le sujet d'un verbe peut voir un de ses modificateurs ou une apposition rejetés après le verbe. Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-

GSD, UD_FRENCH-SEQUOIA, on trouve respectivement 71, 69 et 4 occurrences de cette configuration, soit 24 %, 27 % et 13 % du nombre total de croisements provenant de la minimisation de la longueur des dépendances.

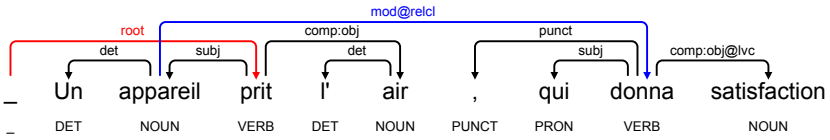

Figure 9. SUD_FRENCH-GSD *fr-ud-train_13250 : Un [second] appareil prit l'air [le 20 mars 1999], qui donna satisfaction*

4.2.4. *Les couples de mots dépendants distants*

Les constructions comparatives en français utilisent en général des couples de mots dépendants dont le premier est un adverbe (*plus*, *moins*, *aussi*, *autant*, *davantage*. . .) ou un adjectif (*autre*, *même*. . .) et le second la conjonction de subordination *que*. Il est en de même avec les constructions consécutives (*si*. . . *que*, *tel*. . . *que*, *tellement*. . . *que*. . .). Dans les deux types de constructions, les deux mots dépendants peuvent être distants l'un de l'autre et c'est cela qui est source de non-projectivité. On retrouve des couples de mots dépendants potentiellement distants pour d'autres constructions : *premier*. . . *à*, *à peine* . . . *que* par exemple.

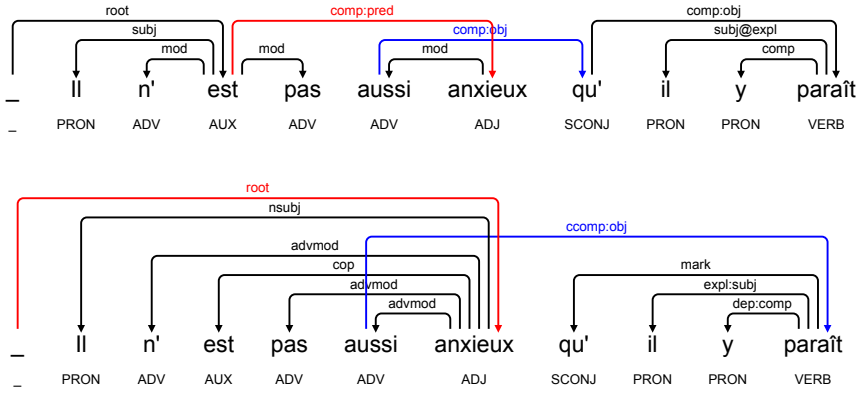

Figure 10. SUD_FRENCH-GSD *et* UD_FRENCH-GSD *fr-ud-dev_00420 : Il n'est pas aussi anxieux qu'il y paraît*

La figure 10 montre un exemple avec la construction *aussi* . . . *que*. Le schéma supérieur présente l'annotation SUD de la phrase qui fait apparaître le croise-

ment de la dépendance (`aussi -[comp:obj]→ qu'`) avec (`est -[comp:pred]→ anxieux`). Dans l'annotation UDprésentée dans la partie inférieure de la figure 10, la relation entre les mots distants *aussi* et *qu'* est exprimée de façon indirecte par la dépendance (`aussi -[ccomp:obj]→ paraît`), qui se croise cette fois avec (`_ -[root]→ anxieux`).

Dans les corpus SUD_FRENCH-GSD, UD_FRENCH-GSD, on trouve respectivement 197 et 194 occurrences de ces croisements. Dans UD_FRENCH-SEQUOIA, nous n'en trouvons aucun car il a été fait un choix différent pour la source de la dépendance exprimant la relation entre les mots distants. La source de cette dépendance n'est pas le premier mot du couple considéré mais son gouverneur. Le lecteur pourra aisément vérifier que dans l'annotation UD de la figure 10, si on déplace la source de la dépendance (`aussi -[ccomp:obj]→ paraît`) sur *anxieux*, on supprime la non-projectivité.

Comme le montre la figure 10, la dépendance principale $\mathtt{H}_2 \rightarrow \mathtt{D}_2$ (en bleu) induit des croisements secondaires avec des dépendances issues de $\mathtt{D}_1$ (la cible de la dépendance en rouge). Dans notre exemple, on n'en trouve aucun pour SUD mais quatre pour UD. Cette constatation est plus générale puisque sur l'ensemble des corpus SUD_FRENCH-GSD et UD_FRENCH-GSD, il y a respectivement 77 et 338 croisements secondaires induits par la dépendance impliquant les mots distants.

5. Conclusion

Notre étude sur les dépendances non projectives dans des corpus du français a fait apparaître la complexité limitée de la non-projectivité, que ce soit en nombre de trous ou de composantes connexes de trous. Elle a fait apparaître aussi le caractère très local du croisement de dépendances. Ces deux résultats sont indépendants du schéma d'annotation. En revanche, l'étude a montré que la fréquence de croisement était en partie déterminée par le schéma d'annotation.

Ensuite, elle a mis en évidence quatre sources linguistiques de non-projectivité : la montée de clitiques, l'extraction profonde, la minimisation de la longueur des dépendances et les couples de mots dépendants distants. Ces phénomènes ont déjà été largement étudiés par le passé (Abeillé et Godard, 2001 ; Candito et Seddah, 2012a ; Ferrer Cancho, 2006 ; Liu, 2008), mais en général indépendamment les uns des autres et pas sous l'angle de leur responsabilité dans la non-projectivité. L'étude que nous avons menée visait à l'exhaustivité en mettant en évidence toutes les sources linguistiques de non-projectivité dans les corpus considérés. Les quatre sources exhibées couvrent 97 % des croisements de dépendances dans les corpus considérés. L'intérêt de l'étude est aussi que les différents phénomènes ont été quantifiés. La seule autre étude à notre connaissance visant à déterminer l'ensemble des sources de non-projectivité dans un corpus du français est celle de Botalla (2014) menée sur le treebank RHAPSODIE, mais elle reste qualitative et elle considère différentes manifestations de la minimisation de la longueur des dépendances sans l'envisager dans toute sa généralité.

Le fait d'avoir considéré deux schémas d'annotation syntaxique très différents met en évidence que la non-projectivité dépend parfois des choix qui sont faits pour la tête des constituants. Plus précisément pour SUD et UD la divergence porte sur la tête des groupes prépositionnels, des propositions subordonnées et des noyaux verbaux avec auxiliaires. La non-projectivité liée à l'extraction profonde et à la minimisation de la longueur des dépendances est pour une bonne part indépendante de ces choix de têtes, alors que la non-projectivité liée à la montée des clitiques est due la plupart du temps au choix de l'auxiliaire comme tête dans le couple qu'il forme avec le verbe. Pour ce qui est des couples de mots dépendants distants, ce qui joue n'est pas le choix de la tête d'un constituant mais celui de la source de la dépendance qui établit une relation plus ou moins directe entre les deux mots distants, comme cela est expliqué avec l'exemple de la figure 10.

