NNT/NL : 0000AIXM0000/000ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 14 décembre 2022 par

Rita Hijazi

Simplification syntaxique de textes à base de représentations sémantiques exprimées avec le formalisme Dependency Minimal Recursion Semantics (DMRS)

Discipline
Sciences du Langage

École doctorale
ED 356 – Cognition, Langage, Éducation

Laboratoire/Partenaires de recherche
Laboratoire Parole et Langage (UMR7309) CNRS-AMU
Laboratoire Informatique et Systèmes (UMR 7020) CNRS-AMU

Composition du jury

Guy PERRIER PR émérite, Université de Lorraine	Rapporteur
Marie CANDITO MCF, Université Paris Cité	Rapporteure
Alexis NASR PR, Aix-Marseille Université	Président du jury
Amalia TODIRASCU PR, Université de Strasbourg	Examinatrice
Bernard ESPINASSE PR, Aix Marseille Université	Directeur de thèse
Núria GALA PR, Aix Marseille Université	Co-directrice de thèse

Affidavit

Je soussignée, Rita HIJAZI, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Bernard ESPINASSE et Núria GALA, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Aix-en-Provence, le 18 octobre 2022

Affidavit

I, undersigned, Rita HIJAZI, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Bernard ESPINASSE and Núria GALA, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place Aix-en-Provence, date 18 octobre 2022

Liste de publications et participation aux conférences

1) Liste des publications réalisées dans le cadre du projet de thèse :

    1. (ACTN) Hijazi, R. (2020). Transformations syntaxiques entre niveaux de simplification dans le corpus Newsela. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL* (pp. 137-150). ATALA; AFCP. Nancy (France).

    2. (ACTI) Hijazi, R., Espinasse, B, Gala, N. (2022). GRASS: A syntactic text simplification system based on semantic representations. In *11th International Conference on Natural Language Processing (NLP 2022).* Copenhagen (Denmark). https://nlp2022.org/

2) Participation à des conférences et évènements scientifiques au cours de la période de thèse :

    1. 11th *International Conference on Natural Language Processing* (NLP 2022) : présentation orale.

    2. Journées du Laboratoire Parole et Langage (LPL 2022), Agay : Poster.

    3. Journée Annuelle des Doctorants (JAD 2021), Laboratoire Parole et Langage, Aix-en-Provence : Poster.

    4. GDR LIFT 2021 (Linguistique Informatique, Formelle et de Terrain), Grenoble : Poster.

    5. 22e *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues* (RÉCITAL, 22e édition, 2020) : présentation orale.

Résumé

La simplification de textes consiste à transformer un texte en une version plus simple à lire et/ou à comprendre et plus accessible à un public cible, tout en conservant son information, son contenu et son sens originaux. On distingue quatre niveaux de simplification, *lexical*, *syntaxique*, *morphologique* et *discursif*, et différents systèmes de Simplification Automatique de Textes (SAT) ont été développés en tenant compte de ces niveaux. Cette thèse se concentre sur la *simplification syntaxique de textes en anglais*, une tâche pour laquelle les systèmes automatiques existants présentent certaines limites.

Pour les dépasser, nous proposons tout d'abord une nouvelle *méthode de simplification syntaxique* exploitant des *dépendances sémantiques* exprimées en DMRS (*Dependency Minimal Recursion Semantics*), une représentation sémantique profonde sous forme de graphes combinant sémantique et syntaxe. La simplification syntaxique consiste alors à représenter la phrase complexe en un graphe DMRS, transformer selon des stratégies spécifiques ce graphe en d'autres graphes DMRS qui généreront des phrases plus simples. Cette méthode permet la simplification syntaxique de constructions complexes, en particulier des opérations de division basées sur des appositives, sur des coordinations et sur des subordinations ; ainsi que la transformation de formes passives en formes actives.

Pour évaluer cette méthode, nous avons développé un *système automatique de simplification syntaxique*. Ce système, nommé GRASS, met en œuvre les stratégies spécifiques de transformation de graphes DMRS par des ensembles de règles en utilisant le système de réécriture de graphe GREW. Ce système de simplification automatique est évalué sur un corpus de référence spécifique à la simplification syntaxique de façon automatique puis en ayant recours à des experts humains.

Les résultats obtenus par ce système de simplification syntaxique sur ce corpus de référence sur les opérations de division de phrases surpassent ceux des systèmes existants du même type dans la production de phrases simples, grammaticales et conservant le sens, démontrant ainsi tout l'intérêt de notre approche de la simplification syntaxique à base de représentations sémantiques en DMRS.

Mots clés : Simplification Automatique des Textes, simplification syntaxique, représentation sémantique profonde, Dependency Minimal Recursion Semantics, DMRS, interface syntaxe-sémantique

Abstract

Text simplification is the task of making a text easier to read and understand and more accessible to a target audience. This goal can be reached by reducing the linguistic complexity of the text while preserving the original meaning as much as possible. There are four levels of simplification, *lexical*, *syntactic*, *morphological* and *discursive*, and different Automatic Text Simplification systems (ATS) have been developed taking these levels into account. This thesis focuses on the *syntactic simplification of texts in English*, a task for which these automatic systems have certain limitations.

To overcome them, we first propose a new *method of syntactic simplification* exploiting *semantic dependencies* expressed in DMRS (*Dependency Minimal Recursion Semantics*), a deep semantic representation in the form of graphs combining semantics and syntax. Syntactic simplification enables to represent the complex sentence in a DMRS graph, transforming this graph according to specific strategies into other DMRS graphs, which will generate simpler sentences. This method allows the syntactic simplification of complex constructions, in particular division operations such as subordinate clauses, appositive clauses, coordination and also the transformation of passive forms into active forms.

To evaluate this method, we developed an automatic *syntactic simplification system*. This system, named GRASS, implements the specific DMRS graph transformation strategies of the method by sets of rules using GREW, a Graph REWriting system. This automatic simplification system is evaluated on a reference corpus specific to syntactic simplification by human experts.

The results obtained by this system of syntactic simplification surpass those of the existing systems of the same type in the production of simple, grammatical sentences and preserving the meaning, thus demonstrating all the interest of our approach to syntactic simplification based on semantic representations in DMRS.

Keywords: automatic text simplification, syntactic simplification, deep semantic parsing, meaning representation, Dependency Minimal Recursion Semantics, DMRS.

Remerciements

Je tiens tout d'abord à remercier sincèrement Guy Perrier et Marie Candito de m'avoir fait l'honneur d'accepter de rapporter mon manuscrit de thèse. Leur lecture fine et leurs remarques pertinentes ont abouti à l'amélioration et à l'enrichissement du contenu du manuscrit. Je remercie également Alexis Nasr et Amalia Todirascu d'avoir accepté de participer à mon jury de thèse.

Je remercie chaleureusement Bernard Espinasse et Núria Gala, mes directeurs de thèse. Je tiens à leur exprimer ma gratitude, non seulement pour leur encadrement, mais aussi pour leur présence permanente, leur patience, leurs encouragements, leur soutien et leur gentillesse dont j'ai grandement bénéficié tout au long de ces quatre ans de thèse. Bernard, merci pour ton humanité, ta confiance en moi accordée dès le départ, ta positivité et bonne humeur à toute épreuve. Tout cela a été un moteur inépuisable de motivation durant cette thèse. Núria, merci pour ton aide à tous les niveaux, ta grande disponibilité. Merci de m'avoir appris à chercher plus pour trouver des réponses, pour tes explications et tes commentaires avisés sans lesquels ce travail n'aurait pu aboutir.

Je remercie Bruno Guillaume et Guy Perrier pour le séjour de recherche à Nancy sur l'outil GREW. Merci pour votre temps précieux, vos nombreuses explications et vos solutions qui m'ont aidée à maintes reprises à me débarrasser des blocages rencontrés avec GREW. Je remercie également Bastien Gastinel et Hamza Ghorfi pour leur aide précieuse.

Merci à l'Université d'Aix-Marseille pour le contrat doctoral de courte durée, au Laboratoire Parole et Langage (LPL) et au Laboratoire Informatique et Systèmes (LIS) pour les différents financements reçus lors de mes déplacements pour les conférences et autres événements dans le cadre de la recherche. Merci à Adrian CHIFU, à Sébastien Fournier et au CROUS pour leur aide financière.

Je remercie les collègues du LIS et du LPL de m'avoir fourni un environnement scientifique et humain de grande qualité pendant quatre ans. Je remercie en particulier Yassine, Dima et Elisa. Avec Yassine, j'ai partagé le même bureau, mais aussi des larmes et des rires. Merci à vous trois pour votre soutien et vos encouragements. Les jours étaient moins lourds avec vous.

Ma famille, votre présence m'a accompagnée tout au long de ce long voyage qui a duré quatre ans. Vous avez toujours été là, malgré la distance. Auprès de vous, j'ai toujours trouvé une écoute attentive, un soutien moral et spirituel inconditionnel, des paroles encourageantes et la motivation pour retrouver le sourire et me remettre au travail quand il fallait. Je ne saurais jamais exprimer véritablement ma gratitude. Je remercie mes

parents : Jamal et Safaa. Merci pour votre écoute, patience et aide au niveau moral et financier. Vous étiez toujours là pour moi. Merci maman pour ton soutien, tes encouragements et tes prières. Merci à mes frères Malek et Mohamad et à ma sœur May pour leur soutien et paroles réconfortantes. Vous avez tant donné pour cette thèse, je vous la dédie. Merci beaucoup, sans vous, je n'en serai pas là ! Je suis contente de faire votre fierté et de vous honorer par l'obtention de ce doctorat.

Enfin, merci à mon très cher Hussein. Merci pour ta patience, ta compréhension et pour ton soutien moral indéfectible.

Table des matières

Introduction

Contexte et motivations

Les études en lisibilité des textes remontent aux années 1940 et se sont traduites par des formules qui évaluent le degré de facilité de lecture sur la base de variables lexicales, morphosyntaxiques et typographiques. Ces formules ont pour but de donner une représentation statistique de la facilité de lecture, évaluant dans le texte des facteurs tels que la complexité et la forme des phrases, la familiarité du vocabulaire, la longueur des mots, la répartition entre mots concrets et abstraits, etc. (Gray and Leary, 1935 ; Pearson, 1974). D'après Sorin (1996), il existe à peu près 200 formules de lisibilité en usage. Les problèmes de compréhension s'expliquent souvent par une complexité des textes. Les difficultés de compréhension peuvent être liée à divers facteurs linguistiques comme la familiarité avec le vocabulaire utilisé ou la complexité de la forme syntaxique de la phrase. Différents domaines de recherche s'intéressent à la difficulté de compréhension d'une phrase (Yan, 2021) : dans le domaine de l'acquisition et l'apprentissage des langues, les phrases de différents niveaux de difficulté peuvent être des ressources 1) pour l'évaluation du degré d'acquisition d'une langue 2) pour l'évaluation des compétences linguistiques dans des situations de troubles du langage et 3) elles peuvent servir comme support aux enseignants dans la préparation de cours ainsi que pour évaluer les résultats d'apprentissage. Parmi les nombreux aspects qui influencent la difficulté de compréhension des phrases, notre thèse s'intéresse à la *complexité syntaxique*.

Les langues contiennent des constructions complexes et pour beaucoup de tests de lisibilité, les phrases qui comportent une grande proportion de propositions subordonnées ou un grand nombre d'enchâssements de subordonnées sont considérées comme complexes. Les phrases longues avec une structure complexe (énumérations, enchâssements, etc.) ne sont pas seulement un obstacle à la compréhension par les humains, mais elles peuvent aussi poser un problème pour les outils et les applications qui traitent le langage naturel (Chandrasekar et al., 1996), tels que l'analyse syntaxique (Tomita, 1985 ; McDonald et Nivre, 2011 ; Chandrasekar et al., 1996), la traduction automatique (Gerber et Hovy, 1998 ; Chandrasekar, 1994), l'extraction d'informations (Beigman Klebanov et al., 2004 ; Evans, 2011), et l'étiquetage des rôles sémantiques (Vickrey et Koller, 2008). Dans toutes ces études, la simplification des phrases a été suggérée comme étape de prétraitement afin d'améliorer les performances de ces systèmes.

Les phrases les plus longues sont généralement composées de plusieurs propositions, reliées par diverses relations de coordination ou de subordination. Ils contiennent souvent des phrases verbales ou nominales complexes avec des énumérations, des propositions relatives, des parenthèses, etc. Afin de faciliter la tâche de traitement de

telles phrases, il est souhaitable de les simplifier structurellement en phrases plus courtes et plus simples tout en préservant le sens et les informations qu'elles contiennent.

La simplification de textes (ST) a pour objectif de réduire la complexité linguistique d'un texte, tout en conservant les informations et le sens d'origine. Il s'agit d'un domaine diversifié avec différents publics cible par exemple, les apprenants de langues étrangères (Petersen et Ostendorf, 2007), les personnes ayant une déficience intellectuelle (Feng, 2009 ; Saggion et al., 2011), et les personnes ayant des troubles du langage comme les personnes atteintes d'autisme (Martos et al., 2012), d'aphasies (Carroll et al., 1998 ; Devlin, 1998) et dyslexie (Rello, 2014). Pour chacun de ces publics, les objectifs de simplification sont différents. Par exemple, pour les personnes souffrant d'autisme la simplification vise la réduction de la complexité syntaxique (Martos et al., 2012), alors que pour les débutants en lecture, la recherche en ST a tenté de fournir des méthodes pour réduire la longueur et l'enchâssement des phrases et la simplification lexicale (De Belder et Moens, 2010).

Depuis les années 1990, diverses initiatives ont proposé des lignes directrices pour la rédaction de textes faciles à lire (Freyhoff et al., 1998 ; PlainLanguage, 2011 ; Gala et al., 2020a). Cependant, la simplification du matériel écrit existant par des éditeurs humains est à la fois très coûteuse et prend du temps, en particulier dans le cas d'articles d'actualité qui sont constamment générés. Par conséquent, de nombreuses tentatives ont été faites pour automatiser complètement ou au moins partiellement ce processus.

Plusieurs approches de simplification automatique du texte (SAT) ont été proposées pour l'anglais (Siddharthan, 2006 ; Zhu et al., 2010 ; Woodsend et Lapata, 2011), pour l'espagnol (Saggion et al., 2011), pour le français (Brouwers et al., 2014), pour le portugais (Aluísio et al., 2008) et pour l'italien (Barlacchi et Tonelli, 2013). Cependant, les approches statistiques et neuronales utilisées récemment dans la simplification de textes sont confrontées à plusieurs problèmes.

En simplification syntaxique, décider *quand* et *où* diviser une phrase en deux phrases plus simples, associées à des évènements distincts, peut être déterminé par la syntaxe même, par exemple, des phrases avec des subordonnées. Par exemple, Shieber et Schabes (1991) utilisent des grammaires synchrones qui spécifient des opérations de transformation entre les arbres syntaxiques. Cependant, la simplification syntaxique basée uniquement sur la syntaxe (Zhu et al., 2010 ; Woodsend et Lapata, 2011) n'arrive souvent pas à reconstruire correctement l'élément partagé, ni à identifier correctement le point de découpage. Des informations supplémentaires s'avèrent utiles, par exemple, lors de l'examen de la portée et de la nature des modifieurs adverbiaux. Les représentations syntaxiques ne font pas la distinction entre modifieurs adverbiaux modifiant un verbe et ceux modifiant la proposition entière. La considération de constituants sémantiques en plus des constituants syntaxiques rend le découpage des phrases plus pertinent, notamment pour les tâches qui se soucient de la compréhension du texte, par ex. la traduction et le résumé automatiques (Muszyńska, 2016). Cependant, les constituants sémantiques ne sont pas suffisants pour traiter tous les phénomènes syntaxiques. L'idéal est alors de tirer profit des avantages des deux représentations, sémantique et syntaxique.

Dans le cadre de ce travail de thèse, nous avons retenu la représentation sémantique *Dependency Minimal Recursion Semantics* (DMRS ; Copestake, 2009). Notons que la plupart des représentations sémantiques reposent sur des analyses syntaxiques, de sorte qu'il existe un fort chevauchement entre les constituants sémantiques et syntaxiques. Cela s'applique particulièrement pour DMRS, qui est fortement compositionnel par définition.

Objectifs et hypothèses

Dans cette thèse, nous proposons une nouvelle méthode de simplification syntaxique de texte basée sur les dépendances sémantiques et utilisant une grammaire à large couverture en tirant parti de connaissances linguistiques encodées dans cette grammaire, appelée l'*English Ressources Grammar* (ERG ; cf. 3.1.2.2).

Cette thèse se concentre autour de deux questions principales : (1) comment exploiter une représentation sémantique profonde qui unit sémantique et syntaxe pour améliorer la simplification de phrases ? (2) L'approche proposée utilisant des règles développées manuellement est-elle prometteuse pour la tâche de simplification syntaxique automatique des textes ?

Contributions

Nous avons proposé d'utiliser une information linguistique riche sous la forme de représentations sémantiques profondes afin d'améliorer la tâche de simplification des phrases. Nous utilisons la représentation sémantique exprimée en DMRS. La méthode prend la représentation sémantique profonde comme entrée pour appliquer un ensemble de règles de transformation de graphes afin d'obtenir deux (ou plusieurs) graphes DMRS. Les graphes transformés sont générés afin d'obtenir des phrases simplifiées sans modifier leur sens. Nous montrons que les informations sémantiques facilitent la complétion des phrases lors de la réécriture de l'élément partagé dans les phrases divisées. Nous avons comparé notre système par rapport à plusieurs systèmes de référence et nous avons montré que notre approche produit une sortie plus simple et préserve à la fois la syntaxe et la sémantique.

Cette thèse apporte deux contributions principales à la simplification de textes.

1) Tout d'abord nous proposons une nouvelle *méthode de simplification syntaxique* utilisant des informations linguistiques plus riches, les dépendances sémantiques, exprimées en DMRS (*Dependency Minimal Recursion Semantics*), une représentation sémantique profonde sous forme de graphes. La simplification consiste alors à transformer ces graphes DMRS. Cette méthode traite notamment de la simplification syntaxique de constructions complexes, en particulier les opérations de division telles que les constructions appositives, les coordinations et les subordinations ; ainsi que la transformation des formes passives en formes actives. À notre connaissance, nous sommes les premiers à exploiter des *dépendances sémantiques* pour la simplification de phrases.

2) Ensuite, afin de l'évaluer, nous avons développé un *système automatique de simplification syntaxique* implémentant cette méthode. Ce système est basé sur des règles de transformation de graphes DMRS utilisant le système de réécriture de graphe GREW. Ce système de simplification syntaxique est évalué sur un corpus de référence pour la simplification syntaxique de façon automatique et en ayant recours à des experts humains. Les résultats obtenus par ce système de simplification syntaxique sur ce corpus de référence surpassent ceux des systèmes existants du même type dans la production de phrases simples, grammaticales et conservant le sens, démontrant ainsi tout l'intérêt de notre approche de la simplification syntaxique à base de représentations sémantiques en DMRS.

Structure du manuscrit

Ce manuscrit se compose de deux parties. **La première partie** (les trois premiers chapitres) constitue l'état de l'art de la thèse. Elle décrit les fondements théoriques et les notions de base telles que le domaine de Simplification de Textes (ST) en général (chapitre 1) et plus particulièrement le domaine de Simplification Syntaxique (SS) sur lequel nous avons travaillé (chapitre 2). Le deuxième axe est la tâche d'analyse sémantique qui constitue l'élément-clé de notre approche (chapitre 3).

Les trois autres chapitres constituent **la deuxième partie** et développent nos contributions. Elle est constituée de trois sous-parties : la méthode linguistique (chapitre 4), l'implémentation informatique de cette méthode (chapitre 5) et l'évaluation de la méthode proposée (chapitre 6).

Le **Chapitre 1** introduit l'importance et les enjeux en simplification de textes, présentant d'abord la notion de complexité linguistique (section 1.1) et certaines lignes directrices existantes pour l'adaptation de textes et la rédaction de textes faciles à lire pour les personnes ayant des déficiences intellectuelles ou des difficultés de lecture (section 1.2). Nous décrivons dans ce chapitre les domaines d'application (section 1.3) et identifions les différents types de simplification possibles (section 1.4) : lexicale, morphologique, discursive et syntaxique. Ce chapitre décrit également les ressources et corpus existants pour le développement et l'évaluation des systèmes de SAT (section 1.5). La dernière section du chapitre introduit l'intérêt de la simplification dans d'autres tâches de traitement automatique des langues (TAL), telles que le résumé, la création de paraphrases et la traduction automatique tout en présentant les points de divergence et de convergence (section 1.6).

Le **Chapitre 2** présente la simplification syntaxique de textes, les différentes opérations de transformations appliquées lors du passage d'une version complexe à une version simplifiée (section 2.1). Nous dressons un état de l'art des approches de simplification de textes utilisées dans le développement des systèmes de SAT (section

2.2). Ce chapitre compare différentes approches en SAT et attire l'attention sur leurs principales forces et faiblesses (section 2.3).

Le **Chapitre 3** clôt la partie I du manuscrit et présente un aperçu de la relation entre les niveaux syntaxiques et sémantiques, de l'origine des formalismes de modélisation informatique de la syntaxe et les courants qu'ils forment. Il décrit la différence entre les deux niveaux d'analyse syntaxique : de surface (section 3.1) et profond (section 3.2), ainsi que la structure argumentale d'un énoncé (section 3.3). Il introduit la tâche d'analyse sémantique des textes (section 3.4), ainsi que les différents formalismes sémantiques existants, à base de cadres (section 3.5) et ceux à base de graphes (section 3.6), tout en mettant l'accent sur la famille MRS du projet DELPH-IN et notamment sur DMRS qui fait partie de ce projet. La dernière section (section 3.7) de ce chapitre est consacrée à une comparaison entre ces formalismes, leur représentation, leur structure, les informations qu'annote chacun d'eux.