Enfin, l'intérêt de l'étude effectuée est qu'elle utilise une méthode universelle, valable quelle que soit la langue, quel que soit le schéma d'annotation et quel que soit le type de corpus. C'est vrai pour la première phase d'étude topologique. L'aboutissement de cette première phase est de mettre en évidence un nombre limité de motifs finis couvrant l'essentiel des croisements de dépendance. Bien entendu, ces motifs dépendent de la langue concernée, mais ils sont le point de départ d'une étude exhaustive et quantifiée à l'aide de l'outil GREW-MATCH des phénomènes linguistiques qui sont source de non-projectivité. Comme prolongement immédiat de notre étude, il serait intéressant d'appliquer la méthode au treebank de l'oral SPOKEN[18]. Cela permettrait, par comparaison avec l'étude présentée ici sur des treebanks de l'écrit, de mettre en évidence les points communs et les spécificités de l'oral.

Remerciements

Merci à Sylvain Kahane, Bruno Guillaume et aux relecteurs anonymes pour leurs commentaires pertinents.

6. Bibliographie

Abeillé A., Clément L., Liégeois L., « Un corpus annoté pour le français : le French Treebank », *TAL*, vol. 60, p. 19-43, 2019.

Abeillé A., Godard D., « Deux types de prédicats complexes dans les langues romanes », *Linx. Revue des linguistes de l'université Paris X Nanterre*, n° 45, p. 167-175, 2001.

Béchet D., Lacroix O., « CDGFr, un corpus en dépendances non-projectives pour le français », *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, p. 206-212, 2015.

Bloomfield L., *Language*, New-york, 1933.

18. `https://github.com/surfacesyntacticud/SUD_French-Spoken`

Bonfante G., Guillaume B., Perrier G., *Application de la réécriture de graphes au traitement automatique des langues*, vol. 1 of *Série Logique, linguistique et informatique*, ISTE editions, 2018.

Botalla M.-A., *« Analyse du flux de dépendance dans un corpus de français oral annoté en micro-syntaxe »*, Master's thesis, Université Paris III Sorbonne Nouvelle, 2014.

Candito M.-H., Crabbé B., Denis P., Guérin F., « Analyse syntaxique statistique du français : des constituants aux dépendances », *TALN 2009*, Senlis, France, 2009.

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*, Lisbon, Portugal, 2012a.

Candito M., Seddah D., « Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical », *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France, 2012b.

Corro C., Le Roux J., Lacroix M., Rozenknop A., Calvo R. W., « Dependency parsing with bounded block degree and well-nestedness via Lagrangian relaxation and branch-and-bound », *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 355-366, 2016.

Ferrer Cancho R., « Why do syntactic links not cross ? », *Europhysics Letters*, vol. 76, n° 6, p. 1228-1235, 2006.

Fitialov S. J., « O modelirovanii sintaksisa v strukturnoj lingvistike », *Problemy strukturnoj lingvistiki, Moskva*p. 100-114, 1962.

Gaifman H., « Dependency systems and phrase-structure systems », *Information and control*, vol. 8, n° 3, p. 304-337, 1965.

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*, Brussels, Belgium, 2018.

Gerdes K., Guillaume B., Kahane S., Perrier G., « Improving Surface-syntactic Universal Dependencies (SUD) : surface-syntactic relations and deep syntactic features », *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, Paris, France, 2019.

Gómez-Rodríguez C., Sartorio F., Satta G., « A polynomial-time dynamic oracle for non-projective dependency parsing », *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 917-927, 2014.

Guillaume B., de Marneffe M.-C., Perrier G., « Conversion et améliorations de corpus du français annotés en Universal Dependencies », *Traitement Automatique des Langues*, vol. 60, n° 2, p. 71-95, 2019.

Hajicová E., Havelka J., Sgall P., Veselá K., Zeman D., « Issues of Projectivity in the Prague Dependency Treebank. », *Prague Bull. Math. Linguistics*, vol. 81, p. 5-22, 2004.

Harper K. E., Hays D. G., *The use of machines in the construction of a grammar and computer program for structural analysis*, Rand Corporation, 1959.

Havelka J., Mathematical Properties of Dependency Trees and their Application to Natural Language Syntax, PhD thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, 2007.

Holan T., Kubon V., Oliva K., Plátek M., « On complexity of word order », *TAL. Traitement automatique des langues*, vol. 41, n° 1, p. 273-300, 2000.

Kahane S., Gerdes K., *Syntaxe théorique et formelle, Volume 1 : Modélisation, unités, structures*, Language Science Press, 2020. A paraître.

Kahane S., Yan C., « Advantages of the flux-based interpretation of dependency length minimization », *First international conference on Quantitative Syntax (Quasy)*, 2019.

Kuhlmann M., Nivre J., « Mildly non-projective dependency structures », *Proceedings of the COLING/ACL on Main conference poster sessions*, Association for Computational Linguistics, p. 507-514, 2006.

Kuhlmann M., Nivre J., « Transition-based techniques for non-projective dependency parsing », *Northern European Journal of Language Technology (NEJLT)*, vol. 2, n° 1, p. 1-19, 2010.

Lecerf Y., Ihm P., Eléments pour une grammaire générale des langues projectives, Rapport GRISA n° 1, Euratom, 1960.

Liu H., « Dependency distance as a metric of language comprehension difficulty », *Journal of Cognitive Science*, vol. 9, n° 2, p. 159-191, 2008.

Mambrini F., Passarotti M., « Non-projectivity in the Ancient Greek dependency treebank », *Proceedings of the second international conference on dependency linguistics (Depling 2013)*, p. 177-186, 2013.

Marcus S., « Sur la notion de projectivité », *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, vol. 11, p. 181-192, 1965.

McDonald R., Pereira F., Ribarov K., Hajič J., « Non-projective dependency parsing using spanning tree algorithms », *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 523-530, 2005.

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'cuk I. A. *et al.*, *Dependency syntax : theory and practice*, SUNY press, 1988.

Miletic A., Urieli A., « Non-projectivity in Serbian : Analysis of formal and linguistic properties », *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, p. 135-144, 2017.

Nivre J., « Constraints on non-projective dependency parsing », *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Nivre J., « Non-Projective Dependency Parsing in Expected Linear Time », *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, Suntec, Singapore, p. 351-359, August, 2009.

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.