Le **Chapitre 4** présente notre première contribution : une méthode de simplification syntaxique de textes en anglais, qui utilise une représentation sémantique. Nous justifions notre choix d'utiliser une représentation sémantique profonde en entrée, *Dependency Minimal Recursion Semantics* (DMRS) et précisons les avantages d'utiliser une telle représentation par rapport aux d'autres représentations existantes (section 4.1). Nous décrivons notre méthode, les grandes étapes et le système de réécriture de graphes choisi, GREW (*Graph-REWriting system*) (section 4.2). Nous développons comment nous avons procédé à la simplification de certaines constructions syntaxiques de division de phrases telles que *l'apposition* (section 4.3), la *coordination* (section 4.4), la *subordination* (section 4.5), la *relative* (section 4.6), ainsi qu'une opération de substitution morpho-syntaxique qui est *le passage de la voix passive à la voix active* (section 4.7). Pour chacune de ces constructions, nous développons les théories linguistiques et les différentes formes qu'elle peut avoir avant de décrire la procédure ou stratégie de simplification par transformation de graphes DMRS.

Le **Chapitre 5** présente GRASS (*GRAph-based Syntactic Simplification*), le système de simplification syntaxique à base de graphes que nous avons développé et qui implémente la méthode proposée au chapitre précédent. L'architecture générale et les composantes informatiques de ce système sont tout d'abord détaillées (section 5.1). Ensuite, pour chacune des constructions syntaxiques traitées au chapitre précédent, nous présentons les règles de transformations de graphes DMRS mises en œuvre par le système de réécriture de graphes *Graph REWriting* (GREW) (section 5.2) implémentant la stratégie spécifique de la méthode de simplification pour chacune des constructions considérées. L'étape de génération des graphes après transformation est décrite dans la section 5.3. La section 5.4 définie la stratégie d'application des règles. Des captures d'écrans d'exécution de notre système GRASS sont présentés dans la section 5.5.

Le dernier **Chapitre 6** présente l'évaluation de notre système de simplification syntaxique. Nous introduisons le corpus que nous avons utilisé pour le développement de nos règles ainsi que le corpus d'évaluation (section 6.1). Nous montrons les résultats de l'évaluation automatique tout en décrivant les métriques et les outils utilisés (section 6.2). Une campagne d'évaluation manuelle a été menée ; le protocole et les résultats sont décrits dans la section 6.3. Nous analysons les erreurs (le bruit et le rappel) dans la section 6.4. Une section de discussion clôt ce chapitre (section 6.5).

Enfin, nous clôturons ce manuscrit par une **Conclusion Générale**, en rappelant nos contributions, ainsi que leurs limitations, et en évoquant quelques perspectives à ce travail de recherche.

Partie I

1. La simplification de textes

La simplification de textes peut être définie comme l'adaptation d'un texte afin de le rendre plus accessible à un public cible, tout en conservant son contenu sémantique intact (Siddharthan, 2014). Saggion (2017) définit la simplification de textes comme étant le processus de transformation d'un texte en un autre texte qui véhicule le même message, afin de le rendre plus facile à lire et à comprendre par un public cible. Adapter un texte, c'est-à-dire, réduire sa complexité, peut s'avérer utile dans les cas de publics confrontés à des difficultés de lecture et/ou de compréhension, par exemple les enfants apprentis lecteurs (Javourey-Drevet et al., 2022).

La simplification automatique de textes (SAT), vue comme une alternative à la simplification manuelle, a commencé à se développer dans le domaine du traitement automatique des langues (TAL) à la fin des années 90. Initialement, cette tâche ne visait pas à adapter les documents à des humains ; elle était considérée comme une étape de prétraitement destinée à faciliter d'autres tâches de TAL comme l'analyse syntaxique (Chandrasekar et Srinivas, 1997), l'annotation sémantique (Vickrey et Koller, 2008), le résumé automatique (Vanderwende et al. 2007), la recherche et l'extraction d'information (Beigman Klebanov et al., 2004) et la traduction automatique (Mishra et al., 2014 ; Štajner et Popovic, 2016). Ces travaux ont établi la simplification de textes comme axe de recherche et ont inspiré des études ultérieures qui ont abordé la tâche comme une technique destinée à divers groupes cibles. Par exemple, les longues phrases, les noms composés et les longues séquences d'adjectifs comme dans « *twenty-five-year-old blond-haired mother-of-two Jane Smith* » peuvent causer des problèmes aux personnes atteintes d'aphasie et d'autisme (Carroll et al., 1998).

Ainsi, de nombreux projets de recherche ont vu le jour pour des publics variés et dans différentes langues, par exemple les enfants dyslexiques (Rello et al., 2014), les adultes illettrés (Aluísio et Gasperin, 2010), les aphasiques (Devlin and Tait, 1998 ; Carroll et al., 1998), les apprenants d'une langue seconde (Petersen et Ostendorf, 2007) et les personnes atteintes d'autisme (Martos et al., 2012 ; Evans et al., 2014). Ces travaux existent pour plusieurs langues, comme l'anglais (Woodsend et Lapata, 2011), le français (Brouwers et al., 2014), l'italien (Barlacchi et Tonelli, 2013), l'espagnol (Saggion et al., 2011), le portugais (Aluísio et al., 2008) et le japonais (Inui et al., 2003).

Du point de vue de la construction des phrases, la simplification peut être considérée comme une forme de transformation de texte impliquant cinq types d'opérations majeures, telles que : la division, la suppression, la substitution, le rajout et le déplacement. Considérons l'exemple ci-dessous, tiré du corpus Newsela[1] (Xu et al., 2015), un corpus en anglais de 1,130 textes, où quatre niveaux de simplification ont été produits manuellement par les éditeurs. (a) est la phrase originale et (b) correspond à la version

[1] https://newsela.com/

simplifiée avec le niveau de simplification (*Simp_2*). Les substitutions sont marquées en bleu, les suppressions en rouge et les rajouts d'information (majoritairement des explications) en vert.

a) Brown, who started out in the energy industry, sees mock meat not only as a much healthier alternative but also as a way to reduce methane and other harmful gases emitted by animals and to ease a meat shortage predicted to occur by midcentury.
b) Brown started out in the energy industry. To him, mock meat is not only much healthier than the real thing. It is also a way to reduce methane and other harmful gases emitted by animals. Some of these gases are believed to cause climate change. The effect of those gases could change weather patterns and nature. It can also ease a meat shortage predicted to occur by midcentury. *(Simp_2)*

Afin d'identifier les constructions syntaxiques complexes pour les simplifier, il nous semble important de comprendre d'abord la notion de *complexité linguistique* (section 1.1), sa relation avec la compréhension et la lecture et les éventuels obstacles linguistiques qui doivent être transformés afin de rendre le texte plus facile à comprendre pour les humains et plus facile à traiter pour les machines. Par la suite, nous présentons les lignes directrices existantes qui aident les rédacteurs à produire un matériel lisible qui répond aux besoins de lecteurs en difficultés afin de faciliter la compréhension et les initiatives d'adaptation de textes (section 1.2), ainsi que les domaines d'application, notamment les domaines de l'administration publique, médical et pharmacologique (section 1.3). Ensuite, nous décrivons les différents niveaux de simplification, lexicale, morphologique, discursive et syntaxique tout en se basant sur les études antérieures qui analysent des corpus (originaux et simplifiés manuellement) pour différentes populations cibles et dans différents domaines, dans le but de définir et de déterminer les transformations nécessaires à la simplification de textes (section 1.4). La section 1.5 est consacrée aux différents corpus et ressources de simplification existants. La dernière section (section 1.6) décrit la relation entre la simplification de textes et autres techniques de transformation de textes en TAL, telles que la traduction, le résumé et la création/détection de paraphrases.

1.1. Complexité linguistique

Du point de vue linguistique, la complexité est avant tout un phénomène difficile à définir (Kusters et Muysken, 2001). À notre connaissance, il n'y a pas une définition consensuelle de la notion de *complexité linguistique*.

Rescher (1998) trouve que la complexité d'un système est fonction « du nombre et de la variété de ses éléments constituants et de la richesse des interrelations, organisationnelles ou opérationnelles, qu'il contient ». Romero (2013) s'appuie sur la notion « intuitive » de difficulté et sur la définition donnée par le dictionnaire Larousse, pour la définir ainsi :

« Dans ces domaines, la complexité peut être mise en rapport avec la notion de difficulté pour le locuteur (…) un phénomène sémantique qui semble intuitivement complexe – au sens courant du terme (…) Le premier sens que donne le Petit Larousse de complexe est : 'Qui se compose d'éléments différents, combinés d'une manière qui n'est pas immédiatement saisissable » (Romero, 2013 : 172-173).

De leur côté, Vecchiato et Gerolimich (2013 : 84) s'appuient sur la définition du Trésor de la langue française qui définit l'adjectif « complexe » comme suit :

« Composé d'éléments qui entretiennent des rapports nombreux, diversifiés, difficiles à saisir par l'esprit, et présentant souvent des aspects différents ».

La difficulté pour les auteurs demande un effort cognitif (2013:118) : « l'emploi du style nominal ; dans ce cas, la compréhension de l'énoncé requiert un effort cognitif majeur, qui en ralentit le décodage ». Pour eux, la complexité se trouve dans les unités linguistiques et leur ordre (2013:125) :

« Nous pouvons considérer la complexité linguistique sur un continuum, allant :

a) *d'un emploi quelque peu complexe : l'emploi d'une structure « composée » d'une suite de subordonnées, ce qui peut poser des problèmes mémoriels au lecteur.*
b) *d'un emploi assez complexe : l'emploi du style « nominal » ; dans ce cas, la compréhension de l'énoncé requiert un effort cognitif majeur, qui en ralentit le décodage.*
c) *d'un emploi très complexe, où l'opacité lexicale de la terminologie est importante ».*

Pour Ibrahim (2013), il y a deux critères pour décrire qu'une telle unité est complexe ou simple : celui de la *fréquence* (qui rend l'accès facile et par suite la rend simple à utiliser) et celui de la *composition* (une forme indécomposable en unités ou formes plus petites est une forme simple).

Dans le domaine d'apprentissage des langues seconde, la littérature réfère fréquemment à la complexité *cognitive* et à la complexité *linguistique*. La complexité *cognitive* réfère aux difficultés éprouvées par les apprenants lors de l'acquisition ou de la production de la L2 alors que la complexité *linguistique* est indépendante de l'apprenant et réfère aux propriétés formelles et fonctionnelles du système ou de sous-systèmes de la L2 (Housen et al., 2012). Selon Monville-Burston (2013), la complexité *cognitive* se définit dans la perspective de l'apprenant : elle fait partie de son expérience personnelle de l'apprentissage d'une langue. C'est le résultat de facteurs subjectifs psychologiques ou sociaux (mémoire, aptitudes, motivation, etc.), ce qui en fait une propriété variable mais elle englobe aussi la complexité linguistique, à savoir la complexité inhérente des traits de la L2 à s'approprier, et en cela elle est une propriété objective. La complexité linguistique (qui nous développons ici) est déterminée par la présence d'un marquage plus ou moins spécifique, plus ou moins élaboré ou encore plus ou moins contraint et comprend

différentes sous-catégories, dont la complexité lexicale et la complexité structurelle (qui nous intéresse ici).

Par l'expression « *complexité linguistique* », on pose les questions suivantes : quels sont les facteurs qui déterminent qu'un texte ou un énoncé est complexe ? quelle est la différence entre un texte complexe et un texte difficile ? Ces questions posent problème dans le domaine à cause de la subjectivité et la multi-dimensionnalité des notions de complexité et difficulté.

La problématique de la complexité linguistique a été développée par Gala et al. (2014). Les auteurs affirment que la notion de complexité lexicale est perçue différemment en fonction du public qui y est confronté (apprenants de langue seconde, personnes avec difficulté intellectuelle, etc.). La notion de « complexité » est, ainsi, multidimensionnelle. Pareil pour la notion de difficulté qui est subjective et dépend de l'interaction de nombreux facteurs, notamment sociaux (situation de lecture), individuels (lecteur) et linguistiques (texte). Les besoins de simplification et les notions subjectives de difficulté du texte sont très spécifiques non seulement entre groupes de personnes (ex. dyslexiques par rapport aux apprenants de langues étrangères), mais aussi entre individus et en intra-groupe (Gala et al., 2014), d'où, le profil du destinataire est essentiel : en fonction de son profil (enfant/adulte/âgé, normolecteur/faible-lecteur/dyslexique, entendant/sourd, etc.) certaines constructions seront plus difficiles que d'autres. Là aussi, une distinction entre production (parole/écriture) et analyse (écoute/lecture) entraînera des difficultés différentes.

Les problèmes de compréhension s'expliquent souvent par une complexité des textes, en particulier au niveau du *lexique*, du *discours* et de la *syntaxe*. Ces facteurs sont connus comme étant des facteurs principaux de difficultés de lecture (Chall et Dale, 1995), en particulier chez les apprenants d'une langue étrangère, les jeunes enfants ou les personnes présentant des déficiences intellectuelles. L'adaptation des textes aide ces personnes à accéder au contenu des documents auxquels ils sont confrontés.

Au niveau de la *complexité lexicale*, plusieurs travaux ont étudié les caractéristiques qui peuvent qualifier un mot comme étant un mot complexe (Gala et al., 2014 ; Soler et al., 2018). Le premier critère est *la fréquence* : plus un mot est fréquent, plus on a de chances de l'avoir déjà rencontré et ainsi, de correctement le décoder et l'interpréter. Les mots peu fréquents rendent le texte difficile à comprendre pour les personnes aphasiques (Devlin, 1999) et dyslexiques (Norbury, 2005 ; Martos et al., 2012). L'utilisation de mots plus fréquents n'améliore pas la compréhension mais réduit le temps de lecture chez les personnes dyslexiques (Rello et al., 2014). En ce qui concerne les étudiants ayant une déficience intellectuelle, les études existantes montrent des résultats contradictoires. Fajardo et al. (2014) n'ont trouvé aucun effet de la fréquence des mots sur les scores de compréhension chez les élèves ayant une déficience intellectuelle. Un autre critère pour identifier des mots « simples » concerne la *familiarité d'un terme* (Gernsbacher, 1984) ou encore *son âge d'acquisition* (Brysbaert et al., 2000). Plus récemment, Gala et al., (2014) expliquent qu'il existe d'autres facteurs pour qualifier un mot complexe autre que la fréquence ou la familiarité, comme l'orthographe, la structure des syllabes, le nombre de morphèmes, la polysémie, etc.

Au niveau du *discours*, les personnes atteintes d'autisme ou de déficience intellectuelle peuvent également rencontrer des difficultés à trouver l'idée principale, à résoudre les anaphores et à déduire des informations (Martos et al., 2012 ; Feng, 2009). De plus, les personnes ayant une déficience intellectuelle ont des problèmes pour traiter et retenir de grandes quantités d'informations (Feng, 2009 ; Fajardo et al., 2014). Plusieurs études ont montré que les textes longs peuvent affecter l'efficacité et la motivation à lire chez les élèves ayant une déficience intellectuelle (Morgan et Moni, 2008). L'étude de Gernsbacher et Faust (1991) a indiqué que les adultes ayant des problèmes de compréhension rencontrent des difficultés de supprimer les informations non pertinentes. Par conséquent, les systèmes de simplification de texte destinés à ces populations cibles devraient non seulement simplifier le contenu écrit, mais devraient également effectuer une sorte de réduction du contenu (en supprimant les informations non pertinentes) afin de réduire la charge de mémoire nécessaire à la compréhension du texte donné.

Hoeks (1999) affirme que lors du traitement d'une phrase dans un contexte de discours, l'information contenue dans la phrase doit être intégrée dans une représentation du discours précédent. Afin d'effectuer cette intégration, le processeur doit d'abord déterminer quelles parties de la phrase font référence à des éléments déjà présents dans la représentation du discours (c'est-à-dire établir une référence). Après cela, les nouvelles informations sur les éléments référents dans la phrase doivent être liées aux référents correspondants dans le discours, ce qui permet d'intégrer ces nouvelles informations dans le modèle de discours (Clark et Sengul, 1979 ; Erteschik-Shir, 1997 ; Haviland et Clark, 1974 ; McKoon et al., 1993).

Au niveau de la *complexité syntaxique*, de nombreuses études sur la relation entre la complexité linguistique et la compréhension en lecture prennent la complexité syntaxique comme variable linguistique. La syntaxe peut être considérée comme un "véhicule" de sens lorsqu'il s'agit de compréhension de la lecture (Scott, 2009). La complexité syntaxique indique la difficulté avec laquelle les lecteurs humains peuvent attribuer une structure syntaxique à une phrase et peuvent utiliser cette structure pour déterminer sa signification (Caramazza et Zurif, 1976 ; Norman et al., 1991 ; Just et al., 1996 ; Meltzer et al., 2009). Elle est un facteur qui accroît la difficulté de déterminer qui a fait quoi à qui dans un énoncé. D'après la Grammaire Méthodique du français (Riegel et al., 1998), une phrase est syntaxiquement complexe si elle comprend plusieurs propositions qui se trouvent en relation de dépendance ou d'association. Selon Jakubowicz (2007) et Jakubowicz et Tuller (2008), la complexité syntaxique augmente avec le nombre et la nature des constructions syntaxiques au sein d'une phrase. Jakubowicz (2007) avance que, plus il y aura de fusions, c'est-à-dire plus il y aura d'assemblages d'éléments lexicaux, plus l'énoncé sera considéré comme complexe. La complexité syntaxique va augmenter avec le nombre d'assemblages d'éléments lexicaux et le nombre de déplacements effectués au sein d'un énoncé.

La notion de *complexité syntaxique* a été définie dans les domaines de la psycholinguistique, la psychologie cognitive et la pédagogie dans des recherches sur l'apprentissage de la lecture pour évaluer la lisibilité des textes destinés aux jeunes lecteurs (Lecocq et al., 1996 ; Boyer, 1992) dans le but de calculer la *densité de*

l'information dans les textes (Chuquet, 2000). Ces mesures visent à calculer la longueur des phrases en fonction du nombre de mots qu'elles contiennent : par exemple, si une phrase contient plus de 20 mots, dans n'importe quel domaine, elle est considérée comme longue et donc complexe (Blanche-Benveniste, 2013a). Selon les études psychologiques sur la mémoire de travail, il existe des limites sur celle-ci : d'après Miller (1956), le nombre d'éléments que la personne peut garder temporairement est limité entre 5 et 9 ; ce nombre a été actualisé à 4 par Cowan (2001). Ainsi, « non seulement des structures de phrases différentes entraînent des degrés de complexité syntaxique différents, mais il y aurait une limite à la complexité syntaxique des phrases. Les structures complexes qui dépassent cette limite ne peuvent pas être comprises par le locuteur » (Yan, 2021).

La notion de complexité est aussi étudiée dans l'apprentissage des langues étrangères. La complexité syntaxique peut être évaluée en examinant le nombre de constituants interconnectés dans une structure, qui est le principe derrière trois mesures telles que la longueur de la phrase, le nombre de phrases par proposition et le nombre de propositions par phrase (Pallotti, 2015). Dans l'apprentissage de l'anglais langue seconde (ALS) par exemple, la dimension syntaxique la plus fréquemment étudiée est la longueur moyenne d'une unité syntaxique ou le nombre de propositions subordonnées dans un texte (Lambert et Kormos, 2014 ; Ortega, 2012), d'où la *densité propositionnelle*. Dans de nombreux cas, la densité propositionnelle d'une phrase est proportionnelle à sa longueur. Caplan et Waters (1999) rapportent que les résultats de nombreuses expériences psycholinguistiques en compréhension de phrases peuvent être expliqués par référence au nombre de propositions véhiculées par les phrases présentées aux lecteurs. En bref, plus le nombre de propositions exprimées dans une phrase est grand, plus il est difficile pour les lecteurs d'effectuer des tâches de mémoire simultanées.

Il existe des constructions syntaxiques qui peuvent être source de complexité. Halliday et al. (2014) et Biber (1991) considèrent que certaines formes grammaticales sont plus difficiles à comprendre que d'autres, comme les verbes à la voix passive, les subordonnées antéposées, l'inversement de sujet et les nominalisations des verbes. Thompson et Shapiro (2007) ont identifié quatre variables qui contribuent à la complexité des phrases, notamment le nombre de propositions, le nombre d'enchâssements, l'ordre dans lequel les éléments apparaissent dans la phrase, qu'ils soient canoniques ou non canoniques et la distance entre les éléments de la phrase. Plusieurs travaux de recherches prouvent que la structure syntaxique SVO est la plus simple (Berman, 1984 ; Skehan, 1991). Parmi les difficultés de compréhension figurent, selon Dumortier (2001), la densité des informations, l'élimination des redondances, la diversité des procédés syntaxiques permettant de multiplier les assertions dans une même phrase, les anaphores fondées sur des inférences et les ruptures thématiques.

La complexité de la phrase est directement liée à la structure grammaticale interne de la phrase (Bram, 1978). Ce n'est cependant pas le seul facteur qui rend une phrase complexe en termes de forme syntaxique de surface. Le niveau de lisibilité d'une phrase donnée est considéré en fonction du nombre de mots qu'elle contient, de la fréquence des mots, le nombre de propositions subordonnées ou de phrases prépositionnelles, la proportion de mots concrets par opposition aux mots abstraits, etc. (Edwards, 1980).