Straka M., Hajic J., Straková J., Hajic Jr J., « Parsing universal dependency treebanks using neural networks and search-based oracle », *International Workshop on Treebanks and Linguistic Theories (TLT14)*, p. 208-220, 2015.

Tapanainen P., Jarvinen T., « A non-projective dependency parser », *Fifth Conference on Applied Natural Language Processing*, Association for Computational Linguistics, Washington, DC, USA, p. 64-71, 1997.

Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Emily M. BENDER, Alex LASCARIDES. Linguistic Fundamentals for Natural Language Processing II : 100 Essentials from Semantics and Pragmatics. *Morgan & Claypool publishers*. 2020. 250 pages. ISBN 978-1-68173-073-8.

Lu par **Jocelyn AZNAR**

ZAS, Berlin

Emily M. Bender et P. C. Alex Lascarides proposent un ouvrage destiné aux personnes issues de l'informatique, un ouvrage qui synthétise en cent points l'essentiel de la sémantique et de la pragmatique pour élaborer des systèmes de traitement informatisés de la langue naturelle.

Emily M. Bender et P. C. Alex Lascarides proposent aux personnes élaborant des systèmes de traitement de la langue naturelle de se familiariser à travers cent une vignettes[1] aux questions de sémantique et de pragmatique. Ce volume fait suite à un premier opus par Bender portant sur la syntaxe et la morphologie. Comme le titre du livre l'annonce, deux parties peuvent être distinguées : une première sur la sémantique, et une deuxième sur la pragmatique. Un premier chapitre introductif présente l'enjeu interdisciplinaire du livre qui s'adresse aux personnes issues du monde de l'informatique, le chapitre conclusif fournit des références sur les ressources informatisées.

La sémantique : chapitres 2 à 8

La première approche de l'étude du sens donné est logique, en termes de validité et de référence. Les autrices poursuivent par différentes acceptions du sens : conventionnel, commun, savoir général, émotionnel et social. Elles abordent également les actes du langage, le savoir partagé ou « *common ground* », et examinent la relation entre les actes et la cohérence. La problématique de l'ambiguïté pose les bases pour amener, lors d'une comparaison entre la stratégie d'analyse des systèmes de traitement informatisé et celle d'un être humain, la « défaisabilité » d'une analyse. La conversation en face à face conduit à discuter de la référence à des évènements ou à la situation et aux interactions avec les évènements non linguistiques. Le chapitre se termine sur trois niveaux de

1 Une vignette est un passage de longueur variable, d'un paragraphe à six pages, consacré à une problématique bien définie.

représentations du sens : dérivées des formes linguistiques, l'engagement public du locuteur et les effets cognitifs de l'engagement public.

Les chapitres 3 et 4, sur la sémantique lexicale auraient pu être fusionnés en un seul chapitre, ce qui aurait été plus concis. Après avoir situé la sémantique lexicale vis-à-vis de la grammaire formelle, les autrices définissent trois concepts clefs : le sens des mots, le rôle sémantique et leur connotation. Le chapitre 4 débute par les différentes ressources sur les sens d'une forme lexicale. Les différentes formes de la polysémie sont détaillées : celle dite régulière, la constructionnelle et enfin l'extension de sens. Nous voyons aussi que la représentation du sens lexical facilite l'analyse de la polysémie constructionnelle. À partir de la description de l'homonymie et de la productivité de la dérivation, nous voyons que certaines formes, soit parce que courantes, soit de par des changements diachroniques, se comportent de manière irrégulière. L'analyse des cooccurrences est décrite notamment à travers les espaces vectoriels résultant de l'adaptation de l'analyse distributionnelle à la sémantique. Les approches vectorielles présentent l'avantage de quantifier la similarité entre des mots. Les autrices abordent ensuite la métaphore et montre qu'elle repose sur le sens littéral d'un terme pour désigner un autre objet et le transfert de sens, phénomène à l'origine de bruits dans les modèles sémantiques ou d'erreurs lors de tâches telles que l'inférence. Elles terminent le chapitre sur l'importance de la « défaisabilité » de l'analyse d'une unité lexicale.

Le chapitre 5 porte sur les rôles sémantiques, les autrices détaillent l'intérêt du concept pour représenter les sens d'un mot, en particulier pour les verbes. Elles décrivent le rôle des arguments vis-à-vis des prédicats, leur analyse, les restrictions lors de la sélection des arguments et les ressources implémentant ces enjeux. Le chapitre 6 introduit les concepts de collocations et d'expressions polylexicales, ces expressions dont la totalité présente des particularités non prédictibles *via* leurs composants. Les autrices avancent que ces particularités résultent de leurs formes linguistiques. L'argument ne prend cependant pas en compte leurs connotations ni la diachronicité du phénomène. Si les collocations présentent des ambiguïtés, elles en possèdent moins que les termes simples. Cette « force » de l'ambiguïté est mesurée grâce à l'information mutuelle de ses termes[2]. Les expressions varient quant à leur flexibilité, leur compositionnalité ou selon qu'elles suivent ou non les règles dérivationnelles. La dernière vignette conclut sur la difficulté de représenter le rapport de sens entre une expression polylexicale et ses composants.

Les chapitres 7 et 8 portent tous deux sur la sémantique compositionnelle. Après avoir défini la sémantique compositionnelle et sa représentation, les autrices reviennent sur ses usages : l'analyse des comparaisons, des coordinations, des quantifiants et notamment des déterminants. Elles abordent ensuite la question de la résolution de la portée des opérateurs, comment la syntaxe permet de la déterminer

2 Si l'application de la méthode d'information mutuelle en linguistique est pertinente, l'exemple donné comme n'étant intuitivement pas une collocation : « *old man* », surprend davantage car le terme fait l'objet de multiples entrées dans les dictionnaires anglophones, montrant son statut lexical particulier.

et la sélection du sens puis elles concluent sur les limites de l'approche compositionnelle. Le chapitre 8 correspond à une proposition de Bender d'étendre l'acceptation de la sémantique compositionnelle comme subsumant la morphologie et la syntaxe. Cette proposition intéressante n'aide cependant pas les lecteurs néophytes à resituer les concepts habituellement associés à la syntaxe et à la morphologie qui sont traités dans ce chapitre (le marquage des cas et la grammaticalisation, le temps, l'aspect et l'évidentialité).

La pragmatique : chapitres 9 à 13

Le premier chapitre sur la pragmatique débute par l'importance d'une fonction de mise à jour des informations dans la compréhension du discours, il poursuit sur la pertinence d'une représentation dynamique transmettant des informations entre les parties du discours. Les autrices décrivent également le fonctionnement de la cohérence et les liens qui existent entre le niveau lexical et la cohérence. Le chapitre se termine sur une modélisation du discours à travers la théorie du jeu.