Néanmoins, les formules de lisibilité ne montrent pas pourquoi une phrase donnée est complexe. Comme le soutient Huggins et Adams (2017), les mesures de lisibilité présentent des descriptions statistiques plutôt que structurelles de la complexité ; cependant, la corrélation n'implique pas de prouver la causalité, et bien que les passages très lisibles aient tendance à avoir des phrases courtes et peu de phrases ou de propositions prépositionnelles par phrase, il ne s'ensuit pas que l'écriture de phrases courtes avec peu de phrases prépositionnelles donne un texte très lisible. Ainsi, bien que la longueur de la phrase soit en corrélation avec la complexité syntaxique, elle ne peut pas être utilisée pour l'expliquer. Il est possible d'avoir des phrases longues mais structurellement assez simples et donc faciles à lire. D'autres part, il y a des phrases assez courtes mais dont la structure syntaxique les rend assez difficiles à lire au début (von Glasersfeld, 1970). Comme exemple d'une phrase qui, bien qu'anormalement longue, a une structure grammaticale interne simple. Considérons l'exemple (1) ci-dessous.

(1) [The museum]$_S$ [contains]$_V$ [$_O$ various specimens of tropical fish], [$_O$ the skeletal remains of dinosaurs from the prehistoric age], [$_O$ a variety of butterflies and insects of all shapes and sizes], [$_O$ authentic models of African and Indian elephants and other wild animals], [$_O$ an array of the fossilized remains of plants and small fish] , [$_O$ a very impressive collection of meteorites and rock forms;] and, finally, [$_O$ a display of military equipment and weaponary and the latest developments in laser technology.]

Bien que l'exemple (1) contienne soixante-seize mots, ce n'est pas une phrase difficile à lire. La raison en est que la structure grammaticale de la phrase peut être considérée simple, composée d'un sujet, d'un verbe transitif et de sept objets "groupes nominaux" (énumération). Cette structure crée chez le lecteur une forte attente de ce qui va suivre dans la phrase (c'est-à-dire un (ou plusieurs) objet(s)). En revanche, c'est dans les cas où la syntaxe de la phrase altère cette attente du lecteur que la complexité et la difficulté syntaxiques peuvent survenir dans le texte.

Du point de vue des jeunes lecteurs, l'exemple (2a) poserait plus de difficultés de compréhension que l'exemple (2b). Les exemples (1) et (2) illustrent ainsi l'importance de la forme syntaxique comme facteur influençant le niveau de lisibilité d'une phrase donnée. Cela ne veut pas dire, cependant, que c'est le seul facteur qui influence la lisibilité ; évidemment, la densité sémantique et la facilité de distinguer le sujet de la phrase doivent également jouer un rôle important. Néanmoins, la syntaxe de la proposition est un déterminant premier de sa lisibilité. Dans la phrase (2a), il s'agit de deux syntagmes prépositionnels, un verbe intransitif et un sujet. L'ordre syntaxique de la phrase affiche ainsi deux syntagmes prépositionnels antéposés et une inversion du sujet et du verbe. L'ordre syntaxique le plus courant serait comme dans l'exemple (2b).

(2) a. [$_{SP}$ Down the street] and [$_{SP}$ around the corner] $_V$ran, $_S$Louis.
b. $_S$Louis $_V$ran [$_{SP}$ down the street] and [$_{SP}$ around the corner.]

Pour la plupart des théories, deux opérations sont facteurs de complexité dans l'analyse des phrases (Blache, 2010) : l'intégration d'un nouvel élément à une structure syntaxique et la mémorisation (nombre de dépendances syntaxiques incomplètes). Pour sa part, Yan (2021) utilise trois métriques pour évaluer la complexité syntaxique : (1) la *longueur de dépendance* qui est le nombre de mots entre le gouverneur et son dépendant dans une relation de dépendance ; (2) le *flux d'information* qui représente le nombre d'éléments maintenus simultanément dans la mémoire de travail[2]; 3) le *flux de dépendance* qui est l'ensemble des dépendances qui relient un mot à gauche de cette position à un mot à droite dans une position donnée (entre deux mots dans une phrase) (Kahane, 2001). Il capture l'état du traitement au fur et à mesure que chaque mot de la phrase est traité.

La simplification syntaxique vise à transformer des structures syntaxiques complexes comme les propositions négatives, les formes passives, les tournures impersonnelles et les subordonnées (Brouwers et al., 2014) ; diviser une phrase longue contenant plusieurs propositions (une densité propositionnelle) en plusieurs phrases indépendantes.

À mesure que les structures syntaxiques d'une phrase deviennent plus complexes, le nombre d'analyses possibles augmente, ce qui conduit inévitablement à une ambiguïté accrue et à une plus grande probabilité d'analyse incorrecte (Chandrasekar et al., 1996). Muszyńska (2016) note que certaines approches de l'analyse syntaxique ont des exigences d'espace et de temps avec les phrases longues. Cela peut entraîner des difficultés pratiques de traitement. Par exemple, le processeur ACE exécutant *l'English Resource Grammar* (ERG) (Copestake et Flickinger, 2000) nécessite environ 530 Mo de RAM pour analyser la phrase (3a). En fait, des phrases plus longues et plus compliquées peuvent entraîner l'expiration du temps d'attente de l'analyseur ou un manque de mémoire avant qu'une solution ne soit trouvée. Lorsque la phrase (3a) est divisée en quatre phrases plus courtes (3b), Muszyńska note que chacune des phrases plus courtes peut être analysée avec moins de 20 Mo, nécessitant au total moins d'un cinquième de la RAM nécessaire pour analyser la phrase complète. Les phrases transformées en phrases plus courtes et plus simples avant d'être analysées automatiquement, le niveau d'ambiguïté sera réduit et les performances de l'analyseur devraient devenir plus robustes et moins sujettes aux erreurs. Cette observation fournit une motivation supplémentaire pour le développement d'un outil de simplification des phrases qui réduira la *densité propositionnelle* des phrases d'entrée comme celles contenant des propositions enchâssées.

(3) a. Marcellina has hired Bartolo as her counsel, since Figaro had once promised to marry her if he should default on a loan she had made to him, and she intends to enforce that promise.

[2] L'auteur reconstruit le flux d'informations qui doivent être gardées dans la mémoire de travail lors du traitement des mots. Ces informations sont importantes lorsque la structure de la phrase est linéarisée en une chaîne, ou que l'auditeur essaye de prédire la structure de la phrase entière.

b. Marcellina has hired Bartolo as her counsel. Figaro had once promised to marry her. He should default on a loan she made to him. She intends to enforce that promise.

Bentin et al., (1990) suggèrent que les difficultés de compréhension et de traitement de la syntaxe peuvent être des raisons importantes d'échec en lecture, y compris la fluidité et la compréhension. Plusieurs initiatives d'adaptations de textes ont été développées dans ce but ; nous les mentionnons dans la section suivante.

1.2. Adaptation de textes

Pour les lecteurs humains, le besoin de simplification de textes varie en fonction de la population cible et ses besoins : le niveau de simplification pour les personnes en difficulté est différent de ceux des systèmes automatiques pour les personnes sans difficultés. Deux grandes stratégies ont été explorées pour la simplification manuelle des textes (Gala et al., 2018).

La première consiste à écrire « *from scratch* » à partir de recommandations linguistiques et textuelles. Des guides aident les rédacteurs à produire un matériel lisible qui répond aux besoins de lecteurs en difficultés afin de faciliter la compréhension. Ces lignes directrices (guides) ont été créés afin de normaliser le processus de simplification. Par exemple, on suggère aux écrivains d'augmenter la lisibilité en gardant les phrases à la forme active, privilégier les phrases simples et éviter les nominalisations qui rendent les phrases plus complexes parce qu'elles sont « indirectes » (Klare, 1985 ; Price, 1984). C'est le cas de la collection *La Traversée*[3] à l'initiative de *Lire et Écrire Luxembourg*. Il s'agit d'une collection d'une vingtaine de romans courts s'adressant à des adultes débutants en lecture ou faibles lecteurs.

La seconde stratégie de simplification consiste à transformer un texte original, considéré comme difficile à comprendre, en un texte plus simple, pour un public cible. L'exemple le plus connu est Vikidia[4], un ensemble d'articles avec des contenus simplifiés destinés aux enfants et inspirés des articles de Wikipédia. On peut citer aussi le projet TextToKids[5] qui vise à faciliter l'écriture et le filtrage de textes pour les enfants (la tranche d'âge 7-12 ans). Le consortium, qui réunit des linguistes, des informaticiens et des journalistes spécialisés, cherchera à caractériser les contraintes linguistiques à respecter pour une telle finalité et à proposer des outils d'aide (analyse textuelle automatisée, moteur de recherche, reformulation, bonnes pratiques). En termes de bénéfices, le projet va dans le sens d'un « Internet des enfants » et ouvre la voie à d'autres modalités (parole, images). Dans le contexte de ce projet, Battistelli et al., (2022) étudient la question de la mesure de la complexité d'un texte pour les enfants en âge de lire, au travers de la mise en place d'une chaîne de traitements. Cette chaîne vise à extraire des phénomènes linguistiques, issus de recherches en psycholinguistique et des études sur la lisibilité, mobilisables pour appréhender la complexité des textes. Elle permet d'étudier des

[3] https://lire-et-ecrire.be/latraversee?lang=fr
[4] https://fr.vikidia.org/wiki/Vikidia:Accueil
[5] http://texttokids.irisa.fr/

corrélations entre certains phénomènes linguistiques et les tranches d'âges associées aux textes par des éditeurs. Ces corrélations permettent de valider la pertinence de la catégorisation en âges par les éditeurs. Ainsi, elle justifie la mobilisation d'un tel corpus pour entraîner à partir des âges un modèle de prédiction de l'âge cible d'un texte.

On peut mentionner également des adaptations de type FALC[6] (Facile À Lire et à Comprendre) destinées aux personnes ayant des troubles cognitifs, dans le but d'adapter et simplifier certains textes administratifs, juridiques, ou relevant de la vie quotidienne, dans des secteurs aussi variés que le tourisme, la santé, la culture afin de rendre l'information plus accessible. Par exemple, lors de la pandémie du COVID-19 (printemps 2020), les citoyens français ont dû remplir une attestation pour sortir de leur domicile. Sa version adaptée est représentée dans la Figure 1.

Figure 1 - Exemple de transcription d'un document FALC.
À gauche, le certificat original de l'attestation dérogatoire de déplacement lié au COVID-19. À droite son équivalent FALC. La version simplifiée utilise un langage simple, des phrases plus courtes, une police plus grande et des pictogrammes pour améliorer la lisibilité.

L'amélioration de l'accessibilité à l'aide de la SAT a reçu de plus en plus d'attention avec des initiatives telles que le projet portugais PorSimples (Aluísio et Gasperin, 2010), le

[6] https://www.inspiredservices.org.uk/wp-content/uploads/EN_Information_for_all.pdf

projet espagnol Simplext (Saggion et al., 2011, 2015), le projet belgo-français AMesure (François et al., 2020), ou le projet français ALECTOR (Gala et al., 2020a) (cf. 1.5.1.3) et Cap'FALC (Martin, 2021). Des sites web proposant des informations faciles à lire sont désormais courants dans de nombreux pays, par exemple, Planeta fácil[7] en Espagne et DR[8] au Danemark. La simplification de textes construit une version plus simple de textes complexes afin qu'ils soient plus facilement compréhensibles et intelligibles qui peut grandement faciliter la lisibilité et la compréhension pour certains groupes de personnes. Cette tâche a été développée dans plusieurs domaines de recherche et dans différentes spécialités. La section suivante décrit différents domaines d'application.

1.3. Domaines d'application de la simplification de textes

La SAT se développe considérablement dans le domaine du TAL. Dans tout domaine spécialisé, des notions et des termes techniques et spécifiques sont utilisés mais ils ne sont généralement pas compréhensibles par tout le monde, par exemple pour les publics illettrés. Cela concerne divers domaines d'application, notamment l'administration publique et le domaine médical et pharmacologique.

Dans l'administration publique, les textes qui proviennent d'une administration (département, région, préfecture, etc.) possèdent des caractéristiques et des difficultés propres en tant que textes administratifs et juridiques (François et al., 2018) : prédominance d'un jargon professionnel avec des termes techniques/spécialisés, présence de mots rares, emploi de phrases longues et/ou syntaxiquement complexes, etc.

Plusieurs initiatives et projets ont été développés dans le but d'utiliser un langage clair dans le domaine administratif. Parmi ces projets, le *Plain Language Action and Information Network* (PLAIN)[9], une organisation basée aux États-Unis qui utilise un langage clair dans les communications gouvernementales. Leurs directives sur la façon d'écrire du texte en anglais facile indiquent diverses méthodes pour simplifier le texte et conseillent également aux rédacteurs d'éviter ou de minimiser le jargon, les abréviations et les définitions. Une autre initiative est *Plain English Campaign*[10], un effort au Royaume-Uni pour simplifier le langage et prévenir les malentendus. Ils ont publié un ensemble similaire de lignes directrices sur l'écriture en langage claire, par exemple, privilégier la voix active, les phrases courtes, utiliser les instructions et les listes, etc.

Des textes mal écrits ou riche en jargon bureaucratique peuvent dresser des barrières entre lecteurs et communicants. La norme ISO[11] devrait contribuer à rendre les communications plus accessibles. Cette nouvelle norme (ISO 24495) définira des lignes directrices pour des communications claires et fournira aux rédacteurs une approche leur

[7] https://planetafacil.plenainclusion.org/
[8] https://www.dr.dk/ligetil/
[9] https://www.plainlanguage.gov/guidelines/
[10] http://www.plainenglish.co.uk/how-to-write-in-plain-english.html
[11] https://www.iso.org/fr/news/ref2566.html

permettant de communiquer efficacement avec différents publics dans plusieurs langues. Elle établira des principes, des techniques et des lignes directrices afin d'aider les rédacteurs à produire des communications claires.

Cependant, peu de travaux ont été faits sur ce domaine pour proposer des outils qui rendent les informations publiques plus accessibles aux personnes faiblement lettrées. François et al., (2020) présentent AMesure[12], une plateforme web dédiée à l'évaluation de la difficulté des textes administratifs, dont le but principal est d'aider les rédacteurs de ce type de textes à produire des documents plus accessibles. AMesure est le premier outil librement accessible pour le français qui offre à la fois une évaluation de la lisibilité d'un texte et un diagnostic plus précis des différents indicateurs des difficultés présents dans les textes. La Figure 2 montre une capture d'écran de la plateforme.

Figure 2 - Capture d'écran de la plateforme AMesure pour un texte administratif

Cette plateforme vise un double objectif : d'une part, elle propose une estimation générale de la difficulté d'un texte, c'est pourquoi les auteurs ont développé une formule de lisibilité spécialisée pour les textes administratifs, grâce à une technique d'annotation qui se base sur la mesure de la vitesse de lecture moyenne d'un groupe de lecteurs. D'autre part, elle identifie les éléments lexicaux et syntaxiques difficiles du texte, en intégrant divers indicateurs des difficultés présentes dans les textes administratifs relevant de plusieurs dimensions : complexité lexicale du texte, complexité de ses structures syntaxiques, taux de cohérence, etc.

[12] https://cental.uclouvain.be/amesure/

Dans le domaine médical et pharmacologique, les patients sont confrontés à des documents et des informations médicales techniques difficiles à comprendre, comme les informations sur la prise de médicaments et les documents cliniques. Chapman et al. (2003) et Lerner et al. (2000) montrent que les termes médicaux peuvent être un obstacle à la compréhension pour les patients. Ces difficultés peuvent avoir un impact négatif sur la communication entre les patients et les médecins, et les soins offerts aux malades (Tran et al., 2009). L'adaptation des documents techniques peut apporter des solutions à ce problème.

Ramadier et Lafourcade (2018) ont proposé une méthode pour simplifier des comptes rendus radiologiques en français grâce à une base de connaissance de sens commun qui contient à la fois de la connaissance générale mais aussi de spécialité. L'approche concerne la simplification lexicale. Ils utilisent pour cela non seulement des synonymes mais aussi des termes liés par des relations hiérarchiques et/ou sémantiques (comme l'hyperonymie). Cette dernière peut être très utile car un terme peut être expliqué comme une incidence spécifique de ses parents. Par exemple, le terme *carcinome hépatocellulaire* est un *cancer du foie* (relation *is-a*). La base de connaissance sur laquelle se base leur méthode de simplification est le réseau lexico-sémantique JeuxDeMots[13] (JDM) (Lafourcade, 2007). Koptient et Grabar (2020) ont développé un modèle de simplification automatique, à base de règles, des documents médicaux en français afin d'aider les utilisateurs à mieux comprendre ces textes, et ainsi de les aider à mieux comprendre leur pathologie ou leurs traitements. Ils utilisent un corpus existant construit avec des documents de différents genres (notices pharmaceutiques trouvées sur le site du gouvernement, extraits de revues systématiques et articles d'encyclopédie en ligne).

La simplification d'un texte peut être réalisée à différents niveaux linguistiques : lexical, morphologique, discursif et syntaxique. Ils sont détaillés dans la section suivante.

1.4. Les différents types de simplification

Brouwers et al (2012) ont établi une typologie articulée selon trois grands niveaux de transformations : lexical, sémantique et syntaxique. Gala et al., (2018) proposent une typologie à quatre niveaux linguistiques : *lexical, syntaxique, morphologique* et *discursif*. Dans ce qui suit, des exemples des niveaux de simplification sont présentés. Les exemples sont tirés du corpus Newsela. L'alignement est fait en utilisant MEDITE[14], un outil qui permet d'aligner deux textes afin d'exposer les invariants et les différences entre eux. Il détecte les blocs de caractères supprimés, insérés, remplacés et déplacés. La version originale (a) est suivie de la version simplifiée (b). Dans tous les exemples, la version simplifiée correspond au niveau 2 de simplification.

Dans ce qui suit nous précisons ces différents types de simplification, sachant que la simplification syntaxique sur laquelle porte notre travail de recherche, sera étudiée plus en détail dans le Chapitre 2.

[13] http://www.jeuxdemots.org/jdm-accueil.php
[14] http://obvil.lip6.fr/medite/

1.4.1. La simplification lexicale

La simplification lexicale a pour but soit de modifier le vocabulaire complexe du texte en remplaçant les unités lexicales par des paraphrases (explications) ou par des synonymes plus fréquents (généralement plus courts), soit en incluant des définitions appropriées (exemple 4). Dans les exemples de 4 à 9, les remplacements sont en bleu, les insertions en vert et en rouge les suppressions.

(4) a. Noah is part of a pilot study for CogCubed, a Minneapolis-based startup that develops games to help diagnose and ameliorate cognitive disorders such as ADHD.
b. Noah is part of a early study for CogCubed, a Minneapolis-based business that makes games to help decide if people have brain disorders such as Attention Deficit Hyperactivity Disorder (ADHD).

On distingue généralement deux types de simplifications lexicales (Siddharthan, 2014) :

- Expliquer les unités lexicales complexes au moyen d'une définition ou d'une paraphrase explicative dans le cas de mots techniques (exemple 5).

(5) a. They have tuberculosis.
b. They have tuberculosis, a disease that most often affects the lungs.

- Remplacer les unités lexicales complexes par un synonyme plus simple, en se basant sur des informations présentes dans certaines ressources. Par exemple, pour le français, il est possible de se référer à Lexique 3 (New et al., 2001), Manulex (Lété et al., 2004) ou ReSyf (Billami et al., 2018). L'objectif de cette étape est d'identifier le mot dans son contexte et de le remplacer par un substitut alternatif approprié. Pour la phrase 6 ci-dessous, le mot à substituer est *atrocities*. Les candidats synonymes sont *abomination*, *cruelty*, *enormity*, violation. Le choix de référence est *cruelty* (Cardon, 2018).

(6) a. Hitler committed terrible *atrocities* during the second World War.
b. Hitler committed terrible *cruelties* during the second World War.

La simplification lexicale peut être divisée en quatre étapes importantes (Shardlow, 2014). Ces mêmes étapes sont expliquées par Paetzold et Specia (2017) et Cardon (2018) :

1. *L'identification des mots complexes* consiste à décider, pour une phrase donnée, quels sont les mots qui risquent de ne pas être compris par la population cible.
2. *La génération de synonymes* ou autres substituts possibles consiste à trouver des mots ou des expressions pouvant remplacer le mot complexe cible.

3. *La sélection du substitut* consiste à décider quels substituts seraient les meilleurs candidats pour remplacer le mot complexe sans altérer la signification de la phrase et son contexte.
4. *La hiérarchisation des substituts* consiste à ordonner les candidats restants selon leur simplicité, le plus simple étant celui choisi pour remplacer le mot complexe.

1.4.2. La simplification morphologique

De façon générale, la simplification morphologique (morphèmes grammaticaux) concerne notamment le changement du temps verbal ou le remplacement d'une unité lexicale par une autre de plus fréquente dans sa même famille morphologique (exemple 7).

(7) a. Women who have desired longer hair have depended on hair straightening products, hot combs, flat irons, weaves, extensions and wigs.
b. Women who want longer hair depends on hair straightening products, hot combs and flat irons.

La simplification morphologique consiste à modifier les morphèmes grammaticaux, notamment changement des flexions verbales, comme le remplacement du passé simple par le présent (lorsque c'est possible, en tenant compte que les formes au présent sont plus fréquentes que les formes au passé).

1.4.3. La simplification discursive

La simplification discursive porte notamment sur le remplacement de pronoms personnels par le syntagme nominal désignant leur référent (explicitation de liens logiques, exemple 8).