Le chapitre 10 présente la problématique de la résolution des références et aborde la détermination de l'antécédent d'un pronom et poursuit sur le rôle des informations grammaticales. Il continue sur l'utilisation d'une représentation logique pour résoudre la résolution des références, sur le rôle des modalisateurs et termine sur le rôle de la structure du discours.

Les deux autrices commencent dans le chapitre 11 par distinguer les implications des présuppositions à l'aide notamment du test de projectivité. Elles détaillent ensuite la diversité des termes qui enclenchent une présupposition puis développent les mécanismes qui amènent ou non à la projection. La question de l'accommodation d'un présupposé est ensuite abordée, mettant en avant le rôle de l'anaphore. Le chapitre conclut sur la relation entre la cohérence du discours et les présuppositions. Le chapitre 12 décrit le concept de statut informationnel et montre comment il peut prendre des formes variées que ce soit au niveau morphologie ou syntaxique. Les autrices introduisent le concept de structure informationnelle et associent son analyse à celle du discours. Elles intègrent également la problématique de la diversité linguistique et le rôle de la prosodie. La dernière vignette du chapitre décrit les interactions entre la structure informationnelle et les conditions de vérité.

Le chapitre 13 porte sur les implicatures et leurs caractéristiques conversationnelles ou conventionnelles, sûres ou non assumées par le locuteur. Les autrices discutent également du rôle du silence, de la prosodie et de sa représentation et de son évaluation en termes logiques. Elles montrent enfin que la dérivation des relations entre les propositions ne nécessite pas de connaître ce que pense une personne.

Conclusion

L'objectif de fournir des bases en sémantique et en pragmatique tout en réfléchissant à leurs traitements automatiques est atteint. Chaque vignette suit une progression claire tout en étant reliée aux vignettes traitant de sujets connexes, ce qui permet une lecture aussi bien linéaire que thématique. Les concepts sont le plus souvent définis avec des termes clairs et leur pertinence mise en avant par des

exemples. Les problèmes sont discutés à travers des références récentes, mettant en avant les travaux des deux autrices, mais également ils sont mis en perspective à l'aide de textes plus anciens. Toutefois, certains passages, étrangement intégrés au texte ou propositions originales, ne s'insèrent pas dans cet objectif d'aller à l'essentiel. On peut également regretter certains choix de Bender et Lascarides, comme de ne pas avoir évoqué d'approches intégrant des questionnements sociohistoriques. Ce choix perpétue une préférence générale dans le cadre du traitement informatisé des langues à ne pas considérer ces questions.

Michael McTEAR. Conversational IA. Dialogue Systems, Conversational Agents, and Chatbots. *Morgan & Claypool publishers*. 2021. 190 pages. ISBN 978-1-63639-031-4.

Lu par **Fabrice LEFÈVRE**

Avignon Université – LIA-CERI

Un ouvrage de bonne qualité, embrassant largement le domaine, mais dont la couverture des avancées de recherche actuelles reste lacunaire.

L'ouvrage du professeur McTear se propose de nous offrir une introduction au large domaine de l'« IA conversationnelle ». Afin que les choses soient claires d'emblée le sous-titre nous confirme qu'il s'agit bien de « systèmes de dialogues, agents conversationnels et chatbots ». Sans réellement nous donner de justification à l'introduction d'un nouveau terme pour son titre (le besoin de renouveau suffit sûrement), l'auteur propose dès la préface de revenir ensuite dans l'ouvrage au terme mieux défini de « systèmes de dialogues ». L'histoire des technologies des systèmes de dialogues est présentée selon un triptyque : systèmes à base de règles (ou experts), systèmes probabilistes et approches neuronales de bout en bout.

Après un chapitre d'introduction de facture très classique sur les systèmes de dialogues, une partie plus courte s'attache aux systèmes à base de règles. Cette présentation est justifiée par le nombre encore grand de solutions industrielles reposant sur ces principes simples, mais souvent efficaces. Si la palette complète des techniques qui ont été proposées et expérimentées est largement sous-estimée dans cette présentation, la part importante accordée aux outils de développement actuels est très pertinente et intéressante. À juste titre, le chapitre se poursuit jusqu'à débusquer les avatars de ce paradigme dans les très récentes campagnes d'évaluation, tel le Alexa Prize Challenge organisé depuis 2017 par Amazon pour mettre en avant sa solution d'interaction vocale pour enceintes connectées.

Les systèmes de dialogues par approches probabilistes sont présentés plus rapidement dans le chapitre 3. Les grandes lignes et la motivation sont dessinées très rapidement en deux sous-parties, qui sont poursuivies par une présentation plus spécifique de l'apprentissage par renforcement appliqué au gestionnaire de dialogues (peu convaincante, par ailleurs, mais j'y reviendrai). À la suite de ce chapitre, et au lieu de poursuivre sur la lancée qui conduit à l'apparition des

approches neuronales, l'auteur prend un chemin de traverse pour nous entretenir du problème de l'évaluation des systèmes de dialogues. Ce chapitre 4, avec de nombreux exemples, est très informatif, bien qu'il doive à sa position de laisser en suspens certains aspects récents qui ne pourront être traités qu'après le chapitre suivant. Pour le lecteur éclairé, on y gagne du rythme à la lecture, le néophyte pourra s'y perdre un peu.

Enfin le chapitre 5 est consacré à la revue des approches basées sur les réseaux de neurones appliqués à l'interaction humain machine. Il faut comprendre ici les approches neuronales qui concernent la modélisation du dialogue. Celles utilisées pour les modules internes (compréhension, génération de texte…) ont déjà été présentées dans le chapitre 3. La présentation reprend la chronologie d'apparition des technologies nouvelles, des variantes progressives et raffinées de l'encodeur-décodeur pour les systèmes conversationnels aux modèles plus complexes permettant des systèmes guidés par la tâche. La présentation de quelques systèmes récents (Meena, BlenderBot, GPT-3…) vient agrémenter cette partie, qui étonnamment se termine par la présentation de bases de données et campagnes d'évaluation qu'on perçoit plus en rapport avec le chapitre 4.

Enfin, l'ouvrage s'achève avec un chapitre complet dédié à l'ouverture sur les défis actuels et futures directions qui est, sans conteste, le plus intéressant du livre, et qui, malgré un découpage discutable, offre un point de vue argumenté sur la déclinaison des enjeux en cours pour le domaine de recherche.

Globalement un des points forts majeurs du livre est assurément la qualité de la rédaction. Avec précision et clarté, les concepts sont introduits sans difficulté. La longueur globale de l'ouvrage reste raisonnable pour une lecture informative dans un temps raisonnable.

L'auteur développe sa vision très large du domaine, en remontant dans le temps pour mettre en perspective les tendances récentes. Sa très grande expertise des travaux menés jusqu'au début des années 2000 apparaît clairement. Et ainsi, çà et là, on apprécie le rappel de travaux « anciens » en lien avec des besoins semblant nouveaux, pour recadrer légèrement cette « nouveauté » un peu vite revendiquée. Par endroits on aimerait que cet adossement à une littérature injustement occultée dans les publications récentes soit un peu plus appuyé, c'était une occasion unique.

La présentation des paris récents est pertinente, ainsi que la liste finale des ressources systèmes, données… même si elle est loin d'être exhaustive. Ainsi, les grands acteurs du domaine (GAFAM en tête) sont peu présents, alors que leur participation au domaine, qu'on la loue ou la déplore, est énorme. Liste qui présente aussi le risque de n'être plus représentative très rapidement.

Dans la perspective d'ensemble de l'ouvrage, on peut regretter qu'il ne soit pas mieux explicité que l'approche neuronale n'est qu'une poursuite opportune des recherches en *data-driven dialogue systems*. L'introduction des propositions neuronales de type *seq2seq* avait aussi son pendant dans le monde préneurones (bien que l'on connaissait et utilisait déjà les réseaux de neurones à l'époque, mais ils étaient moins performants). En effet, le recours aux modèles de langages statistiques

(y compris au niveau phrastique) est bien plus ancien que 2011. On tend donc à oublier (ou du moins à minimiser) le fait que les modèles d'architectures présentés ici comme des panacées sont étudiés et proposés depuis… le siècle précédent ! Seule l'évolution des performances est un fait marquant et ayant un très fort impact sur ce domaine.

À ce titre, à la lecture du texte, il semble d'ailleurs vraisemblable que certaines notions récentes n'aient pas été bien assimilées par l'auteur. Quelques (très) bons tutoriaux sont cités dont l'auteur aurait pu (encore) mieux s'inspirer pour mettre en avant les évolutions récentes de la discipline. En effet, si un modèle de type encodeur-décodeur ne présente pas d'avancée conceptuelle importante pour le domaine, mais l'accès à un nouveau palier de performance, l'articulation avec des modèles à base de mémoire (*Mem2Seq*), par exemple, introduit des enjeux nouveaux sur l'organisation formelle des systèmes méritant une présentation plus détaillée.

De même, il existe des faiblesses dans la présentation de l'état de l'art. Ainsi, on s'étonne de ne pas trouver une plus grande attention apportée à l'apprentissage par renforcement appliqué aux systèmes d'interactions humain machine. En effet, depuis la dernière décennie, il s'agit clairement d'une des voies les plus travaillées par la communauté pour amener à maturité les systèmes basés sur les données. Cette voie de recherche conduit à entrevoir l'apprentissage continu des systèmes et beaucoup d'autres aspects primordiaux pour le déploiement dans des applications réelles.

De ce fait, certains enjeux sont quelque peu minimisés : les systèmes multidomaines, l'apprentissage en ligne, l'adaptation au locuteur... Et, de façon plus générale, il aurait été plus intéressant que l'essentiel des discussions apparaissant dans les « *Future Directions* » ait été rapatrié dans le corps du livre et mieux développé. Car beaucoup de discussions correspondent, en fait, à des recherches déjà bien établies et pour lesquelles un point sur l'état de l'art aurait été plus enrichissant qu'une simple présentation rapide du thème et de ses enjeux.

Si l'ouvrage n'a pas forcément fait les bons choix quant à son découpage thématique, le rapport à la littérature est aussi parfois discutable. Cela reste de la liberté de l'auteur de choisir ses références, mais beaucoup de citations ne correspondent pas réellement aux travaux séminaux des sujets concernés, mais à leur reprise ultérieure (mais on peut toujours se couvrir sur ce point : « Ils l'ont bien mieux fait ! »), surtout la surreprésentation injustifiée des travaux d'un des collègues de l'auteur dépasse toutes les limites habituelles (plusieurs apparitions à chaque page de la bibliographie !)

En conclusion le plus grand reproche qui peut être fait à cet ouvrage est sûrement ne pas respecter le parti pris de sa collection. Dans les *Synthesis on HLT* un choix fort, et fort appréciable, consiste à se focaliser sur un point (très) précis. Ceci afin d'en donner un panorama le plus clair, le plus précis et le plus actuel possible. En comparaison « *Conversational AI* » semble embrasser trop large et, de ce fait, ne convainc pas. Sur les aspects anciens, il reste lacunaire et peu explicatif (par exemple, comment fonctionne une approche par plan, un analyseur sémantique par règles probabilistes, comme Phoenix...), et sa tentative d'appréhender le monde

nouveau (surtout neuronal, on l'aura compris) risque de n'atteindre personne : la présentation des techniques est bien trop limitée pour un non-expert et inutile pour un familier de l'IA récent.

Xavier AIMÉ, Frank ARNOULD. Modélisation ontologique & psychologies. Une influence réciproque. *Éditions Matériologiques*. 2021. 238 pages. ISBN 978-2-37361-260-8.

Lu par **Éric HENNEKEIN**

EPHE (École Pratique des Hautes Études), Paris

---

*En plaçant de façon opportune, en hors-d'œuvre, un extrait de l'*Encyclopédie *de Diderot et d'Alembert (1765) faisant la distinction entre ontologie*[3] *naturelle et artificielle, Xavier Aimé et Frank Arnould expriment d'emblée que, dans leur ouvrage, ils vont présenter une synthèse de leurs travaux sur les relations entre psychologie et modélisation ontologique. Le sujet est d'autant plus passionnant qu'ils proposent une réflexion interdisciplinaire que doivent partager les tenants de ces deux disciplines. Et, afin de participer aux développements des technologies de l'intelligence artificielle, de l'apprentissage automatique, du traitement du langage, de façon scientifiquement fructueuse, dans le cadre d'une participation transdisciplinaire, ils invitent les chercheurs en neurosciences cognitives, les linguistes et les philosophes à s'inscrire, pour le moins, dans une démarche de clarification et d'uniformatisation des terminologies et des concepts qu'ils utilisent.*

De même, et ce n'est pas anodin, ils partent de l'exemple des travaux pionniers en psychologie préscientifique d'Ebbinghaus sur la mémoire, au tournant du XX^e^ siècle, puis de l'attention, pour mettre en évidence l'émergence de leurs nombreuses définitions et conceptualisations. Les approximations de sens, vernaculaire *versus* scientifique, qui en découlent, vont mettre en évidence des divergences, voire des controverses, au cours de l'appropriation par les diverses disciplines qui abordent ces notions, ce qui devient crucial avec le développement récent des sciences cognitives.