(8) a. Manning really truly, genuinely believed that this information could make a difference and that "he believed that the information couldn't be used to harm the United States.
b. Manning believed that this information could make a difference. Manning also "believed that the information couldn't be used to harm the United States."

La simplification discursive consiste à reformuler les chaines de référence afin de mettre en évidence l'élément déjà évoqué : par exemple le remplacement de pronoms personnels par leur référent. Les pronoms peuvent être ambigus en induisant des inférences et notamment anaphoriques. Lorsque des référents multiples sont présents dans le texte, la tâche de résolution des pronoms est encore plus complexe (Martins et Le Bouédec, 1998). La reconnaissance de l'antécédent peut poser des difficultés pour un certain nombre de publics cibles, notamment les personnes atteintes de troubles de l'autisme. Pour ce public, il peut être difficile d'interpréter l'indice donné par le pronom qu'un référent "a récemment été discuté et est disponible en mémoire, mais n'est pas

actuellement dans l'attention" en raison de leur difficulté à diriger et à gérer l'attention (O'Connor et Klein, 2004).

Dans le cadre du projet français ALECTOR (cf. 1.5.1.3), Wilkens et Todirascu (2020) étudient les restrictions de cohésion liées aux chaînes de référence et leur applicabilité au domaine de la simplification automatique. Les auteurs proposent un corpus original d'évaluation, pour aborder le niveau de discours parmi différents niveaux de simplification. Pour chaque phrase, le corpus présente plusieurs simplifications alternatives (annotées), au niveau discursif, lexical et syntaxique. Wilkens et al., (2020) évaluent le module de simplification discursive en modifiant la structure des chaînes de référence, en réduisant les inférences nécessaires pour identifier liens anaphoriques et coréférentielles, car ces liens sont des éléments de difficulté pour les enfants faibles lecteurs. Les personnes atteintes d'autisme ou ayant une déficience intellectuelle ont des difficultés à trouver l'idée principale, à résoudre les anaphores et à déduire des informations (Martos et al., 2012 ; Feng, 2009).

1.4.4. La simplification syntaxique

La simplification syntaxique a pour but de simplifier la structure des phrases, par exemple, en désenchâssant les subordonnées dans le but de créer des propositions indépendantes ou en supprimant des incises jugées comme supplémentaires sur le plan informationnel. Par exemple, transformer une proposition relative en proposition indépendante ; transformer la voix passive en voix active, etc. Dans l'exemple (9), il s'agit de découpage de phrases.

(9) a. The company, which has its North American headquarters in Portland, Oregon, also said it will be a founding member of a coalition that addresses Native American mascots in sports.

b. Adidas also said it will be a founding member of a group that deals with the problem of Native American mascots in sports. The German company makes shoes and clothing. Its North American headquarters is in Portland.

Cette thèse porte sur la simplification syntaxique. Ce type de simplification sera développé en profondeur dans le chapitre suivant.

1.5. Ressources et corpus existants

Les approches de SAT s'appuient sur des grandes quantités de données parallèles avec des phrases complexes et leurs phrases simples associées. De telles ressources peuvent être construites manuellement, ou bien être acquises à partir de données issues de gros corpus parallèles (Nisioi et al., 2017). Cependant, le nombre de textes simplifiés manuellement augmente chaque jour en raison du nombre de sites Web qui fournissent des documents faciles à lire. Malheureusement, ces matériaux et leurs versions originales (non simplifiées) ne sont généralement pas mis à disposition ni à des fins de recherche ni

à des fins commerciales (Štajner, 2021). Nous aborderons les ressources disponibles gratuitement dans cette section.

1.5.1. Les corpus parallèles

Il existe quelques corpus parallèles alignés, obtenus à partir de corpus comparables. Ils sont principalement le résultat de simplifications manuelles. En anglais, deux corpus sont fréquemment utilisés : Wikipédia et sa version simplifiée (EW-SEW) et Newsela (Xu et al., 2015).

1.5.1.1. English Wikipedia et Simple English Wikipedia (EW-SEW)

SEW (*Simple English Wikipedia*) est une version simple de Wikipédia anglais (*English Wikipedia* EW) où la grande majorité des articles qui apparaissent se trouvent également dans EW, fournissant ainsi un alignement au niveau du document des textes complexes-simples (exemple 10). Des paires de phrases complexes-simples sont ensuite extraites des articles correspondants en alignant automatiquement les phrases ayant une signification similaire à l'aide d'heuristiques de similarité basées sur les termes.

(10) a. In 1940 Ellis became a member of MAUD who were investigating the possibility of using nuclear fission to develop new weapons.
b. During World War II he worked on the possibility of using nuclear fission to develop new weapons.

Zhu et al. (2010) introduisent PWKP (*Parallel Wikipedia*), un ensemble de données de 108,016 phrases complexes-simples parallèles extraites de EW-SEW. Pour permettre à l'opération de division de phrase d'être représentée dans leur ensemble de données, ils fusionnent des paires où les phrases complexes sont les mêmes et les phrases simples sont adjacentes, ce qui donne un alignement 1 à n. Woodsend et Lapata (2011) alignent également EW-SEW d'abord en alignant les paragraphes, puis au niveau de la phrase. Ils utilisent en outre l'historique de révision de SEW pour créer des paires de phrases complexes-simples. La version initiale de la phrase est utilisée comme source et la version éditée comme cible. Kauchak (2013) a ensuite mis à jour cet ensemble de données avec des données Wikipédia plus récentes et un traitement de texte amélioré pour créer 167,689 paires de phrases alignées.

Les alignements de phrases de tous ces trois travaux ont ensuite été combinés dans l'ensemble de données WikiLarge (Zhang et Lapata, 2017). L'ensemble de données résultant combine 296 402 phrases, utilisé dans plusieurs travaux ultérieurs (Dong et al., 2019 ; Vu et al., 2018 ; Mallinson et Lapata, 2019 ; Kriz et al., 2019).

Il a été démontré que l'utilisation d'alignements automatiques de EW-SEW produit des données d'apprentissage contenant des bruits avec des alignements où la version simplifiée est conservée et n'est pas plus simple (33%) ou n'est pas liée à la phrase complexe (17%) (Xu et al., 2015). En conséquence, l'ensemble de données Newsela a été proposé.

1.5.1.2. Newsela

La qualité du corpus d'entraînement est un facteur important et a été largement discutée, en particulier pour le corpus Wikipédia simple (Xu et al., 2016 ; Scarton et al., 2018). Le corpus Newsela[15] (Xu et al., 2015) est devenu une des principales ressources car il propose des simplifications manuelles à plusieurs niveaux, du niveau 0 comprenant le texte original au niveau 4 comprenant le texte le plus simplifié (exemple 11). Newsela est composé de 1,130 articles d'actualité qui ont été réécrits en 4 niveaux de simplicité différents par des rédacteurs professionnels. Les systèmes de SAT basés sur les approches neuronales exploitent ce corpus. Ces travaux seront détaillés dans le chapitre 2 (section 2.2.2.3).

(11) a. Original : The athletic shoe and apparel maker said Thursday it will provide free design resources to schools looking to shelve Native American mascots, nicknames, imagery or symbolism. The German company also pledged to provide financial support to ensure the cost of changing is not prohibitive.

b. **Simp_1** : It also pledged to provide financial help to ensure the cost of changing is not excessive.
Simp_2 : It also promised to help cover the cost to make sure that the change is not too expensive.
Simp_3 : The shoe and clothing company will also help to pay for the change. New uniforms, mascots and signs can be expensive for schools.
Simp_4 : Adidas will help schools design new uniforms. It will also help them to design new logos. Logos are the pictures on uniforms or signs. It costs a great deal of money to change logos and mascots. Adidas will help schools pay for it.

1.5.1.3. ALECTOR

Pour le français, le corpus ALECTOR[16] (Gala et al. 2020) est un corpus de textes parallèles, versions originales et simplifiées pour les enfants de sept à neuf ans. Ce corpus a été créé dans le cadre du projet du même nom visant à tester les effets de la simplification à différents niveaux linguistiques. ALECTOR est constitué de 183 textes différents, dont 79 textes originaux et leurs équivalents simplifiés. Il s'agit d'un ensemble de textes proposés dans les classes de CE1 à CM1, genre littéraire (contes, fables, histoires) et documentaire (sciences de la vie et de la terre). L'ensemble de ces textes a été simplifié manuellement et proposé comme tests de lecture lors d'une étude longitudinale de trois ans visant à évaluer les effets de la simplification de textes dans

[15] https://newsela.com/
[16] Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques : https://alectorsite.wordpress.com/

l'apprentissage de la lecture (exemple 12). Tous les corpus originaux ont subi des transformations à tous les niveaux (lexical, morpho-syntaxique, discursif).

(12) a. Un murmure de protestation s'élève dans la classe et une fille d'environ huit ans, aux longs cheveux tout bouclés, se dresse comme un ressort.
b. Un murmure de contestation s'élève dans la classe et une fille de 8 ans, aux cheveux bouclés, se lève vite.

Notons aussi que plusieurs corpus parallèles alignés ont été créés pour les travaux de simplification dans plusieurs langues :

- Espagnol : Simplext (Saggion et al., 2015)
- Danois : D-Sim (Klerke & Søgaard, 2012)
- Portugais du Brésil : PorSimples (Caseli et al., 2009)
- Italien : TERENCE et TEACHER[17] (Brunato et al., 2015), PaCCSS-IT[18] (Brunato et al., 2016).
- Russe (Dmitrieva et Tiedemann, 2021).

Certains de ces corpus comme PaCCSS-IT et ALECTOR proposent un schéma d'annotation pour la simplification avec plusieurs classes de modifications : découpage, fusion, réorganisation, insertion, suppression et transformation (substitution lexicale, changement de voix, etc.) (Brunato et al., 2014). Ce schéma d'annotation couvre la simplification lexicale et la simplification syntaxique.

1.5.2. Corpus comparable : CLEAR

CLEAR (Grabar et Cardon, 2018) est un corpus composé de textes comparables[19] en français qui se différencient par leur technicité : des textes techniques et les textes simplifiés correspondants. Le corpus est composé de documents issus de trois sources différentes : des notices de médicaments, des articles d'encyclopédies et des résumés de revues systématiques. Chacun de ces sous-corpus comporte deux versions du même texte : une version « technique » et une version « simple ». Pour les articles d'encyclopédies, les documents techniques en médecine proviennent de Wikipédia et les documents simples proviennent d'articles issus de Vikidia (destinée aux enfants de 8-13 ans). Pour les résumés de revues systématiques, les documents techniques sont les versions techniques des résumés et les documents simples aux versions simplifiée de ces résumés. Au total, ce corpus contient 16 190 paires de documents, avec plus de 15M d'occurrences de mots dans la version technique et 35M d'occurrences dans la version simplifiée.

[17] http://www.italianlp.it/resources/terence-and-teacher/
[18] http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/
[19] http://natalia.grabar.free.fr/resources.php

Cardon et Grabar (2020a ; 2021) et Grabar et al. (2021) ont comme objectif de préparer les ressources nécessaires pour la création de documents médicaux simplifiés pour le grand public. Ils proposent une méthode pour construire un corpus parallèle à partir de corpus comparables, en se basant sur le corpus CLEAR. La méthode proposée repose sur une étape de filtrage, qui ne garde que les meilleures phrases candidates à l'alignement, et une étape d'alignement considérée comme un problème de catégorisation. Il s'agit de décider si une paire de phrases est alignable ou non.

1.5.3. Ensembles de données pour l'évaluation humaine multi-références

Afin d'évaluer les simplifications générées automatiquement, des travaux antérieurs ont validé manuellement un sous-ensemble du corpus Wikipédia. Ils ont comparé la simplification générée avec des simplifications de référence de haute qualité à l'aide de métriques automatiques. Dans cette section, nous présentons les ensembles d'évaluation humaine qui sont traditionnellement utilisés dans la simplification des phrases.

1.5.3.1. TURKCORPUS

Xu et al., (2016) ont proposé Turkcorpus, un jeu de données composé de 2 359 phrases complexes (2 000 pour validation et 359 pour test) extraites de Wikipédia où, pour chaque phrase complexe, 8 simplifications de référence ont été collectées à l'aide d'*Amazon Mechanical Turk*. La plupart des phrases simplifiées sont cependant très similaires à la phrase complexe avec seulement quelques simplifications lexicales ou suppressions de mots, c'est-à-dire qu'elles ne sont pas adaptées à l'évaluation de systèmes de simplification de phrases à part entière effectuant des divisions de phrases et des opérations de réécriture plus complexes.

1.5.3.2. HSPLIT

Axé uniquement sur la division des phrases, l'ensemble d'évaluation HSplit (Sulem et al., 2018b) a été créé en utilisant les mêmes 2 359 phrases complexes que Turkcorpus et fournit quatre références humaines par phrase source. Chaque référence a été créée en opérant uniquement la division de la phrase sur la phrase complexe d'origine. Il s'agit donc d'un ensemble de données pour l'évaluation du découpage des phrases, mais il ne se généralise pas à la simplification des phrases en général.

1.6. Intérêt de la simplification pour d'autres tâches de TAL

La simplification des phrases présente de nombreuses similitudes, mais aussi des différences marquées avec d'autres tâches de réécriture de texte. Dans cette section, nous donnons un aperçu des tâches de TAL similaires telles que la traduction automatique, le résumé, la compression de phrases et la création de paraphrases, et nous décrivons en

quoi elles diffèrent de la simplification des phrases, et en quoi elles peuvent être complémentaires.

1.6.1. Résumé de textes

La simplification des phrases partage des similitudes avec le résumé de textes. L'un des moyens les plus explorés pour réduire automatiquement la difficulté de lecture des phrases consiste à les raccourcir en supprimant des mots et des phrases de manière à ne conserver que les informations essentielles. Le résumé et la simplification impliquent tous deux de transformer un texte. De plus, pour les deux tâches, les textes source et cible sont dans la même langue. Le résumé et la simplification de textes sont deux techniques utilisées pour adapter les textes aux personnes possédant des compétences faibles en lecture, y compris les enfants, les locuteurs non natifs et les illettrés. Cependant, le raccourcissement de phrase vise à aider les lecteurs et améliorer leur temps de lecture en filtrant les parties les moins informatives d'un texte. Cet objectif est différent de celui de la simplification syntaxique, qui vise à aider les personnes à atteindre une meilleure compréhension du texte.

Siddharthan (2006) affirme qu'il convient de mettre en contraste la simplification du texte et la tâche du résumé automatique de texte : l'objectif de la simplification du texte est de préserver le contenu des informations tout en réduisant la complexité linguistique. Cependant, l'objectif du résumé est de réduire le contenu des informations en ne conservant que les informations les plus importantes, ce qui peut être utile dans la simplification dans les cas où le texte à simplifier contient trop de détails et d'informations supplémentaires. La simplification consiste à simplifier les textes au niveau de la phrase, alors que le résumé fonctionne au niveau d'un texte. Bien que les textes simplifiés soient généralement plus courts, ce n'est pas nécessairement le cas et la simplification peut entraîner une sortie plus longue, en particulier lors de la génération d'explications.

Les méthodes de simplification de phrase ressemblent davantage aux méthodes de résumé abstractif, car elles doivent réécrire au moins une partie du texte original, car seule l'extraction du contenu d'entrée ne peut pas gérer toutes les opérations de simplification. En outre, le résumé abstrait n'est souvent pas principalement compris comme une stratégie de simplification, mais plutôt considéré comme une « étape importante vers la compréhension du langage naturel » (Chopra et al., 2016).

Les utilisateurs cibles de systèmes de résumé doivent souvent compenser un déficit de cohérence du texte dans les textes de sortie par leur capacité cognitive à reconstruire des connexions logiques à partir de leur connaissance réelle des mots. Cela ne peut pas être attendu des utilisateurs de la simplification de texte, car généralement ils ne peuvent pas traiter les connexions logiques facilement (Saggion et al., 2015). En outre, les textes peuvent être simplifiés à des niveaux linguistiques très différents : la simplification peut avoir comme objectif de réduire la longueur de la phrase, la complexité lexicale, la variété lexicale, le niveau de détail de l'information transmise, etc.

1.6.2. Création de paraphrases

L'objectif de la paraphrase est la reformulation de la signification d'une phrase en utilisant d'autres mots. Il s'agit d'une tâche de réécriture de texte monolingue qui vise à exprimer le sens original mais avec une autre formulation. Une phrase simplifiée peut être considérée comme un type spécifique de paraphrase où la phrase doit être plus facile à lire et à comprendre. Comme pour la simplification des phrases, les paraphrases parallèles sont difficiles à trouver en grande quantité. Cette méthode a été utilisée pour créer de grandes bases de données de paraphrases au niveau des mots (Pavlick et al., 2015), des simplifications lexicales (Pavlick et Callison-Burch, 2016, Kriz et al., 2018) ou des corpus de paraphrases au niveau de la phrase (Wieting et Gimpel, 2018).

Bouamor (2012) soutient que les locuteurs d'une langue utilisent les paraphrases afin, par exemple, de simplifier et de communiquer un énoncé le plus clairement possible. La paraphrase est une opération de reformulation d'un énoncé qui conserve le sens. Il s'agit de produire un texte cible à partir d'un texte source afin d'expliciter, clarifier ou développer certains aspects. Selon Bouamor, la définition de la paraphrase ressemble à celle de la simplification, les deux sont un moyen alternatif qui exprime le même contenu sémantique, la même information ou la même idée que la forme originale (Barzilay et McKeown, 2001 ; Fujita, 2005 ; Callison-Burch, 2007).

Dras (1999) a développé une étude approfondie des paraphrases syntaxiques. Il établit une catégorisation sur cinq axes :

1. Changement de perspective : la façon dont les éléments sont représentés, tel que le remplacement d'un verbe par un adjectif dans une phrase.
2. Changement d'emphase : changement de la structure syntaxique en modifiant son focus comme le passage de la voix active à la voix passive.
3. Changement de relation : changement de connexion entre les propositions des phrases (lors du découpage ou de la fusion des phrases).
4. Suppression d'éléments secondaires de la phrase.
5. Déplacement de la position de quelques propositions dans la phrase.

Bouamor (2012) définit trois classes de paraphrases :

1. Paraphrase *lexicale* : des unités lexicales individuelles ayant la même signification sont généralement appelés paraphrases lexicales (par exemple, « manger » ↔ « consommer » et « bouquin » ↔ « livre »).
2. Paraphrase *sous-phrastique* : des syntagmes ou fragments d'une phrase sont en relation d'équivalence sémantique dans un contexte donné (par exemple, « envisage-t-elle » ↔ « a-t-elle l'intention »).
3. Paraphrase *phrastique* : Deux énoncés véhiculant le même contenu sémantique (par exemple « Elle a grondé son enfant » ↔ « Elle s'est fâchée contre son enfant »).

L'utilisation de méthodes automatiques pour générer des paraphrases a été appliquée avec succès pour la simplification de texte parmi d'autres tâches de TAL. La paraphrase

est utilisée pour supprimer les structures syntaxiques difficiles pour les apprenants sourds de l'anglais et du japonais écrits (Inui et al., 2003). Des méthodes de paraphrase ont été appliquées pour simplifier les textes de journaux pour les personnes atteintes d'aphasie (Carroll et al.,1998, 1999) et atteintes de Syndrome de Down (Saggion et al., 2011) ainsi que pour simplifier l'information en ligne pour les personnes atteintes d'aphasie (Devlin et Unthank, 2006).

1.6.3. Traduction automatique

L'objectif de la traduction automatique est de traduire un texte vers une autre langue. La plupart des méthodes utilisées en Traduction Automatique (TA) sont également adaptés à la simplification de phrase comme la TA Statistique où la simplification est considérée comme traduction automatique monolingue (les langues source et cible sont identiques). Cette notion sera détaillée dans la section 2.2.2.

1.7. Conclusion

Nous avons présenté dans ce chapitre la notion de complexité linguistique et les travaux de recherche relatifs à des adaptations ou simplifications pour différents publics cibles et dans différents domaines. Plusieurs facteurs ont une incidence sur la compréhension de textes et la densité propositionnelle d'une phrase est proportionnelle à sa longueur, d'où les phrases qui contiennent des enchâssements de propositions (des subordinations, des relatives et des coordinations) sont généralement les plus longues.

La disponibilité limitée des ressources pour la simplification de textes demeure un problème pour le domaine. Les contributions individuelles telles que le développement et la publication d'un corpus de simplification plus ciblé comme Newsela et ALECTOR ont eu un impact notable et ont en partie résolu certains des problèmes surtout ceux qui concernent la qualité de simplification. Nous avons également montré la relation et les intérêts de la SAT dans d'autres tâches de TAL, notamment le résumé, la création des paraphrases et la traduction automatique, ainsi que les points communs et les différences.

Notre travail de thèse a comme objectif le développement d'un système de simplification spécifique à la syntaxe, dont le but est de transformer les textes qui contiennent ces types de constructions en autres qui ne les contiennent pas. C'est pourquoi nous avons dédié le chapitre suivant à ce domaine. Ainsi, le chapitre suivant dresse un état de l'art des différentes opérations de transformations, les approches utilisées, ainsi que la relation entre la simplification de textes et les autres domaines de TAL, tels que la traduction et le résumé automatique.

2. Simplification syntaxique (automatique) de textes

De façon générale, la simplification syntaxique a pour but de transformer la structure des phrases. Il s'agit d'effectuer des opérations de substitution, de suppression, de division et de regroupement qui changent, par exemple, les phrases dépendantes en phrases indépendantes et en transformant les phrases à la voix passive en autres en voix actives. Dans ce chapitre, nous décrivons ce type de simplification, tout en dressant un état de l'art sur les opérations, ainsi que les approches adoptées dans la simplification syntaxique automatique de textes. Dans notre travail, nous avons traité l'opération de découpage et le passage de la voix passive en voix active. Nous n'avons pas traité les opérations de fusion, de réorganisation, d'insertion et de suppression.

La section 2.1 s'intéresse aux opérations de transformation syntaxique, telles que le découpage, la réorganisation et la suppression de phrases (ou de fragments). La section 2.2 décrit les différentes approches de simplification syntaxique. Les avantages et inconvénients de chacune de ces approches sont présentés à la section 2.3.

2.1. Opérations liées à la simplification syntaxique

Comme nous l'avons déjà évoqué, la **simplification syntaxique** a pour but de simplifier la structure des phrases, par exemple, en désenchâssant les subordonnées afin de créer des propositions indépendantes. Le Tableau 1 décrit les types de transformations de phrases les plus fréquentes (Štajner, 2016). Ces transformations correspondent aux appositives, relatives, participiales, coordinations, adverbiales, subordinations et les voix passives. Dans notre travail, nous nous sommes inspirés de cette typologie pour déterminer les constructions à traiter.

Type	Original	Simplified
Appositions	"John Smith, a New York taxi driver, won the lottery."	"John Smith is a New York taxi driver. John Smith won the lottery."
Relative clauses	"The mayor, who recently got a divorce, is getting married again."	"The mayor recently got a divorce. The mayor is getting married again."
Participial phrases	"The participants (...) will be presented with a book, edited by the town council (...)"	"The participants (...) will be presented with a book. This book is edited by the town council (...)"
Coordinate clauses	"The problem is difficult and there is probably no right answer."	"The problem is difficult. There is probably no right answer."
	"The problem is difficult and has no easy solution."	"The problem is difficult. The problem has no easy solution."
Adverbial clauses	"Needing money to pay my rent, I forced myself to beg my parents."	"I needed money to pay my rent. I forced myself to beg my parents."
Subordinate clauses	"Though all these politicians avow their respect for genuine cases, it's the tritest lip service."	"All these politicians avow their respect for genuine cases. However, it's the tritest lip service."
Passives	"Mary was punched by John."	"John punched Mary."

Tableau 1 - Types de transformations de phrases les plus fréquentes (Štajner, 2016)

Avant de développer un système de SAT, toutes les transformations de simplifications se définissent suite à une étude de corpus afin d'étudier les structures de phrases qui peuvent poser un problème de lecture ou de compréhension. Il existe des outils de simplifications de textes qui reposent sur des typologies de transformations réalisées sur des corpus. Dans le cadre du projet français ALECTOR[20] (cf. 1.5.1.3), Gala et al., (2020b) proposent une typologie de variations syntaxiques, à la suite de l'analyse de deux corpus parallèles pour les enfants faibles lecteurs et dyslexiques (Figure 3). L'objectif de ce projet était d'identifier les phénomènes qui ont un impact dans l'amélioration de la lecture et de la compréhension des textes par les jeunes enfants. Chacun des trois objectifs de la simplification syntaxique, présentés dans des rectangles aux angles droits colorés, est lié aux opérations effectuées manuellement dans les corpus analysés : privilégier l'ordre SVO ; supprimer les informations « secondaires » et privilégier les phrases courtes. Dans les rectangles aux angles arrondis, il est possible d'identifier les moyens d'atteindre les objectifs de simplification visés.

[20] Aide à la LECTure pour améliORer l'accès aux documents pour enfants dyslexiques : https://alectorsite.wordpress.com/

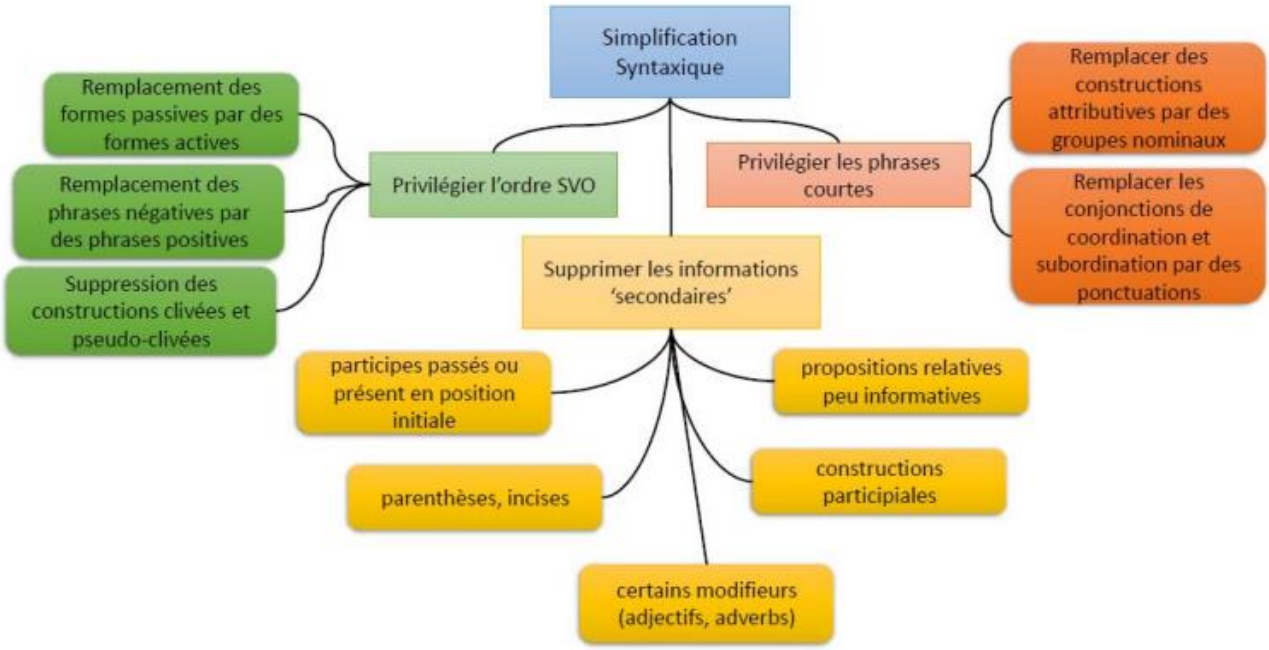


Figure 3 - Typologie des transformations syntaxiques en vue d'une simplification de textes (Gala et al., 2020b)

Brunato et al., (2014) proposent des macro-catégories d'opération de simplification automatique : découpage des conjonctions et des relatives, fusion de deux (ou plusieurs) phrases dans une seule phrase, changement d'ordre, suppressions d'informations, insertion d'éléments, des transformations lexicales, syntaxiques et morpho-syntaxiques. Koptient et al., (2019) proposent une typologie de simplifications adaptée aux textes de spécialité (techniques, et plus particulièrement les textes médicaux). Parmi les transformations appliquées, ils identifient la substitution lexicale, l'ajout d'explications, la substitution syntaxique (transformer la phrase passive vers la phrase active), la suppression d'informations et la pronominalisation.

Xu et al., (2015) ont effectué une analyse systématique de l'ensemble du corpus Newsela (2016), un corpus parallèle de textes en langue anglaise avec quatre niveaux de simplification, tout en se focalisant sur l'aspect lexical. Ce corpus propose des simplifications manuelles à plusieurs niveaux, du niveau 0 comprenant le texte original au niveau 4 comprenant le texte le plus simplifié. Nous avons proposé une typologie linguistique des transformations syntaxiques à partir de ce même corpus (Hijazi, 2020). Nous avons étudié les différentes transformations caractérisant le passage d'un niveau de simplification à l'autre sur un sous-ensemble de textes (107 phrases tirées de 6 textes choisis d'une façon aléatoire du corpus). Le but est d'analyser ce corpus en se focalisant sur les aspects syntaxiques, c'est-à-dire, étudier les changements syntaxiques qui ont été faits lors du passage d'un niveau de difficulté à un autre plus simple. Une analyse qualitative a été réalisée afin de cibler différents types d'opérations de simplification. Sur la base de différentes classifications des opérations de simplification proposée dans la littérature (Brunato et al., 2015 ; Bott et Saggion, 2014 ; Coster et Kauchak, 2011 ; Medero et Ostendorf, 2011 ; Gala et al. 2020b), nous avons identifié six classes principales

d'opérations observées dans le corpus Newsela, à savoir : les *découpages*, les *fusions*, les *réorganisations*, les *insertions* (rajouts), les *suppressions* et les *substitutions*.

Dans ce qui suit, nous présentons en détail ces opérations de transformation de phrases relatives à la simplification syntaxique. Des exemples issus du corpus Newsela seront présentés pour chaque opération avec (a) est la phrase originale et (b) est la simplifiée. *Simp_x* marque le niveau de simplification dans Newsela.

2.1.1. Opérations de découpage de phrases

C'est l'opération la plus fréquente en SAT, pour les applications à la fois humaines et automatiques. En règle générale, le découpage supprime les conjonctions, les deux points, les points-virgules, les énumérations et les appositives afin d'obtenir deux phrases indépendantes (exemple 13). Nous nous sommes concentrés précisément sur cette opération dont le but est de passer d'une phrase complexe en deux ou plusieurs phrases plus simples.

(13) a. The advocacy group says about a dozen schools have dropped Native American mascots over the past two years and an additional 20 are considering a change.

b. The advocacy group says about a dozen schools have dropped Native American mascots over the past two years**. An** additional 20 are considering a change. (*Simpl_1*)

2.1.2. Opération de fusion de phrases

Cette opération est conçue comme l'inverse de la division, c'est l'opération par laquelle deux (ou plusieurs) phrases originales sont fusionnées en une phrase simplifiée unique. Une telle transformation est moins fréquente que la division des phrases (exemple 14).

(14) a. Eric Liedtke, Adidas head of global brands, *traveled to the conference*. He said sports must be inclusive.

b. Eric Liedtke, Adidas head of global brands, said sports must be inclusive (*Simp_1*).

2.1.3. Opération de réorganisation

Lorsque certaines structures complexes ne sont pas supprimées, elles sont souvent déplacées ou modifiées dans le texte dans le but de maintenir une structure SVO (Gala et al., 2020b). Des psycholinguistes avaient montré que la disposition syntaxique SVO aide au déchiffrage pour les jeunes apprenants de l'anglais langue seconde (Skehan, 1991). D'autres chercheurs ont vérifié l'importance de l'ordre SVO et le rôle qu'il joue sur le plan linguistique et cognitif (Hosenfeld, 1984 ; Blanche-Benveniste, 2013a ; Chuquet 2000). L'expérience de Blanche-Benveniste montre que toutes les autres dispositions (SV, VSO et VOS) étaient considérées comme difficiles, même lorsque les apprenants avaient dans

leur langues maternelles ces dispositions. Cette opération marque le changement de position de mots entre la phrase d'origine et son équivalent simplifié (exemple 15).

(15) a. *According to the group Change the Mascot*, there are about 2,000 schools nationwide that have Native American mascots.

b. About 2,000 schools nationwide have Native American mascots, *according to the group Change the Mascot* (*Simpl_2*).

2.1.4. Opération d'insertion

Le processus de simplification peut entraîner une phrase plus longue, en raison de rajout de mots ou d'expressions qui fournissent des informations de clarification. Nous avons observé un seul type d'informations pour marquer les insertions : le rajout d'explications et de définitions (exemple 16).

(16) a. The NFL's Washington Redskins have resisted appeals by Native American and civil rights groups to change their name and mascot.

b. Americans have asked the Washington Redskins to change the team's name. [*The Redskins is a football team. Redskins is a very old word for Native Americans. It is not a nice word*]. The football team has refused again and again. (*Simpl_4*)

2.1.5. Opération de suppression

Les informations secondaires ou redondantes, généralement considérées comme supplémentaires au niveau syntaxique, ne sont pas incluses dans les textes simplifiés. Un texte devrait être simplifié en éliminant les informations redondantes. Les phrases simplifiées contiennent moins d'adverbes ou d'adjectifs que les phrases originales, par exemple. Nous proposons six types d'informations qui peuvent être supprimées (Štajner, et al., 2013 ; Drndarević et al., 2013) : les informations entre parenthèses, les exemples, les constructions appositives, certains modifieurs, quelques relatives ainsi que les expressions temporelles et locatives dans certains cas (exemple 17).

(17) a. Some colleges kept their nicknames by obtaining permission from tribes, *including the Florida State Seminoles and the University of Utah Utes*.

b. Some colleges were able to keep their names, though. They received permission from tribes (*Simpl_3*)

2.1.6. Opération de substitution morpho-syntaxique

Nous avons observé des substitutions de nature morpho-syntaxique : le changement des formes passives en formes actives, les propositions négatives en positives, les formes impersonnelles en personnelles, le discours indirect en discours direct. Dans le cadre de la thèse, nous avons uniquement traité le passage de la voix passive en voix active (exemple 18).

(18) a. A few hundred protesters gathered in scattered demonstrations in Sao Paulo, Rio de Janeiro, Porto Alegre, Brasilia and Belo Horizonte, *but they were controlled by police.*

b. A few hundred protesters gathered in scattered demonstrations in Sao Paulo, Rio de Janeiro, Porto Alegre, Brasilia and Belo Horizonte. *Police kept them under control (Simp_3).*

En plus de ces variations linguistiques, les versions simplifiées de Newsela présentent des variations typographiques notamment avec des variations des nombres (exemple 19). Ce procédé est assez courant pour alléger la charge cognitive pendant la lecture (particulièrement pour les lecteurs en difficulté, cf. Rello, 2014 ; Gala et al. 2020a).

(19) a. The advocacy group says about *a dozen* schools have dropped Native American mascots over the past two years and an additional 20 are considering a change.

b. In the last two years, about *12* schools stopped using them. Another 20 are thinking about it. (*Simpl_4*).

Dans le cadre de cette thèse, nous nous intéressons essentiellement aux deux opérations suivantes : le découpage des phrases et le passage de la voix passive en voix active.

2.2. Différentes approches de la simplification syntaxique automatique

Il existe trois grandes catégories d'approches de simplification de phrases : (1) les approches à base de règles, (2) les approches basées sur la traduction automatique statistique (*Statistical Machine Translation* SMT) y compris la traduction automatique basée sur les arbres (*Tree-Based Machine Translation* TBMT) (Zhu et al., 2010 ; Woodsend et Lapata, 2011), la traduction automatique basée sur des syntagmes (*Phrase-based Machine Translation* PBMT) (Coster et Kauchak 2011 ; Wubben et al., 2012 ; Stajner, et al., 2015) et la traduction automatique basée sur les arbres syntaxiques (*Syntax-based Machine Translation* SBMT) (Xu et al. 2016) et (3) les approches basées sur l'apprentissage automatique neuronal (*Neural Machine Translation* NMT). La combinaison des deux approches est appelée *approche hybride* (couplage d'un système de simplification lexicale neuronal avec un système de simplification syntaxique à base de règles, par exemple).

Les premiers travaux sur la simplification automatique sont des approches à base de règles ciblant des constructions syntaxiques. Un peu plus tard, la simplification a été considérée comme une traduction monolingue, elle a été opérée en utilisant toutes les techniques de traduction automatique et plus récemment diverses architectures neuronales.

Dans ce qui suit nous présentons plusieurs approches en nous intéressant plus particulièrement à la simplification syntaxique. Pour chacune de ces approches nous évoquerons divers travaux de SAT existants.

2.2.1. Approches à base des règles

Les approches à base de règles sont les premières approches adoptées en SAT, elles ont été proposées pour des cas d'applications spécifiques et pour une population bien ciblée. Il s'agit par exemple des systèmes qui ne traitent que les coordinations et les relatives. La simplification syntaxique implique généralement deux étapes de travail : étudier des sources de difficulté syntaxiques dans un corpus afin d'identifier constructions syntaxiques qui peuvent être simplifiées, puis création d'un ensemble de règles de simplification (soit élaborées à la main, soit générées automatiquement).

Les travaux de Chandrasekar et al., (1996), sont considérés comme à l'origine de la discipline. Ils visaient à diviser les phrases longues d'un texte, comme étape de pré-traitement pour l'analyse syntaxique automatique en vue d'améliorer les performances de l'analyseur. Cette approche était basée sur des règles capables de rendre indépendantes les coordinations, les propositions relatives ou appositives, fournissant le fondement des approches actuelles de simplification basée sur des règles.

Chandrasekar et Srinivas (1997) ont proposé un système à base de règles d'apprentissage automatique fonctionnant en deux étapes : la description structurelle de la phrase et la simplification de celle-ci en utilisant la description réalisée à la première étape. Les règles, quant à elles, sont créées à partir d'un algorithme permettant d'induire automatiquement des règles à partir d'un corpus de phrases alignées et de leurs versions simplifiées manuellement.

Peu après, est développée la « simplification pour les humains », au travers du projet PSET (*Practical Simplification of English Text*) (Carroll et al., 1998). Cette étude a adopté une approche similaire à Chandrasekar et al., (1996) : transformation des phrases en arbres syntaxiques (en utilisant un analyseur statistique) sur lesquels sont ensuite appliquées des règles de simplification. Ce projet ne simplifie que les coordinations et les passives, mais fait appel à un système de résolution d'anaphores afin de remplacer les pronoms par leur antécédent dans les phrases simplifiées. Les mots rares sont ensuite remplacés par des synonymes plus fréquents.

Les systèmes basés sur ces approches soutiennent que les règles de simplification syntaxique sont difficiles à apprendre à partir de corpus, car une morphologie complexe (riche) et des variations de temps doivent être apprises à partir d'instances spécifiques vues dans le corpus.

Selon Max (2005), l'utilisation de règles permet d'exprimer des conditions sur leur applicabilité. Elles offrent une meilleure maîtrise du contexte de simplification qui peut être plus ou moins spécifié. Le développement des règles peut être incrémental, ce qui autorise une évaluation progressive du système sur un corpus de test, afin de contrôler que l'ajout de règles ne dégrade pas les performances. Par ailleurs, si une règle n'est pas appliquée sur une phrase, cette dernière reste inchangée, ce qui garantit, au cas de la non-

application des règles, la conservation de la complexité syntaxique au lieu de changer la rendre agrammaticale ou asémantique.

Siddharthan (2006) soutient que l'application par hasard de certaines règles de simplification peut nuire à la cohésion du texte, comme dans l'exemple suivant où la phrase (20), qui contient une conjonction et une subordonnée relative, est transformée en une séquence de trois phrases plus simples (20b), (20c) et (20d). Siddharthan explique que cette transformation affecte la cohésion du texte puisque la subordonnée concessive (20d) est liée à la phrase (20c) au lieu de (20b) comme elle devrait être. Le pronom « *it* » pourrait également être mal interprété comme désignant « *employment agency* ».

(20) a. Mr. Anthony, who runs an employment agency, decries program trading, but he isn't sure it should be strictly regulated.
b. Mr. Anthony decries program trading.
c. Mr. Anthony runs an employment agency.
d. But he isn't sure it should be strictly regulated.

Siddharthan propose un système basé sur des règles. Il décompose la tâche de simplification en trois étapes : *analyse*, *transformation* et *régénération* (Figure 4) :

1. La première étape (*analyse*) utilise divers analyseurs de phrases (taggeurs, analyseurs morphologiques, syntaxiques, etc.) pour fournir une description structurelle de l'entrée.
2. La simplification s'effectue dans la deuxième étape (*transformation*), sur la base de la description de la phrase obtenue dans la première étape et selon un ensemble de règles de réécriture, qui effectuent les opérations de simplification, telles que le découpage, la réorganisation et la suppression de phrases ou de propositions. Bien que des techniques automatisées pour appliquer ces règles existent, la plupart des systèmes de simplification syntaxique utilisent des règles de réécriture écrites à la main, car cela élimine le besoin de corpus annotés et conduit généralement à une meilleure précision (Shardlow, 2014).
3. Après transformation, une phase de *régénération* peut également être effectuée, au cours de laquelle de nouvelles modifications sont apportées au texte pour en améliorer la cohésion et la lisibilité.

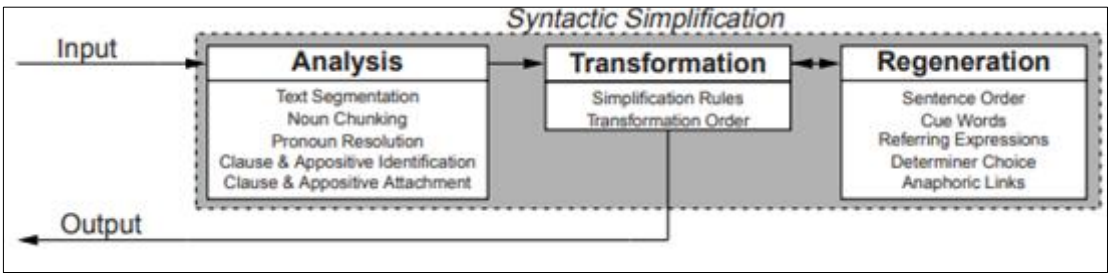


Figure 4 - Architecture du système de simplification syntaxique de texte (Siddharthan, 2006)

Le travail de Seretan (2012), s'est concentré sur la réduction de la complexité syntaxique en français. Elle a choisi des articles de journaux comme corpus d'étude pour le développement du système de simplification. Quarante règles ont été définies manuellement. Les transformations syntaxiques modifient la structure d'une phrase complexe comme les subordonnées, les relatives, les appositives, ainsi que la suppression de certaines expressions (p. ex. *par ailleurs*, *dans ce contexte*, etc.).