Les auteurs relèvent une controverse propre aux ontologies cognitives et computationnelles faisant débat en ingénierie des connaissances : faut-il envisager les ontologies computationnelles comme une représentation plus ou moins rationnelle que nous partagerions ou comme des représentations de la réalité, que la science décrit ? Faut-il adopter une méthode *top-down* ou *bottom-up* ? La théorie des concepts, définie dès l'Antiquité par les philosophes, a été sérieusement remise en cause, dans les années 1950 par Wittgenstein, puis par la psychologie cognitive, laissant apparaître l'importance d'une révision des modèles, des effets de typicalité vers une proximité sémantique entre les concepts. Mais, reprenant l'argumentation d'auteurs (Declerk et Charlet) pour qui les ontologies computationnelles ne doivent

---

3 « Ontologie : description formelle des entités censées exister dans un domaine et leurs relations », d'après le glossaire de l'ouvrage.

pas chercher à reproduire le fonctionnement cognitif humain, et la façon dont il catégorise les objets, mais plutôt à les considérer comme des artéfacts permettant d'augmenter la cognition humaine.

La structure conceptuelle de la cognition fait débat au sein de la communauté scientifique en général. Mais l'ingénierie ontologique propose aux chercheurs des techniques et des outils, pour formaliser explicitement les concepts de la cognition, et leurs représentations – sans se substituer à la recherche scientifique ! –, afin de dépasser les ambiguïtés conceptuelles, taxonomiques propres au domaine de la cognition (y compris dans leurs sous-concepts et les inférences qu'ils induisent). Grâce aux ontologies computationnelles, on pourrait vérifier la consistance logique des modèles théoriques et évaluer leur interopérabilité. Pour cela, il faut disposer de référentiels sachant raisonner afin de faciliter l'activité de recherche, mais Aimé et Arnould relèvent plusieurs difficultés qui sont à surmonter : 1) les chercheurs doivent s'accorder sur une modélisation formalisable ; 2) ils doivent être capables d'envisager une traduction de l'anglais des terminologies employées afin, d'une part, de permettre l'applicabilité à des contextes et corpus et, d'autre part, de concevoir une ontologie de référence ; 3) il faut envisager une structuration des diverses communautés scientifiques autour d'institutions de référence afin d'accompagner le développement d'une ontologie de la cognition ; 4) il faut favoriser, dans l'interdisciplinarité, l'apprentissage et la maîtrise de logiciels et de formats du Web sémantique, y compris en *machine learning*, afin de développer une ontologie computationnelle.

La psychologie, cognitive, mais aussi sociale, les sciences comportementales et les neurosciences doivent définir un cadre théorique général, décrivant de façon explicite, et formalisée, les réseaux de concepts qu'ils utilisent : ce qui permettrait, notamment, non seulement une réplication des expériences, mais aussi d'envisager une prédictibilité de résultats (fiabilité de la donnée, son indexation, etc.). Aimé et Arnould permettent au lecteur de se familiariser aux ontologies computationnelles, aux technologies du Web sémantique et à leurs fonctionnements, tant sur un plan didactique, qu'au travers d'exemples que l'ingénierie ontologique met en pratique. Celle-ci est en capacité de fournir aux chercheurs en psychologie et en neurosciences cognitives un langage formel qui permet de raisonner, d'assurer des inférences de classes et d'évaluer la consistance logique de la modélisation. Aimé et Arnould poursuivent en disant qu'ayant posé une première réflexion sur les entités de la cognition de haut niveau, ils poussent à envisager d'autres entités et d'autres relations primitives de la cognition. C'est ainsi qu'ils proposent notamment d'envisager des ontologies cognitives alternatives, à savoir, citant Hutchinson et Barrett (2019), un cadre prédictif de la cognition humaine se basant sur un traitement de haut niveau de l'information, interne et permanent, qui maintient une homéostasie corps monde, sauf lors de traitement de stimulus inattendu nécessitant un traitement cognitif spécifique.

C'est cette approche de modélisation ontologique, qu'à titre personnel nous soutenons, qui doit permettre d'intégrer et de réconcilier de façon fort intéressante de nombreux pans des différentes théories cliniques, notamment, de la psychologie,

tout en participant à l'émergence d'une nouvelle dialectique homme et intelligence artificielle.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Rémi CARDON : remi.cardon@protonmail.com

Titre : Simplification automatique de textes spécialisés et techniques

Mots-clés : simplification automatique de textes, corpus et ressources, textes biomédicaux.

Title: *Automatic Text Simplification of Specialized and Technical Texts*

Keywords: *automatic text simplification, corpora and resources, biomedical texts.*

Thèse de doctorat en sciences du langage, Savoirs, Textes, Langages, UMR 8163, Université de Lille, sous la direction de Natalia Grabar (CR HDR, CNRS) et Anne Carlier (Pr, Sorbonne Université, Paris). Thèse soutenue le 19/04/2021.

Jury : Mme Natalia Grabar (CR HDR, CNRS, codirectrice), Mme Anne Carlier (Pr, Sorbonne Université, Paris, codirectrice), Mme Cécile Fabre (Pr, Université de Toulouse Jean Jaurès, rapporteuse), M. Thomas François (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Emmanuelle Canut (Pr, Université de Lille, présidente), M. Pascal Denis (CR, Inria, examinateur), M. Thierry Hamon (MC, Université Sorbonne Paris Nord, examinateur), M. Horacio Saggion (MC, Universitat Pompeu Fabra, Barcelone, Espagne,).

Résumé : *La simplification automatique de textes est un domaine du traitement automatique des langues (TAL) qui vise à traiter des textes difficiles à lire pour un public donné de façon à les rendre plus accessibles. Notre objectif consiste à simplifier automatiquement les textes médicaux et de santé. Nous présentons l'ensemble de notre travail sur cette question, qui va de la collecte et analyse de corpus jusqu'aux expériences en simplification automatique.*

Nous commençons par la collecte d'un corpus comparable de textes médicaux. Ce corpus est constitué de couples de documents qui traitent du même sujet : l'un s'adressant à un public spécialiste et l'autre à un public néophyte. Le corpus contient trois types de textes : des informations sur les médicaments, des revues systématiques de littérature médicale et des articles encyclopédiques. Une fois les documents collectés, nous annotons un sous-ensemble de ces documents et analysons les transformations linguistiques qui y sont mises en œuvre lors de la simplification.

À partir du corpus comparable, nous mettons en place une méthode pour en extraire un corpus parallèle, c'est-à-dire un corpus comprenant des couples de phrases qui ont le même sens, mais diffèrent par leur degré de difficulté. Ce type de corpus représente le matériau principal pour les méthodes de simplification automatique. Notre méthode d'extraction de phrases parallèles comporte deux étapes : (1) le préfiltrage de paires de phrases candidates à l'alignement selon des heuristiques syntaxiques et (2) la classification binaire permettant de distinguer les phrases en relation de simplification. Nous évaluons différents classifieurs ainsi que l'influence du déséquilibre des données sur les performances. Afin de valoriser ce corpus parallèle, nous créons également un corpus de paires de phrases annotées selon leur similarité sémantique, avec des scores allant de 0 (sémantique indépendante) à 5 (même sémantique). Les deux corpus sont disponibles pour la recherche.

Enfin, nous présentons une série d'expériences en simplification automatique de textes médicaux en français. Ainsi, nous mettons en œuvre une méthode neuronale issue de la traduction automatique. Nous utilisons plusieurs ressources : le corpus parallèle médical construit par nous, le corpus parallèle de langue générale automatiquement traduit par nous de l'anglais vers le français ainsi qu'un lexique qui apparie des termes médicaux avec des termes ou paraphrases accessibles au grand public. Nous décrivons le protocole expérimental et menons une évaluation en deux volets, quantitatif et qualitatif. Les résultats sont comparables à l'état de l'art de la simplification en langue générale et montrent que les simplifications produites peuvent être exploitées dans le cadre d'une tâche de simplification assistée par ordinateur.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03343769

Hélène FLAMEIN : helene.flamein2@gmail.com

Titre : Étude de la perception d'une ville : repérage automatique, analyse et visualisation

Mots-clés : lieu, perception, ESLO, traitement automatique des langues, visualisation de l'information.

Title: *Study of the Perception of a City: Automatic Identification, Analysis and Visualization*

Keywords: *location, perception, ESLO, natural language processing, information visualization.*

Thèse de doctorat en sciences du langage, Laboratoire Ligérien de Linguistique, UMR 7270, UFR Lettres, Langues et Sciences Humaines, Université d'Orléans, sous la direction de Iris Eshkol-Taravella (Pr, Université Paris Nanterre) et Gabriel Bergounioux (Pr, Université d'Orléans). Thèse soutenue le 10/12/2019.

Jury : Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, codirectrice), M. Gabriel Bergounioux (Pr, Université d'Orléans, codirecteur), M. Olivier Baude (Pr, Université Paris Nanterre, président), M. Thierry Poibeau (DR, CNRS, rapporteur), M. Mathieu Roche (CR HDR, CIRAD, rapporteur).

Résumé : *À l'heure où de plus en plus de corpus et de données sont accessibles, le travail initié s'interroge sur l'exploitation de données linguistiques dans un corpus oral à dimension sociolinguistique avec l'objectif d'en extraire automatiquement du contenu subjectif. À partir de l'exploitation du corpus de transcriptions de français parlé ESLO (enquête sociolinguistique à Orléans), l'objectif est de modéliser, détecter et visualiser la perception qu'ont les locuteurs de la ville d'Orléans. Pour cela, une approche pluridisciplinaire associant la linguistique, le traitement automatique des langues (TAL) et la géographie a été suivie.*

Après avoir discuté la définition du lieu, la première partie de la thèse décrit la méthodologie employée pour la détection de ce type d'information. La démarche proposée suit une approche symbolique fondée sur des lexiques et des règles élaborées grâce à une analyse approfondie des lieux et de leur nommage. Si des normes existent pour le nommage des lieux, tout individu est à même de faire varier cette norme et de se référer aux espaces de son environnement par des moyens détournés. Dans ce travail, les lieux sont identifiés tels qu'ils sont nommés, c'est-à-dire sous la forme utilisée quotidiennement par les gens. Le module développé est évalué et obtient une F-mesure de 0,91 avec un rappel de 0,90 et une précision de 0,93.

La deuxième partie de la thèse apporte un éclairage sur la notion de subjectivité et surtout sur l'articulation des émotions, des sentiments et des opinions avec la notion de perception. À partir de la détection des lieux, les transcriptions sont analysées par apprentissage automatique supervisé afin d'identifier leur caractère subjectif ou objectif ainsi que leur polarité positive ou négative. Les expériences menées ont permis d'entraîner un modèle obtenant une macro-average *de 0,77 pour la tâche de détection de la subjectivité et de 0,76 pour celle de détection de la polarité dans les segments analysés. La détection de la subjectivité et de la polarité oriente l'analyse de la perception, mais ne suffit pas à en rendre compte. Pour aller plus loin, des propositions typologiques sont réalisées au sujet de la cible de la perception et de la manière dont les locuteurs font part de leur perception.*

Afin de confronter les segments subjectifs extraits des transcriptions avec les différents éléments détectés à leur sujet, les résultats obtenus sont projetés dans un système d'information géographique (SIG). Cette visualisation cartographique permet d'avoir une

vision synthétique de l'information, mais aussi de créer de la connaissance en passant du texte à l'image. La troisième et dernière partie décrit les méthodes employées pour la visualisation de la perception de la ville d'Orléans par les locuteurs du corpus ESLO.

Finalement, la carte finale obtenue offre une nouvelle manière d'accéder au corpus ESLO qui se présente comme le portrait sonore de la ville d'Orléans. La matérialisation de ce portrait de la ville d'Orléans ancre d'une part la dimension patrimoniale et anthropologique du corpus, et d'autre part, le témoignage qu'il représente.

URL où le mémoire peut être téléchargé :
http://theses.fr/s270349

Cédric GENDROT : cedric.gendrot@sorbonne-nouvelle.fr

Titre : Traitement automatique et analyse de la variation dans la parole : des mesures phonétiques sur grands corpus aux réseaux de neurones profonds

Mots-clés : phonétique, phonologie, TAL, traitement automatique des langues.

Title: *Automatic Processing and Analysis of Variation in Speech: From Phonetic Measurements on Large Corpora to Deep Neural Networks*

Keywords: *phonetics and phonology, NLP, natural language processing.*

Habilitation à diriger des recherches en sciences du langage, Dynamique du Langage, UMR 5596, Institut des Sciences de l'Homme, Université Lumière Lyon 2, sous la direction de François Pellegrino (DR, CNRS). Habilitation soutenue le 08/07/2021.

Jury : M. François Pellegrino (DR, CNRS, directeur), M. Laurent Besacier (Pr, Université Grenoble Alpes, président), Mme Ann Bradlow (Pr, Northwestern University, Evanston, Illinois, États-Unis, rapporteuse), Mme Corinne Fredouille (MC, Avignon Université, rapporteuse), M. Kim Gerdes (Pr, Université Paris-Saclay, examinateur), Mme Christine Meunier (DR, CNRS, rapporteuse).

Résumé : *Dans ce document d'habilitation à diriger des recherches sont présentées mes activités pédagogiques et académiques, ainsi que mes activités de recherche depuis mon recrutement en tant que maître de conférences à l'Université Sorbonne Nouvelle en 2006. Ce résumé se concentre sur le dernier point en suivant le fil rouge de mes travaux : l'utilisation de grands corpus de parole non préparée pour des analyses phonétiques automatiques afin de mieux comprendre la variation présente dans la parole.*

Dans la première section, après avoir présenté des valeurs formantiques de référence pour le français, j'ai montré des phénomènes de réduction acoustique pour toutes les voyelles en fonction de leur durée phonétique, du contexte consonantique et du style de parole. Cette réduction s'observe également dans plusieurs langues avec des contraintes phonologiques différentes. Il a été démontré au cours de ces travaux que

des mesures effectuées de façon automatique sur des corpus alignés automatiquement restent cohérentes à la condition de respecter certaines précautions méthodologiques.