Brouwers et al., (2014) ont suivi une approche similaire pour la simplification syntaxique du français. Les auteurs utilisent des corpus parallèles réalisés pour définir manuellement 19 règles de suppression, de modification et de division de phrases. La suppression procède à l'élimination des éléments syntaxiques secondaires comme les propositions subordonnées. La méthode appliquée passe par un processus de surgénération qui permet d'obtenir, dans un premier temps, les substituts possibles pour une seule phrase. La meilleure candidate est ensuite sélectionnée sur la base de critères de lisibilité intégrés au sein d'un algorithme d'optimisation.

Saggion et al., (2015) développent un outil de simplification capable d'adapter le contenu textuel pour des hispanophones ayant une déficience cognitive. Le système Simplext comprend trois modules : un composant de simplification syntaxique ; un composant de simplification basé sur des synonymes utilisant des ressources externes (thésaurus) ; et un composant de simplification lexicale basé sur des règles implémentées dans GATE permettant de supprimer les informations secondaires (uniquement des informations entre parenthèses) et de réécrire les chiffres.

Cette approche reste d'actualité notamment pour les langues peu dotées pour lesquelles il n'existe pas de corpus parallèles. Elle est adaptée à la simplification syntaxique lorsqu'un système se concentre sur des structures et des phénomènes linguistiques très spécifiques qui sont relativement faciles à gérer avec un ensemble fini et restreint de règles (Siddharthan, 2014). Cependant, ces systèmes basés sur des règles souffrent souvent d'une couverture très limitée et ne parviennent pas à détecter les variations de la forme de la surface, elles consomment beaucoup de temps et elles requièrent beaucoup d'implication humaine pour la définition des règles. Plus récemment, des méthodes basées sur les données inspirées de la traduction automatique ont utilisé des modèles statistiques ou neuronaux pour la SAT. Ces approches sont développées dans la section suivante.

2.2.2. Approches relevant de la traduction automatique

La plupart des approches de simplification automatique de texte considèrent le processus de simplification comme un cas particulier de Traduction Automatique (TA) (MT pour *Machine Translation*) avec des langues source et cible identiques (traduction monolingue). La majorité des recherches en simplification de phrases se sont concentrées sur la langue anglaise, notamment en raison de la disponibilité de corpus d'apprentissage et d'évaluation dans cette langue tels que Wikipédia et sa version simplifiée (EW-SEW), ainsi que Newsela (Xu et al., 2015).

Une première contribution dans ce sens est le travail de Zhu et al., (2010), qui utilise un modèle de traduction basé sur des transformations d'arbres d'analyse syntaxique, exploitant la nécessité d'effectuer des opérations de traduction au niveau lexical dans le cas d'une traduction monolingue. Les développements ultérieurs dans le domaine de la simplification en tant que traduction se sont ensuite concentrés sur le paradigme de la traduction automatique statistique et sont ensuite passés aux méthodes neuronales, reflétant les tendances générales de la traduction automatique (depuis 2015). Nous abordons très brièvement ces différentes méthodes et soulignons un certain nombre de leurs applications à la simplification, d'abord pour la traduction automatique statistique puis pour la traduction automatique neuronale.

Ces approches reposent principalement sur l'utilisation de grands corpus disponibles à la place de connaissances expertes. L'objectif est d'apprendre automatiquement les règles depuis ces ressources de données. Pour qu'un algorithme d'apprentissage automatique puisse apprendre ces transformations, il faut lui fournir suffisamment d'exemples de simplifications de bonne qualité appariant des couples de phrases simples et complexes. La qualité de sorties d'un système de simplification dépendra fortement de la qualité et de la quantité des phrases parallèles.

Les approches de traduction automatique dans la simplification syntaxique regroupent les méthodes suivantes : la traduction automatique statistique fondée sur les syntagmes (*Phrase-Based Statistical Machine Translation*), la traduction automatique statistique fondée sur les arbres syntaxiques (*Syntax-Based Statistical Machine Translation*) et la traduction automatique neuronale (*Neural Machine Translation*).

2.2.2.1. Simplification de phrases basée sur des syntagmes (PBMT)

En TA statistique, en particulier les méthodes basées sur des syntagmes (*Phrase-Based Machine Translation* PBMT ; Koehn et al., 2007), un système apprend d'abord les alignements entre les mots et les syntagmes à partir d'un corpus de phrases parallèles. À partir des mêmes données, les probabilités de traduction entre ces phrases sont ensuite estimées. Au moment du décodage, un système génère des traductions partielles à partir d'une entrée et les évalue en fonction de son modèle de traduction ainsi que d'un modèle de langue censé assurer la fluidité et la grammaticalité de la sortie. Cette approche consiste à segmenter une phrase en syntagmes, ensuite à faire la traduction syntagme par syntagme puis les réordonner pour formuler une phrase de sortie a priori simplifiée. Certains travaux avaient pour objectif le découpage des phrases longues en phrases plus courtes et plus simples (Specia, 2010), d'autres se sont intéressés à la suppression des expressions secondaires (Coster et Kauchak, 2011b).

Specia (2010) a été la première à aborder problème de la simplification de textes en le considérant comme une tâche de traduction dans le cadre de traduction automatique statistique (*Statistical Machine Translation* SMT) pour apprendre à traduire des phrases complexes en phrases simplifiées. Elle s'appuie sur un corpus parallèle de textes originaux et simplifiés, alignés au niveau des phrases.

Coster et Kauchak (2011b) utilisent PBMT pour entraîner un système de simplification de phrases sur 137k paires de phrases extraites du corpus Wikipédia en anglais (EW-

SEW). Ils améliorent le modèle avec un composant de suppression de phrases pour améliorer le modèle sur ce type particulier d'opérations de simplification. Les systèmes de simplification de phrases opèrent souvent trop peu de modifications sur la phrase originale s'ils sont entraînés sans biais inductifs spécifiques.

Wuben et al., (2012) modifient également l'approche pour la simplification des phrases en utilisant les données EW-SEW, mais au lieu d'ajouter un composant de suppression, ils reclassent l'hypothèse en fonction d'une métrique de distance d'édition de Levenshtein pour forcer le modèle à apporter suffisamment de modifications. Cependant, le modèle effectue encore relativement peu de modifications sur la phrase originale et il ne gère pas la division de phrases. Il effectue des opérations de substitution lexicale, de suppression de phrases et de réorganisation de phrases.

2.2.2.2. Simplification de phrases basée sur des arbres syntaxiques (SBMT)

Simplifier une phrase nécessite d'effectuer diverses opérations de réécriture structurelle telles que la division, les transformations des formes passives en formes actives ou la réorganisation des constituants de la phrase. Afin de capturer ce type de transformations, des travaux antérieurs ont effectué des simplifications en s'appuyant sur les représentations d'arbres syntaxiques (*Syntax-Based MT)*. La différence avec la méthode précédente est qu'elle manipule des unités syntaxiques complètes au lieu de mots ou de syntagmes seuls, incorporant ainsi une représentation explicite de la syntaxe dans les systèmes de TA (comme l'ordre des mots par exemple), et permettant donc d'effectuer de meilleures opérations de réorganisations.

L'approche de simplification par transduction d'arbres vise à surgénérer des règles de simplification automatiquement, ensuite à choisir celles qui correspondent le mieux à la phrase d'entrée. Le modèle de simplification de Zhu et al., (2010), est le premier modèle statistique qui gère la division, la suppression, la réorganisation et la substitution, couvrant ainsi la simplification lexicale et syntaxique dans le même modèle. Il opère sur l'arbre d'analyse de la phrase, et les auteurs implémentent chaque opération indépendamment avec un ensemble de règles et de fonctionnalités spécifiques à la tâche.

Woodsend et Lapata (2011) apprennent les règles de réécriture de la grammaire quasi-synchrone en utilisant EW-SEW et l'historique de révision de SEW. Paetzold et Specia (2013) ont développé un système exploitant le *Tree Transducer Toolkit* pour apprendre des règles de transformation syntaxique à partir d'un corpus parallèle de textes syntaxiquement analysés dans leurs formes originales et simplifiées. Les règles acquises sont appliquées aux phrases d'entrée analysées à l'aide de l'analyseur de constituants de Stanford, ce qui donne en sortie les différentes règles candidates (lexicales, syntaxiques, lexico-syntaxiques). Un module sélectionne ensuite les transformations de suppression ou de découpage. Le module de simplification génère à partir d'une phrase complexe en entrée et des règles précédemment sélectionnées, des phrase simplifiées candidates. Finalement, le module de classement attribue des scores aux différentes candidates et les ordonne pour sélectionner la meilleure.

Les conclusions d'Alva-Manchego et al., (2017) suggèrent que la conservation des phrases originales rend difficile l'apprentissage des opérations de simplification abstraite pour un système PBMT standard qui prend de nombreuses décisions locales. En réponse à ces problèmes, Xu et al., (2016) proposent des moyens d'adapter les systèmes de traduction automatique statistique à la tâche de simplification. Ils intègrent des caractéristiques supplémentaires spécifiques à la simplification telles que la longueur en caractères et en mots, le nombre de syllabes ou la proportion de mots anglais courants. Ces caractéristiques sont ensuite utilisées pour développer des règles de paraphrase à des fins de simplification. Leurs contributions incluent le développement de fonctions et métriques objectives spécifiques à la simplification, ainsi qu'une méthode de développement des règles de paraphrase pour la simplification.

Cependant, Narayan et Gardent (2014 ; 2015) montrent que les approches de la simplification syntaxique basées uniquement sur la syntaxe sont confrontées aux problèmes d'identification du point de découpage, de reconstruction de l'élément partagé en phrases découpées et de la suppression des arguments. Ils démontrent que l'usage de la sémantique dans la simplification syntaxique peut s'avérer nécessaire, notamment dans la construction de la deuxième phrase, dans l'identification de l'élément partagé. Leurs travaux seront développés dans la section 2.2.3.

Il a été démontré que l'utilisation d'alignements automatiques de EW-SEW produit des données d'apprentissage contenant des bruits avec des alignements où la version simplifiée est conservée et n'est pas plus simple ou n'est pas liée à la phrase complexe (Xu et al., 2015). Grâce à la disponibilité du corpus Newsela, la traduction automatique neuronale est devenue une alternative pour la simplification de phrases en anglais. Ce corpus est devenu une des principales ressources car il propose des simplifications manuelles à plusieurs niveaux, du niveau 0 comprenant le texte original au niveau 4 comprenant le texte le plus simplifié. La section suivante décrit les travaux qui ont adopté cette approche.

2.2.2.3. Simplification de phrases basée sur les approches neuronales (NMT)

Les systèmes de TA neuronale ont deux composants principaux, un encodeur et un décodeur. Le premier itère sur une séquence de mots représentés par des vecteurs d'intégration de grande dimension et calcule un vecteur d'état, de sorte qu'à la fin de l'itération, le vecteur d'état est une représentation de taille fixe de la séquence entière (généralement une phrase). Le décodeur génère alors des mots de sortie sur la base de cette représentation d'entrée ainsi que de la sortie précédemment générée.

Les avantages de l'approche neuronale par rapport à la traduction automatique statistique résident dans le fait de produire une sortie plus grammaticale ainsi qu'une meilleure capture des dépendances à longue distance. Cependant, un des inconvénients de l'approche neuronale est que l'apprentissage est plus lent, ainsi que le besoin de grands volumes de données en raison du grand nombre de paramètres généralement impliqués.

Le premier travail qui a utilisé des méthodes de traduction automatique neuronale pour la simplification de texte est celui de Nisioi (2017) qui a choisi d'apprendre des

simplifications à partir du corpus *Simple Wikipedia* compilé par Hwang (2015). Bien que ce modèle soit le premier qui peut effectuer conjointement une simplification lexicale et une réduction de contenu, il est encore limité à cause de la conservation de la phrase d'entrée.

Zhang et Lapata (2017) utilisent les données de Newsela et étendent les premiers travaux en optimisant directement les métriques pertinentes pour la simplification, notamment SARI (Xu et al., 2016), grâce à l'apprentissage par renforcement. Scarton et Specia (2018) proposent un modèle de simplification neuronale qui prend comme entrée supplémentaire un niveau de simplicité souhaité, ce qui leur permet dans une certaine mesure d'adapter la sortie à des lecteurs spécifiques. Surya et al. (2018) présentent une première tentative de simplification neuronale non supervisée du texte. Leur motivation était de concevoir une architecture qui pourrait être exploitée pour former des modèles SS pour des langues ou des domaines qui n'ont pas de grandes ressources d'instances originales simplifiées parallèles. Le modèle reconstitue d'une part l'entrée originale complexe à partir de la phrase simplifiée, et de l'autre part, il génère une version simplifiée de l'entrée.

Guo et al., (2018) ont développé un système de SS dans un cadre d'apprentissage multitâche (*Multi-Task Learning* MTL). Considérant SS comme la tâche principale, ils ont incorporé deux tâches supplémentaires pour améliorer les performances du modèle : la génération de paraphrases et la génération d'implications. Le premier aide à induire des remplacements de mots et de phrases, des réorganisations et des suppressions ; tandis que le second garantit que la sortie simplifiée générée suit logiquement la phrase d'origine.

Shardlow et Nawaz (2019) ont développé un système de simplification de texte pour faciliter les communications des spécialistes de la santé avec leurs patients. Leur méthode est basée sur une méthode neuronale de simplification de texte entraînée sur des textes de EW et SEW. Ils ont étendu le modèle en ajoutant un tableau de phrases pour fournir des simplifications de la terminologie médicale spécialisée extraite d'une ontologie médicale.

2.2.3. Simplification syntaxique basée sur les représentations sémantiques

En simplification syntaxique, décider *quand* et *où* diviser une phrase en deux phrases plus simples, associées à des évènements distincts, peut être déterminé par la syntaxe même, par exemple, des phrases avec des subordonnées. Cependant les approches de la simplification syntaxique essentiellement basées sur la syntaxe n'arrivent souvent pas à reconstruire correctement l'élément partagé ni à identifier correctement le point de découpage (Narayan et Gardent, 2014). Ces derniers auteurs présentent une approche combinant une sémantique profonde pour la division et la suppression de phrases et une traduction automatique monolingue pour la substitution et la réorganisation. Cette approche est basée sur la sémantique, en ce sens qu'il prend en entrée une représentation sémantique profonde, *Discourse Representation Structure* (DRS) pour la phrase complexe,

puis modifient cette représentation à l'aide d'un modèle probabiliste qui supprime des parties de la phrase et identifie les points de division afin de produire un ensemble de phrases plus simples. Les représentations DRS sont produites en utilisant Boxer (Curran et al., 2007), un analyseur automatique attribué. Dans une deuxième étape, les composants découpés sont ensuite reformulés en les complétant avec des éléments manquants afin de reconstruire des phrases grammaticalement correctes. Ces phrases plus simples sont encore simplifiées à l'aide d'un système de traduction automatique qui traite la substitution et la réorganisation. Par exemple, dans la phrase (21a), *brick* est impliquée dans deux événements : "*being resistant to cold*" et "*enabling the construction of permanent buildings*."

(21) a. Original: Being more resistant to cold, bricks enabled the construction of permanent buildings.
b. Simplified: Bricks were more resistant to cold. Bricks enabled the construction of permanent buildings.

Narayan et Gardent (2015) proposent une méthode qui ne nécessite pas de phrases originales simplifiées alignées pour former un modèle de simplification de textes. Leur approche utilise d'abord un simplificateur lexical sensible au contexte (Biran et al., 2011) qui apprend les règles de simplification des articles de EW et SEW. Étant donné une phrase originale, ces règles sont appliquées et la meilleure combinaison de simplifications est trouvée en utilisant la programmation dynamique. Ensuite, ils utilisent Boxer (Curran et al., 2007) pour extraire la représentation sémantique de la phrase et identifier les événements/prédicats. Après cela, ils estiment la vraisemblance maximale des séquences d'ensembles de rôles sémantiques qui résulteraient après chaque découpage possible (c'est-à-dire une sous-séquence d'événements). Pour calculer ces probabilités, ils s'appuient uniquement sur les données de SEW. Enfin, ils utilisent une programmation linéaire en nombres entiers pour déterminer quels constituants de la phrase doivent être supprimées.

Sulem et al., (2018a) décrivent un autre cadre de simplification structurelle basé sur la sémantique qui suit une approche similaire. Ils présentent un algorithme de division basé sur un analyseur sémantique automatique. Après la division, des opérations de simplification lexicale sont appliquées. Ils proposent une méthode pour effectuer la division de phrases, en présentant *Direct Semantic Splitting* (DSS), un algorithme basé sur un analyseur sémantique de division de phrase en ses constituants sémantiques principaux, un formalisme appelé *Universal Conceptual Cognitive Annotation* (UCCA ; Abend et Rappoport, 2013). Pour générer les structures de l'UCCA, l'analyseur TUPA basé sur la transition est utilisé.

Niklaus et al., (2019) présentent une approche pour diviser et reformuler des phrases anglaises complexes en une nouvelle hiérarchie sémantique de phrases simplifiées. Ces phrases ayant une structure plus simple et plus régulière sont plus faciles à traiter pour les tâches de TAL, telles que la traduction automatique ou l'extraction d'informations. Le cadre de simplification proposé prend une phrase en entrée et effectue une étape de transformation récursive basée sur 35 règles décrites manuellement, construite suite à

une analyse linguistique approfondie des phénomènes syntaxiques (les coordinations, les propositions adverbiales, les relatives, les phrases rapportées, les appositions, etc.). Chaque règle définit comment découper et reformuler l'entrée en phrases structurellement simplifiées (sous-tâche 1), établir une hiérarchie contextuelle entre les éléments décomposés (sous-tâche 2) et identifier la relation sémantique qui existe entre ces éléments (sous-tâche 3).

L'ensemble de ces travaux a montré que les informations sémantiques sont importantes dans la SS. Les opérations de division et de suppression sont conduites par la sémantique : la division est déterminée par les rôles sémantiques qui sont associés à un élément alors que la suppression d'un nœud est déterminée par ses relations sémantiques avec les événements divisés, d'où un modèle de suppression qui doit distinguer entre les arguments et les modifieurs. Ces travaux constituent des pistes pour notre travail, dans le sens où nous utilisons une approche basée sur l'analyse sémantique pour la simplification syntaxique.

2.3. Synthèse des différentes approches

Comme le soutient Siddharthan (2006), les approches basées sur des règles sont utiles dans le domaine de la simplification du texte lorsqu'un système se concentre sur des structures et des phénomènes linguistiques très spécifiques qui sont relativement faciles à gérer avec un ensemble de règles limité. L'avantage de cette approche est que le développeur peut facilement généraliser les phénomènes observés ou mettre en œuvre des phénomènes non vus dans un corpus mais susceptibles de se produire. De plus, contrairement aux approches à base d'apprentissage, les résultats des approches à base de règles permettent d'interpréter, d'expliquer comment les transformations des phrases à simplifier sont réalisées par le système, permettant ainsi d'analyser les erreurs et leur origine dans le but d'améliorer la méthode et son implémentation. Cependant, les systèmes basés sur des règles ont souvent une couverture très limitée et ne détectent pas de variations fines de la forme de surface. De plus, l'application « aveugle » de certaines règles de simplification peut nuire à la cohérence du texte : « *There are various discourse level issues that arise when carrying out sentence-level syntactic restructuring. Not considering these discourse implications could result in the simplified text losing coherence, or even changing the intended meaning, in either case, making the text harder to comprehend* » (Saggion, 2017).

Les méthodes d'apprentissage en traduction automatique ont conduit un certain nombre de chercheurs à essayer d'adopter cette technologie pour la simplification de textes. Les avantages des approches basées sur l'apprentissage automatique (surtout les méthodes neuronales) résident dans une sortie plus simple et grammaticale. Par rapport aux approches à base de règles développées manuellement, les approches basées sur les données peuvent effectuer simultanément plusieurs transformations de simplification, ainsi que l'apprentissage de modèles de réécriture très spécifiques et complexes. Cependant, ces méthodes ont besoin de grands volumes de données d'apprentissage requis par le grand nombre de paramètres impliqués. Le volume limité de données

standard/simplifiées parallèles disponibles constituent un défi pour la simplification basée sur l'apprentissage automatique pour la plupart des langues. Alva-Manchego (2020) soutient que les méthodes statistiques (SMT) peuvent effectuer des substitutions, des réorganisations à courte distance et des suppressions, mais ne parviennent pas à produire des divisions de qualité à moins d'être explicitement modélisées à l'aide d'informations syntaxiques ou couplées à d'autres processus, tels que l'analyse sémantique. Les approches basées sur la syntaxe peuvent modéliser les divisions de phrases (et les changements syntaxiques en général) plus naturellement, mais le processus de sélection des règles à appliquer, dans quel ordre et comment déterminer le meilleur résultat de simplification est complexe.

Les techniques de traduction automatique neuronale (NMT) ont dominé le domaine de la simplification de textes au cours des dernières années, en produisant de meilleurs résultats (meilleure préservation de la structure grammaticale et du sens). Cependant, il existe encore des défis avec ces approches, entre autres, la détection d'entités nommées dans le texte source (Štajner et Saggion, 2018).

Les méthodes sémantiques sont aussi utilisées pour la simplification automatique mais sont souvent adoptées dans des approches hybrides : des formalismes pour représenter le texte et des schémas de représentation conceptuelle ont été proposés pour décomposer le texte en unités plus petites pouvant être combinées ultérieurement pour générer d'autres textes en tenant compte des besoins des utilisateurs. Les résultats ont montré que les formalismes sémantiques peuvent être appropriés pour générer de nouvelles variantes et passer d'un texte à sa version simplifiée. Ces méthodes représentent les concepts et leurs relations avec l'ensemble des phrases composant un texte entier. Le Tableau 2 récapitule les points forts et les points faibles de chaque méthode.