Dans la deuxième section, les travaux présentés ont mis en évidence l'importance de la prosodie sur la réalisation acoustique des voyelles. La position dans le mot, le syntagme accentuel et le syntagme intonatif sont trois facteurs de variation récurrents que l'on observe en français, en allemand et en espagnol. La comparaison entre trois langues aux systèmes accentuels différents m'a permis de séparer la structure accentuelle et la structure prosodique, pouvant être mises en avant respectivement soit par des informations spectrales (formants) de façon prépondérante, soit par des paramètres prosodiques (f0 et durée).

Dans la troisième section, je me suis appliqué à traiter des phénomènes linguistiques dont la variation soulève des questions sur la séparation entre phonétique et phonologie. J'ai pu montrer dans le cadre de l'analyse du schwa que la prise en compte de multiples facteurs était possible et souhaitable dans de grands corpus. La mise en évidence de variables différentes pour la réduction du schwa par rapport à son élision complète a permis de conclure à des mécanismes différents, l'un phonétique et l'autre phonologique. L'analyse du /R/ français standard d'après une combinaison de corpus de données articulatoires et de grands corpus de parole a permis de considérer la forme non voisée du /R/ comme la réalisation hyper-articulée de la forme voisée, et a montré que la variation du /R/ est grandement influencée par la position prosodique et par le style de parole, en plus du contexte consonantique. Pour finir, dans une étude postulant que /e/ et /ɛ/ sont entrés dans un processus de fusion, j'ai montré que les grands corpus avec de multiples locuteurs sont des outils appropriés pour repérer des tendances globales dans une langue malgré le maintien de variations inter-locuteurs. Ces études ont également été l'occasion de tester perceptivement les variations mesurées et ainsi valider leur pertinence dans le cadre de la communication parlée. Plusieurs aspects méthodologiques fondamentaux ainsi que des méthodes innovantes sont présentés.

Dans la quatrième et dernière section, une discussion est proposée : l'utilisation des grands corpus y est comparée à celle des petits corpus de parole lue. Une remise en question des méthodes tant pour les données que pour les analyses est également avancée et des solutions sont proposées. Mes travaux récents m'ont guidé vers la recherche de stratégies propres au locuteur et de sa caractérisation phonétique. Depuis moins de dix ans, les réseaux de neurones profonds ont bouleversé le domaine de la classification, et il paraissait indispensable d'essayer de les utiliser pour l'analyse phonétique. En ayant recours à des réseaux de neurones convolutifs (CNN) par le biais de spectrogrammes, le but était double : (1) savoir jusqu'à quel point le spectrogramme permet de caractériser le locuteur au-delà d'une analyse phonétique classique et (2) au moyen de techniques de visualisation, parvenir à localiser les zones du

spectrogramme utilisées par les CNN. Des résultats encourageants présentés dans la discussion finale donnent un aperçu de mes projets de recherche.

URL où le mémoire peut être téléchargé :
https://halshs.archives-ouvertes.fr/tel-03303801

Marine WAUQUIER : marine.wauquier@hotmail.fr

Titre : Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels

Mots-clés : sémantique distributionnelle, morphologie, sémantique lexicale, linguistique de corpus, nominalisation.

Title: *Confrontation of Derivational Processes and Semantic Categories in Distributional Semantic Models*

Keywords: *distributional semantics, morphology, lexical semantics, corpus linguistics, nominalization.*

Thèse de doctorat en sciences du langage, CLLE-ERSS, UMR 5263, Université Toulouse 2 - Jean Jaurès, sous la direction de Nabil Hathout (DR, CNRS). Thèse soutenue le 04/12/2020.

Jury : M. Nabil Hathout (DR, CNRS, directeur), M. Olivier Bonami (Pr, Université de Paris, rapporteur), M. Ingo Plag (Pr, Université de Düsseldorf, Allemagne, rapporteur), Mme Fiammetta Namer (Pr, Université de Lorraine, présidente), M. Laurent Prévot (Pr, Université d'Aix-Marseille, examinateur), Mme Cécile Fabre (Pr, Université Toulouse 2 - Jean Jaurès, examinatrice).

Résumé : *La forme et le sens sont intimement liés en morphologie dérivationnelle, l'affixe d'un dérivé renseignant généralement sur son appartenance à une catégorie sémantique donnée. Cette relation entre affixes et catégories sémantiques n'est cependant pas exclusive, et est étudiée à partir de facteurs phonologiques, syntaxiques, ou encore sémantiques. Ces derniers sont sans doute parmi les facteurs les plus difficiles à évaluer empiriquement, et ont longtemps reposé sur une approche intuitive.*

La sémantique distributionnelle se révèle depuis quelques années comme une des alternatives les plus populaires. Il s'agit d'une approche statistique du sens basée sur les usages en corpus, qui offre une représentation vectorielle du sens des mots. La quantification de la proximité sémantique des mots et la manipulation des représentations permises par les modèles distributionnels ouvrent de nouvelles perspectives sur l'analyse sémantique de la concurrence affixale.

Nous mettons à profit dans cette thèse les modèles distributionnels pour analyser des dérivés morphologiques au regard de ces relations many-to-many*, selon quatre axes. Dans un premier temps, nous quantifions la proximité sémantique entre membres de familles dérivationnelles à l'aide de la proximité distributionnelle dans les espaces*

vectoriels, validant à grande échelle l'hypothèse d'une plus grande proximité du verbe et du nom d'action. Dans un second temps, nous étayons les différences sémantiques entre les noms en -eur, -euse *et* -rice *relatives aux propriétés axiologiques de leurs référents, en comparant les représentations globales de ces trois classes. Dans un troisième temps, nous évaluons l'hétérogénéité morphologique et sémantique de la catégorie lexicale des noms d'agent à partir de l'analyse de la représentation globale de ses représentants prototypiques. Enfin, nous explorons la différenciation sémantique des noms d'action en* -age, -ion *et* -ment, *au regard de leur degré de technicité. Nous combinons des indices distributionnels et statistiques afin de modéliser cette différence de technicité.*

Au travers de ces quatre questions, cette thèse présente différents degrés d'adaptation des modèles distributionnels pour l'analyse linguistique, en tant qu'outil de validation et d'exploration. Nous proposons à ce titre une exploration méthodologique visant à illustrer le potentiel, mais aussi les limites de l'utilisation des modèles distributionnels en linguistique.

URL où le mémoire peut être téléchargé :
http://www.theses.fr/s196917