Approche	Points forts	Points faibles
À base de règles – *Chandrasekar et al., (1996)* – *Carroll et al., (1998)* – *Siddharthan (2006)* – *Seretan (2012)* – *Brouwers et al., (2014) Saggion et al., (2015)*	– Concentration sur des constructions linguistiques très spécifiques faciles à gérer avec un ensemble de règles limité. – Offrir la possibilité d'expression des conditions sur l'applicabilité des règles. – La non-couverture d'une structure syntaxique laisse la phrase inchangée.	– Peut nuire à la cohérence du texte (application aveugle des règles). – Couverture très limitée et ne détectent pas de variations fines de la forme de surface. – Coûteuse : consomment beaucoup de temps et elles requièrent beaucoup d'implication humaine pour la définition des règles.
Traduction Automatique Statistique (SMT) – *Specia (2010)* – *Coster et Kauchak (2011 a,b)* – *Wubben et al. (2012)* – *Zhu et al., (2010)*	– Contrairement aux approches à base de règles : – Possibilité d'effectuer simultanément plusieurs transformations de simplification (substitutions/ réorganisation/suppression). – Le système peut effectuer plusieurs taches : simplification lexicale et syntaxique.	– Besoin de grands volumes de données d'apprentissage parfois non disponibles. – Mal dans la production de divisions de qualité (sauf si couplé avec d'informations syntaxiques ou sémantique). **Pour PBMT :** Nombre élevé de phrases conservées, non simplifiées. **Pour SBMT :** – Problèmes d'identification du point de découpage – Reconstruction incorrecte de l'élément partagé parfois dans les phrases découpées – Suppression des arguments.
Traduction Automatique Neuronale (NMT) – *Nisioi et al. (2017)* – *Zhang and Lapata (2017)* – *Surya et al. (2018)* – *Shardlow et Nawaz (2019*	Production une sortie plus grammaticale ainsi qu'une meilleure capture des dépendances à longue distance par rapport à la SMT.	L'apprentissage est plus lent, ainsi que le besoin de grands volumes de données en raison du grand nombre de paramètres généralement impliqués
À base de représentations sémantiques – *Narayan et Gardent (2015)* – *Narayan et al., (2017)* – *Sulem et al. (2018a)*	– Résolution du problème de réécriture du sujet partagé et de la reconnaissance du point de découpage. – Distinction entre les arguments et les modifieurs. Les arguments ne seront pas modifiés.	– Ne peut s'appliquer que pour les opérations de suppression et de découpage de phrases. – Un nombre d'erreurs élevé lié aux outils relatifs au formalisme utilisé (analyseur, générateur, etc) qui peut influencer négativement sur les résultats même si le système de simplification est performant.

Tableau 2 - Les points forts et faibles des méthodes utilisées en SS

Dans le but de comparer les performances des différents systèmes de simplification de textes existants, Alva-Manchego (2020) a évalué les sorties de ces systèmes en utilisant un ensemble de métriques d'évaluation automatique. Il utilise des métriques inspirées de la Traduction Automatique (par exemple, BLEU), des métriques de lisibilité (par exemple, Flesch Kincaid) et des métriques de simplicité (par exemple, SARI). Il explique que les mesures automatiques sont faciles à calculer, mais elles ne fournissent que des scores de performance globaux qui ne peuvent pas expliquer les forces et les faiblesses spécifiques d'une approche de SS. Par conséquent, il propose d'évaluer également les modèles SS en fonction de leur efficacité à exécuter des transformations de simplification spécifiques. En comparant les sorties des modèles sur l'ensemble de test PWKP, le système à base de représentations sémantiques Hybrid (Narayan et Gardent, 2014) est le meilleur modèle global, car il obtient les scores BLEU (cf. 6.2.1.1), SARI (cf. 6.2.1.2) et SAMSA (cf. 6.2.1.3) les plus élevés proches des plus élevés, et un FKGL pas trop éloigné de la référence.

2.4. Conclusion

Nous avons présenté dans ce chapitre les travaux sur la simplification syntaxique. Nous avons décrit les opérations de transformation de phrases, telles que la division, la fusion, la réorganisation de phrases, l'insertion et la suppression. Nous avons également dressé un état de l'art des différentes approches utilisées dans les systèmes de SAT, notamment les approches à base de règles développées manuellement, les approches relevant de la traduction automatique statistique et neuronale. Enfin, une synthèse sur les différentes approches en simplification syntaxique a été proposée dans le but de déterminer les forces et les faiblesses de chacune des approches utilisées du domaine. Les méthodes à base de représentations sémantiques montrent que les informations sémantiques sont importantes dans la SS. Les opérations de division et de suppression sont conduites par la sémantique : la division est déterminée par les rôles sémantiques qui sont associés à un élément alors que la suppression d'un nœud est déterminée par ses relations sémantiques avec les événements divisés, d'où un modèle de suppression qui doit distinguer entre les arguments et les modifieurs. Dans cette thèse, nous développons un système de simplification syntaxique en utilisant une approche utilisant une représentation sémantique à base de graphes exprimée en *Dependency Minimal Recursion Semantics* (DMRS).

Dans le chapitre suivant, nous présenterons un aperçu sur les formalismes grammaticaux à l'origine des formalismes de représentations sémantiques de texte. Nous développons ces formalismes tout en mettant l'accent sur ceux à base de graphes, nous en décrivons leurs caractéristiques, leurs forces et leurs faiblesses.

3. Relation entre la syntaxe et la sémantique

Tesnière (1959) et Chomsky (1965) postulent que la syntaxe est indissociable de la sémantique d'un énoncé. Selon ce dernier (1955), "*Linguistic theory has two major subdivisions: syntax and semantics*". La relation entre les deux peut être établie par le biais d'une interface syntaxe/sémantique. Chaque domaine aurait deux niveaux d'analyse.

Selon la Théorie Sens-Texte (Mel'čuk, 1988), au niveau de la syntaxe on peut distinguer : le niveau *profond* (appelé niveau tectogrammatical) et le niveau *superficiel* (Sgall et al., 1986 ; Mel'čuk, 1988). Dans le cadre de cette théorie, Mel'čuk définit la structure syntaxique profonde comme une structure intermédiaire entre la structure sémantique et la structure syntaxique de surface. Elle associe à une phrase une ou plusieurs structures qui représentent le sens de la phrase. Le but principal est de proposer, toujours avec des relations syntaxiques entre les unités lexicales, une représentation plus proche du sens de la phrase. La structure syntaxique profonde est représentée par un graphe acyclique dirigé dont les nœuds sont étiquetés par des lexèmes sémantiques (Kahane, 2003). À l'origine, cette représentation faisait usage d'un arbre de dépendances, mais les liens de coréférence explicités à ce niveau produisent un graphe acyclique dirigé et non un arbre. De plus, les lexèmes sémantiques sont dits « pleins » ; les auxiliaires, les conjonctions et les prépositions sont considérés comme sémantiquement vides et ne se retrouvent que dans la structure de surface. Les lexies profondes et les grammèmes profonds sont les nœuds et les relations universelles de dépendance syntaxique sont les branches.

Polguère (1998) définit la *syntaxe de surface* comme un arbre de dépendances, où toutes les lexies de la phrase y sont présentes et proches de la forme phonologique de la phrase. Elle fait intervenir tous les mots en prenant en compte leur ordre. Cette représentation consiste en un arbre de dépendances syntaxiques, dont les nœuds représentent des lexèmes et les arcs des dépendances syntaxiques de surface. La syntaxe profonde est une abstraction de la syntaxe de surface qui vise à se rapprocher du niveau sémantique (Perrier et al., 2014).

Pour ce qui est de la sémantique, les deux niveaux d'analyse sont : 1) l'analyse *sémantique de surface*, où la question est de déterminer les prédicats d'un texte et préciser pour chacun qui a fait quoi à qui, où et comment ; 2) l'analyse *sémantique profonde*, qui associe à une phrase une ou plusieurs structures de représentation de sens. La sémantique profonde qui nous intéresse dans le cadre de cette thèse est la sémantique vériconditionnelle compositionnelle. Une telle structure est qualifiée de représentation sémantique sous-spécifiée. Une phrase est caractérisée par ses conditions de vérité, où la formule logique est construite compositionnellement, à partir de la contribution du sens des mots de la phrase et du mode de composition syntaxique de ces mots.

Dans la plupart des formalismes grammaticaux, l'analyse sémantique et syntaxique sont étroitement liées. On distingue trois principales façons pour ce faire : « **(i) l'intégration** des processus d'analyse sémantique et syntaxique, caractérisée par la présence d'éléments syntaxiques et sémantiques, au sein d'une même structure. Les deux analyses peuvent être conduites dans un même processus » (Morey, 2011). Les exemples les plus typiques qui nous intéressent de cette approche sont fournis par les grammaires du formalisme *Head-driven Phrase Structure Grammar* (HPSG ; cf. 3.2.1), comme la grammaire *English Resource Grammar* (ERG ; cf. 3.2.2) de l'anglais. Chaque structure de la grammaire contient des objets sémantiques et syntaxiques. Un seul processus d'analyse compose ces structures et construit ainsi à la fois les analyses syntaxique et sémantique de la phrase. (ii) **la parallélisation** des deux processus d'analyse et « elle est caractérisée par la séparation des structures sémantiques et syntaxiques et par la synchronisation entre les opérations de composition syntaxique et sémantique ». Les deux analyses peuvent être conduites dans deux processus menés en parallèle qui se contraignent mutuellement, (iii) **la séquentialisation** des deux processus où « l'analyse sémantique est effectuée à partir du résultat de l'analyse syntaxique. Elle est caractérisée par l'absence de synchronisation explicite entre les processus de composition syntaxique et sémantique » (Morey, 2011). La représentation *Minimal Recursion Semantics* (cf. 3.6.3.1) que nous utilisons dans le cadre de cette thèse est construite compositionnellement en parallèle avec l'analyse syntaxique HPSG d'une phrase.

Dans ce chapitre, nous présentons un aperçu de la syntaxe de surface (section 3.1) et de la syntaxe profonde (section 3.2), la différence entre les deux, ainsi que la structure argumentale (section 3.3). Nous décrivons la tâche d'analyse sémantique des textes (section 3.4), ainsi que les différents formalismes sémantiques existants, à base de cadres (section 3.5) et à base de graphes (section 3.6), tout en mettant l'accent sur la famille MRS faisant partie du projet DELPH-IN, une initiative linguistique qui permet la mise en œuvre des grammaires HPSG disponibles pour une variété de langues et sur DMRS qui fait partie de cette initiative. La dernière section (section 3.7) de ce chapitre est consacrée à une comparaison entre ces formalismes, leurs façons de représentation, leur structure et les informations qu'annote chacun.

3.1. La syntaxe de surface

Polguère (1998) définit la structure syntaxique de surface comme un arbre de dépendance non linéairement ordonné : les nœuds représentent des lexies (pleines ou vides) et les arcs des dépendances syntaxiques de surface (liées à la langue en question). La syntaxe de surface fait le lien entre tous les mots de la phrase, les mots grammaticaux et les mots sémantiquement pleins, en ne tenant compte que des dépendances directes et explicites entre eux. Ces dépendances se réalisent par des positions canoniques dans la phrase ou des traits morphologiques telles que les marques de cas ou d'accord (Bonfante et al., 2018). D'après Chaumartin (2008), les structures de dépendance de surface sont des arbres de dépendances syntaxiques non ordonnés. Tous les mots, y compris les marqueurs de ponctuation de la phrase originale, sont représentés par un nœud dans

l'arbre. Les nœuds sont étiquetés avec des lemmes, des étiquettes de POS, des informations morphosyntaxiques telles que le temps et le nombre. Les étiquettes de POS sont presque les mêmes que celle du Penn Treebank POS. Les liens entre les nœuds sont étiquetés avec des étiquettes syntaxiques. Il est à noter que la structure syntaxique de surface est fortement liée à l'ordre linéaire des mots de la phrase, alors que le niveau profond ne l'est pas.

Les corpus annotés avec des informations syntaxiques comme les *treebanks* syntaxiques sont largement utilisés dans diverses tâches de TAL. Ils sont composés des phrases annotées en arbres syntaxiques et une importante quantité de projets existe en langues variées. Des travaux d'annotation de corpus intégrant des informations "non surfaciques" (soit les arbres syntagmatiques, soit des arbres de dépendances) ont été développées. L'un des premiers *treebanks* en dépendance est le *Prague Dependency Treebank* (Böhmová et al., 2003) en langue tchèque. Pour l'anglais, le *Penn TreeBank* (Marcus et al., 1993 ; 1994) est la première banque d'arbres développée à grande échelle et qui utilise l'analyse syntaxique en constituants. Elle fournit des annotations de phrases en anglais : des informations relatives aux parties du discours (POS) et aux relations syntaxiques entre les mots (fonctions syntaxiques, dépendances).

L'un des plus grands projets aujourd'hui est *Universal Dependencies* (Nivre et al., 2016). Ce projet, qui constitue l'ensemble de *treebanks* en dépendance le plus grand et le plus utilisé, propose une collection de *treebanks* dans plus de 100 langues, développées par des équipes du monde entier (d'où le terme « *universal* »). Depuis sa création, UD est exploitable dans les tâches d'analyse et d'annotation automatique et elle est utilisée jusqu'à ces jours dans des projets en typologie linguistique quantitative et l'analyse multilingue. On peut aussi citer le projet *Surface Universal Dependencies Treebanks* (SUD), qui consiste à convertir les treebanks d'UD dans une structure de dépendance de surface (basée sur des critères distributionnels), proposé par Gerdes et al. (2018 ; 2019). Dans SUD, les mots fonctionnels constituent les têtes des syntagmes dans la mesure où ils déterminent leur distribution. Il s'agit des auxiliaires, des prépositions et des conjonctions de subordination. Les conjonctions de coordination et les déterminants 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 » (Perrier, 2021).

Pour le français, on trouve le corpus Sequoia, un corpus arboré annoté en suivant le schéma d'annotation du French Treebank en constituants (Abeillé et al., 2003). Il contient 3 099 phrases (Candito et Seddah, 2012). C'est une ressource exploitable pour des projets de linguistique informatique, théorique ou appliquée. Guillaume et al. (2019) ont abouti à la conversion automatique vers le schéma d'annotation des *Universal Dependencies*, obtenant ainsi un corpus arboré dans le schéma en constituants du FTB, dans le schéma en dépendances FTBdep, et dans le schéma UD, ce qui permet une description fine adaptée au français, ainsi qu'une description adaptée aux traitements multilingues (Candito, 2022).

3.2. La syntaxe profonde

Le niveau tectogrammatical représente la syntaxe profonde de la phrase (Sgall, 1992 ; Tesnière, 1959). La syntaxe profonde étudie le lien entre le niveau syntaxique et le sens. Candito (2022) trouve que le terme syntaxe profonde est « très "chargé" d'un point de vue théorique, puisqu'il rappelle les temps de la grammaire générative transformationnelle, où une représentation profonde était centrale, et la représentation de surface des phrases obtenue par application de transformations ». Dans leur livre, Kahane et Gerdes (2022) définissent la syntaxe profonde comme la représentation qui s'intéresse à la correspondance entre la structure sémantique et la structure syntaxique de surface, c'est-à-dire à *l'interface sémantique-syntaxe*. Cette correspondance est décrite au travers d'une structure qu'on appelle la *structure syntaxique profonde* (Kahane et Gerdes, 2022). Les auteurs affirment que cette structure peut être vue essentiellement comme une « projection de la structure prédicative sur la structure syntaxique de surface et donc comme une structure syntaxique de surface ». Néanmoins la syntaxe profonde indique à la fois les connexions syntaxiques et les relations prédicat-argument, qui ne se superposent pas toujours aux connexions syntaxiques. Ils montrent que la structure syntaxique profonde, à l'inverse, peut être vue comme une « projection de la structure syntaxique sur la structure prédicative, c'est-à-dire comme une *structure sémantique hiérarchisée* ». Elle s'intéresse à la correspondance entre la structure sémantique et la structure syntaxique de surface (Kahane et Gerdes, 2022).

Parmi les projets développés, on peut citer les *enhanced dependencies*[21] (Schuster et Manning, 2016), proposées dans le cadre des *Universal Dependencies*. Ils comprennent toutes les dépendances argumentales qui ne sont pas représentées dans UD, telles que les sujets d'infinitifs. Elles sont appelées des dépendances profondes par opposition à celles des UD qui sont considérées comme des dépendances de surface (Guillaume et al., 2019).

Contrairement au niveau analytique, les structures arborées tectogrammaticales présentent les caractéristiques suivantes (Ribeyre, 2016) :

- Un nœud peut représenter plus qu'un mot. Seuls les mots pleins représentent un nœud de l'arbre. Les auxiliaires et les prépositions coïncident avec les mots pleins.
- Comme la non-projectivité n'est pas autorisée, les nœuds sont alors ordonnés de telle sorte que le critère de projectivité soit respecté.
- Les fonctions de surface sont remplacées par les fonctions profondes (agent, patient, thème, etc.).
- Les nœuds de l'arbre contient le lemme du token et des traits morpho-syntaxiques.
- Un attribut est ajouté pour capturer l'articulation topic-focus (Sgall et al., 1986). Il peut prendre trois valeurs : T(opic), F(ocus), C(onstrast).

Il existe plusieurs formalismes de modélisation informatique de la syntaxe. Nous nous intéressons ici à la syntaxe fondée sur la théorie des modèles (*Model-Theoretic Syntax* MTS) (Rogers, 1996). Ce courant est représenté, selon Morey (2011), par les Grammaires

[21] https://universaldependencies.org/u/overview/enhanced-syntax.html

Syntagmatiques Guidées par les Têtes (*Head-driven Phrase Structure Grammar* - HPSG) (Pollard, 1999) et des Grammaires Lexicales Fonctionnelles (LFG) (Kaplan et Bresnan, 1995), ainsi que des formalismes purement MTS comme les Grammaires de Dépendances eXtensibles (XDG) (Debusmann, 2006), les Grammaires de Propriétés (GP) (Blache, 2004) et les Grammaires d'Interaction (IG) (Perrier, 2003).

Dans le cadre de la thèse, nous nous intéressons aux grammaires HPSG (section 3.2.1), à partir duquel a été conçu un formalisme de représentation sémantique, la MRS puis la DMRS (*Dependency Minimal Recursion Semantics*). Une grammaire HPSG est une grammaire fortement lexicalisée basée sur l'unification, qui utilise des structures de caractéristiques typées pour coder les informations syntaxiques (nous ne rentrons pas dans les détails des autres structures).

3.2.1. Head-driven Phrase Structure Grammar (HPSG)

HPSG est un exemple de formalisme basé sur l'unification, dans laquelle les objets linguistiques sont analysés en termes de structures d'information partielles qui peuvent être intégrées de manière récursive (Pollard et Sag, 1988). Une représentation de structure de caractéristiques HPSG d'une phrase est créée à l'aide d'un analyseur utilisant une unification pour combiner des structures de caractéristiques lexicales et syntaxiques. Le résultat d'une analyse en HPSG est une structure de traits synthétisant l'information dans une phrase, accompagnée de l'arbre des structures de traits construites au cours de l'analyse. Toutes les dépendances se trouvent dans la structure de traits finale. Dans la formulation originale, les contenus sémantiques sont analysés en tant qu'individus, relations, rôles, situations et circonstances. Les individus jouent des rôles dans les relations, qui sont syntaxiques mais assignent également des rôles sémantiques.

Il s'agit d'un formalisme qui est issu de plusieurs courants théoriques et relève des grammaires d'unification dans laquelle les objets linguistiques sont analysés en termes de structures d'information partielles qui peuvent être intégrées de manière récursive (Pollard et Sag, 1988). Il donne des descriptions uniformes des différentes dimensions du langage. Les entrées lexicales, les règles de grammaire, le contexte et les principes généraux de bonne formation fournissent des contraintes supplémentaires sur les structures. HPSG intègre les aspects syntaxiques et sémantiques de la théorie grammaticale sous l'hypothèse qu'aucun ne peut être bien compris isolément de l'autre.

Cette uniformité de la modélisation se manifeste en ce que le modèle de toute unité est construit sur le même patron quel que soit sa taille. Il s'agit d'une grammaire bidirectionnelle[22] se proposant de fournir un cadre de modélisation de principes grammaticaux universels, un formalisme grammatical pour représenter la correspondance entre la syntaxe et la sémantique.

[22] Les grammaires écrites dans ce modèle sont utilisées en analyse comme en génération.

La structure de base de la représentation syntaxique en HPSG est la structure de traits typés, une structure qui est constituée d'un ensemble de traits formant des couples attribut/valeur (cf. Figure 5).

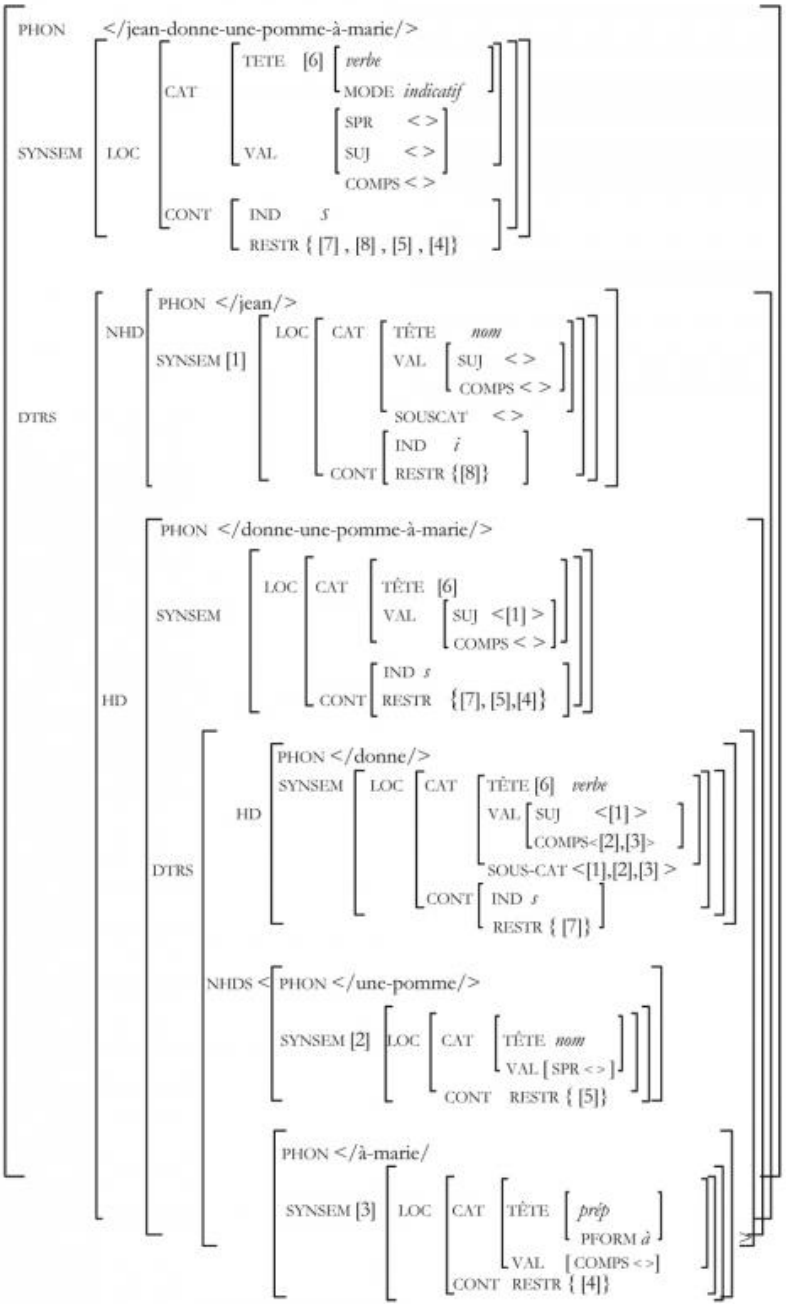


Figure 5 - HPSG de la phrase « Jean donne une pomme à Marie » (Desmets et al., 2003).

Une autre caractéristique essentielle de HPSG réside dans l'intégration des connaissances. L'objectif de regrouper dans une représentation des connaissances homogène des informations linguistiques variées est clairement annoncé. Utiliser les structures de traits comme cadre unique pour représenter des informations linguistiques intégrant des connaissances phonologiques, lexicales, syntaxiques, sémantiques et

pragmatiques. La Figure 5 est la structure HPSG de la phrase « Jean donne une pomme à Marie » (Desmets et al., 2003).

Les caractéristiques essentielles d'une structure de traits sont les suivantes (Johnson, 1988) :

- Les éléments de la structure sont atomiques ou complexes.
- La structure interne d'un élément est définie par ses attributs et ses valeurs.
- Les valeurs peuvent être partagées. Le mécanisme de partage de valeurs est central en HPSG et permet de décrire un grand nombre de relations syntaxiques.

English Resource Grammar (ERG ; Flickinger, 2000 ; 2012 ; Flickinger et al., 2014) est une implémentation de la théorie grammaticale HPSG. Dans la section suivante, nous développons cette ressource.

3.2.2. English Resource Grammar (ERG)

English Resource Grammar (ERG ; Flickinger, 2000) est une grammaire open-source de l'anglais, indépendante du domaine et à large couverture. Elle encapsule les connaissances linguistiques nécessaires pour produire de nombreux types d'annotations de sens compositionnel. L'ERG est une grammaire bidirectionnelle, ce qui lui permet d'être utilisée à la fois pour produire des *analyses* de phrases d'entrée, ainsi que pour produire des phrases à partir d'une représentation de sens d'entrée (c'est-à-dire, *génération*). Le cadre de représentation du sens utilisé par l'ERG et que nous utilisons dans cette thèse est *Minimal Recursion Semantics* (MRS ; Copestake et al., 2005) (cf. 3.6.3.1).

ERG est développé (1) en utilisant un corpus pour découvrir des phénomènes encore non traités par le HPSG, et (2) en étendant HPSG : il s'agit de coder une généralisation des phénomènes en utilisant leurs intuitions linguistiques et en consultant la littérature linguistique (Baldwin et al., 2005). ERG se compose d'un grand ensemble d'entrées lexicales sous une hiérarchie de types lexicaux, avec un ensemble réduit de règles lexicales pour la génération. ERG utilise une forme de représentation Davidsonienne dans laquelle tous les verbes introduisent des événements. Un effet de ceci est que les adverbes sont de deux classes (cf. 4.1.2) : les adverbes scopaux comme « *Probably* », qui prend des prédicats comme arguments et les adverbes non scopaux comme « *Quickly* », qui prend des événements comme arguments.

Les représentations sémantiques produites par l'ERG sont parfois appelées *English Resource Semantics* (ERS), une annotation sémantique associée à ERG. De nouvelles versions de la grammaire sont continuellement développées. Tout au long de cette thèse, nous travaillons avec la version ERG 2018.

3.3. Structure argumentale

La structure argumentale (ou structure prédicat-argument) indique le nombre d'arguments que possède un prédicat et leur relation sémantique envers ce prédicat. « La notion de structure argumentale, quoique globalement partagée, n'est pas univoque et

varie en fonction des théories surtout selon leur nature : plus syntaxique ou plus sémantique » (Ribeyre, 2016). Bresnan (2001) l'a définie comme suit : « *Argument structure is an interface between the semantics and syntax of predicators (which we may take to be verbs in the general case)… Argument structure encodes lexical information about the number of arguments, their syntactic type, and their hierarchical organization necessary for the mapping to syntactic structure* ».

Un cadre de représentation sémantique fournit une notation formelle qui dicte comment et quelles informations doivent être représentées, ainsi qu'une définition des règles pour fusionner diverses unités de sens dans sa notation. Au centre de tous les formalismes, se trouve l'identification de la structure prédicat-argument et la distinction entre les arguments et modificateurs.

L'une des représentations de sens les plus courantes est la *forme logique*, qui est basée sur des prédicats et le lambda calcul. Lorsqu'une phrase ou un paragraphe est complètement analysé et toutes les ambiguïtés sont résolues, sa signification est représentée sous une forme logique unique. Cependant, cela ne résout complètement que quelques cas simples : dans l'analyse sémantique, nous rencontrons souvent des structures complexes qui ne peuvent pas être capturées dans des structures arborescentes ou des expressions logiques simples, nécessitant le développement de représentations sémantiques plus avancées. *Les représentations sémantiques à base de graphes* offrent un moyen flexible de modéliser la sémantique du langage naturel. Elles capturent une vue plus complète de la signification des phrases que les méthodes superficielles (telles que l'étiquetage des rôles sémantiques pour PropBank et FrameNet ; cf. section 3.5), en décrivant les relations entre toutes les entités mentionnées dans une phrase dans une seule structure (qui fait quoi à qui, quand, où, pourquoi et comment dans un énoncé).

Dans le contexte de la tâche partagée *Cross-Framework Meaning Représentation Parsing* (MRP 2019 ; Oepen et al., 2019), nous distinguons différents types d'ancrage (cf. 3.7) de graphes sémantiques en fonction de la nature de la relation qu'ils assument entre la phrase et les nœuds du graphe. Nous appelons cette relation l'ancrage des nœuds sur des sous-chaînes (*anchoring*) ; d'autres termes couramment utilisés incluent l'alignement, la correspondance ou la lexicalisation.

Nous proposons d'exploiter ce type de représentations dans la tâche de simplification syntaxique des textes. Dans la section suivante, nous présentons la définition de l'analyse sémantique et quelques représentations sémantiques existantes.

3.4. Analyse sémantique

L'analyse sémantique (*Semantic parsing*) est une tâche de TAL qui consiste à convertir une phrase en langage naturel en une représentation sémantique (RS) structurée compréhensible par la machine. Kate et Wong (2010) proposent une définition de l'analyse sémantique comme la tâche de « *mapper* » des phrases en langage naturel dans des représentations de sens formelles qu'un ordinateur peut exécuter pour une application spécifique à un domaine. Le but est ainsi de convertir le langage naturel en

une représentation sémantique qui peut ou non être utilisée dans une tâche spécifique. L'exécution de l'analyse sémantique nécessite un formalisme (appelé aussi schéma) de représentation sémantique qui définit comment le sens doit être représenté.

L'analyse sémantique s'est avérée fondamentale dans de nombreuses tâches de Compréhension du Langage Naturel (*Natural Language Understanding* NLU), telles que le résumé de texte (Genest et Lapalme 2011), les systèmes de dialogue (Tur et al., 2005) et la traduction automatique (Liu et Gildea, 2010). Les représentations sémantiques constituent un intermédiaire entre l'expression naturelle d'une phrase et son traitement automatique. Ces représentations définissent des structures qui reflètent le sens de phrase tel qu'il est compris par un locuteur. Leur usage pour le TAL évolue rapidement et plusieurs tâches peuvent bénéficier des informations qu'apportent les structures sémantiques, telles que l'annotation sémantique (Pan et al., 2015), le résumé automatique de textes (Liu et al., 2018) ou l'extraction d'informations (Garg et al., 2015). « En n'utilisant que les informations syntaxiques et lexicales, les systèmes correspondant à ces tâches sont limités, car ils ne modélisent pas les interactions entre les différents éléments qu'ils ont extraits » (Michalon, 2017).

Au cours des dernières années, une grande variété de cadres sémantiques a été proposée, mettant l'accent sur la représentation des informations sémantiques (rôles sémantiques, sens des mots, relations entre les entités). Parmi ces projets, les projets PropBank (Kingsbury et Palmer, 2003 ; Palmer et al., 2005) et NomBank (Meyers et al., 2004) portent sur les propositions verbales et leurs arguments. Le projet FrameNet (Fillmore, 1976 ; Baker et al., 1998 ; Johnson et al., 2002) propose une base lexicale basée sur les cadres sémantiques (cf. 3.5).

Les formalismes diffèrent selon leur interaction avec la syntaxe, ou encore les informations sémantiques qu'ils couvrent. Comme en analyse syntaxique, on distingue deux grandes catégories de représentations sémantiques (Marzinotto, 2019) :

- ***Les représentations sémantiques*** **profondes** : elles associent à une phrase une (ou plusieurs) structures représentant le sens d'une phrase, à partir des contributions sémantiques de chacun des mots qui la composent. En TAL, l'analyse sémantique profonde s'inscrit la plupart du temps dans le cadre particulier de la sémantique vériconditionnelle compositionnelle. C'est pourquoi ce type de représentations est appelé aussi ***logique*** (*Logical Semantic Representation LSR)*. Il s'agit d'analyser des phrases dans un langage de représentation formelle du sens, comme le lambda calcul. L'analyse sémantique logique est assez rare et compliquée en raison du degré de détail dans lequel les textes sont traités et du manque de grands corpus annotés pour entraîner les systèmes d'apprentissage automatique. Pour cette raison, la plupart des approches existantes sont soit basées sur des grammaires (Wong et Mooney, 2007 ; Berant et al., 2013) soit basées sur des transformations de structures syntaxiques (Reddy et al., 2016 ; 2017). Considérons la phrase (22) suivante :

  (22) « Tout homme gentil aime une femme »

L'exemple (22) peut décrire deux situations du monde différentes, qui sont décrites par les deux formules logiques non équivalentes suivantes :

$$\forall x.\text{homme}(x) \wedge \text{gentil}(x) \rightarrow (\exists y.\text{femme}(y) \wedge \text{aime}(x, y))$$
$$\exists y.\text{femme}(y) \wedge (\forall x.\text{homme}(x) \wedge \text{gentil}(x) \rightarrow \text{aime}(x, y))$$

– ***Les représentations sémantiques superficielles*** *(Shallow Semantic Representation SSR)* : consistent à identifier des entités dans une phrase et à leur attribuer les rôles qu'elles jouent dans une situation donnée. Elles consistent à déterminer dans une phrase les arguments d'un prédicat, selon la formule *qui a fait quoi à qui, où, quand et comment*. Cette tâche est appelée « étiquetage des rôles sémantiques » (ou *Semantic Role Labeling* SRL) (Màrquez et al., 2008). Ce type de représentation est également connu sous le nom de représentation sémantique de cadres car elle est basée sur la théorie de cadre (*frame semantics* ; Baker et al., 1998 ; Fillmore, 1982). La phrase (22) est analysée comme suit :

[$_{\text{Arg0}}$ Tout homme gentil] aime [$_{\text{Arg1}}$ une femme].

« Tout homme » est identifié comme l'argument 0 du verbe « aime » et « une femme » comme son argument 1, en utilisant le jeu d'étiquettes du corpus annoté par des rôles sémantiques PropBank. Les arguments 0 et 1 sont attribués aux arguments syntaxiques qui correspondent respectivement à un agent et un patient.

Plusieurs formalismes fournissant une analyse sémantique profonde existent, citons notamment *Semantic Dependency Parsing* **(SDP)** (Oepen et al., 2015), *Universal Conceptual Cognitive Annotation* (**UCCA)** (Abend et Rappoport, 2013), *Abstract Meaning Representation* (**AMR**) (Banarescu et al., 2013), etc. Chacun de ces formalismes a ses avantages et ses inconvénients, et le choix du schéma approprié dépend de l'application.

Il existe deux catégories principales de représentations sémantiques : ***représentations à base de cadre*** et ***représentations à base de graphes***. Elles sont détaillées respectivement dans les sections 3.5 et 3.6.

En TAL, le choix de représentations sémantiques dépend de l'application. Dans les applications de transformation des textes, la représentation sémantique doit prendre en compte certains aspects (Lamercerie, 2021) :

- *Large couverture de situations sémantiques* : une représentation de sens doit modéliser les événements, les relations et les états. Ainsi que l'étiquetage des rôles sémantiques dans une situation.
- *Unicité* : deux phrases qui évoquent le même sens doivent avoir la même RS.
- *Anaphore et coréférence* : si la même entité apparaît plusieurs fois dans la phrase, la RS doit permettre d'identifier les éléments anaphoriques.

- *Les relations temporelles et spatiales* : comprendre l'ordre des événements et déterminez la nature de la relation entre eux (causalité, simultanéité ou continuité).
- *Opérateurs logiques* : une RS doit modéliser les négations, les coordinations et les quantificateurs.

L'état de l'art comprend de nombreuses représentations qui divergent sur plusieurs aspects et les caractéristiques de ces représentations sont variées. Ainsi, leurs relations à la syntaxe et leurs liaisons aux mots s'exercent à des degrés différents. Par exemple, ***l'ancrage*** (relation entre la phrase et les nœuds du graphe) aux tokens d'une phrase donnée est *fort* pour les structures en dépendances, tandis qu'il est *faible*, voire inexistant, pour les formalismes basés sur la logique. Il en résulte des différences importantes dans leur capacité à s'abstraire des variations syntaxiques et conceptuelles, et à intégrer différents phénomènes linguistiques.

Lamercerie (2021) trouve que ces différences se traduisent en termes de formalisme, d'interface syntaxique et de degré d'abstraction : les structures sans ancrage sont généralement plus « faciles » à l'usage, tandis que les correspondances entre les tokens de la phrase et les éléments sémantiques facilitent l'analyse syntaxique.

Dans la suite, nous proposons un aperçu de quelques représentations sémantiques qui nous ont semblé pertinentes pour notre travail. Nous décrivons les caractéristiques les plus importantes de ces cadres, ainsi que leurs caractéristiques formelles et linguistiques spécifiques.

3.5. Formalismes de représentation sémantique de cadres

Dans le but de représenter le sens des phrases, nombreuses structures de représentation du sens sont développées. La sémantique de cadres (*Frame Semantics*) est une théorie sur la signification linguistique, qui relie la sémantique linguistique à la connaissance encyclopédique. Cette théorie stipule que pour comprendre le sens d'un seul mot, il faut avoir accès à toutes les connaissances qui se rapportent à ce mot. Par exemple, pour comprendre le sens de "*envoyer*", il faut être familier avec la situation d'"*envoyer*", et aussi avec les arguments importants de la situation tels que *l'expéditeur*, le *destinataire* et *ce qui est envoyé*. Ces cadres se trouvent dans une encyclopédie des cadres et décrivent le concept spécifique auquel le mot se réfère.

3.5.1. FrameNet

Le projet *FrameNet*[23] (Baker et al., 1998) correspond à la représentation en cadres sémantiques dont la théorie a été développée par Fillmore (1976 ; 1982). Cette théorie permet de représenter une situation et les éléments qui y interviennent sous forme de *cadres* d'une façon que deux situations similaires mais différentes dans leur lexique et leur

[23] https://framenet.icsi.berkeley.edu/fndrupal/

syntaxe soient représentées par le même objet, appelé *cadre sémantique*. FrameNet décrit le sens d'une phrase à travers des événements ou des scènes, ainsi que tous les éléments (ou rôles) qui peuvent être associés à ces événements dans un énoncé, appelés *éléments du cadre* (EC). Les cadres sémantiques sont composés d'un noyau autour duquel graviteraient un certain nombre d'éléments. Lorsqu'ils sont instanciés, le noyau des cadres ainsi que les éléments gravitant autour (partie sémantique) sont chacun associés à des éléments de la phrase (partie lexicale). La partie lexicale est appelée *unité lexicale* (UL), il s'agit du mot (ou groupe de mots) qui évoque le sens du cadre. Les mots associés aux éléments de cadres sont appelés *acteurs*. FrameNet fournit un ensemble de 1 878 cadres.

FrameNet a été utilisé dans des applications telles que les systèmes question-réponse (Agrawal et Mukherjee, 2019 ; Shen et Lapata, 2007), l'extraction d'informations (Barzdins, 2014) ou la paraphrase (Ellsworth et Janin, 2007). FrameNet est une représentation sémantique dépendant de la langue, au départ l'anglais, mais des versions de FrameNet dans d'autres langes ont récemment vu le jour, par exemple pour le français (Candito et al., 2014 ; Djemaa et al., 2016 ; Marzinotto et al., 2018), l'allemand (Burchardt et al., 2009) et l'espagnol (Subirats, 2009). Récemment, de nouvelles tentatives pour construire un FrameNet multilingue (Torrent et al., 2018) ont pris de l'ampleur dans la communauté FrameNet.

La Figure 6 illustre un exemple de phrase avec le Frame « *Sending* ». Ici, le Frame « *Sending* » a un grand nombre d'éléments de cadre[24], ce qui permet de spécifier des informations très détaillées telles que l'agent expéditeur (*Sender*), ce qui est envoyé (*Theme*) et le moyen (*Transport_means*), le lieu (*Place*), etc.

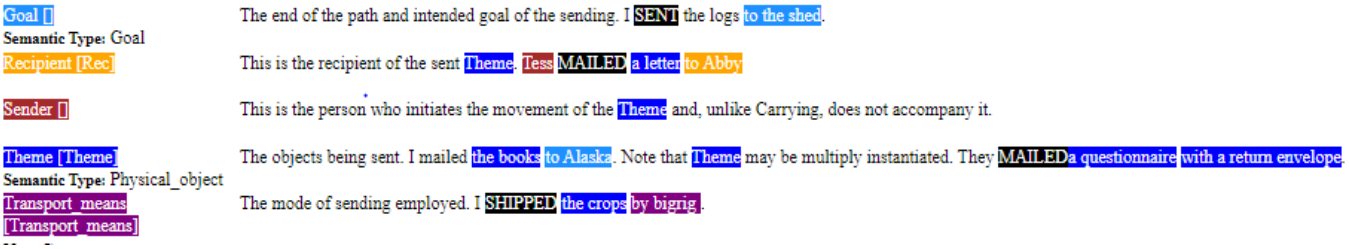


Figure 6 - Le cadre "*Sending*" de FrameNet.

3.5.2. PropBank

Le projet *Proposition Bank* (PropBank ; Palmer et al., 2005) est un autre formalisme de représentation sémantique basé sur les cadres. PropBank est similaire à FrameNet : il s'agit d'une représentation sémantique superficielle qui extrait une représentation du sens d'une phrase à l'aide de cadres et d'arguments, afin d'identifier dans un énoncé *qui a fait quoi à qui, quand, où et pourquoi*. PropBank ajoute une couche d'informations prédicat-argument, ou étiquettes de rôles sémantiques, aux structures syntaxiques du Penn Treebank (Kingsbury et Palmer 2002 ; Palmer et al., 2005). Le projet a commencé par marquer des noyaux de propositions composés de prédicats verbaux et de leurs

[24] https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Sending

arguments (structure prédicat-argument). Plus tard, des "modificateurs de variables d'événement" ont été rajoutés (Babko-Malaya et al. 2004), élargissant les structures prédicat-argument avec des compléments.

PropBank distingue deux types de rôles sémantiques : les arguments principaux communs à tous les verbes et les modificateurs d'arguments. Les arguments principaux sont nommés ARG0 à ARG4 : « A0 » (agent), « A1 » (patient), « A2 » (instrument, bénéficiaire ou attribut), « A3 » (point de départ de l'action, bénéficiaire ou attribut), « A4 » (point d'arrivée de l'action). Les modificateurs d'arguments comme par exemple de lieu (ARGM-Loc) ou de temps (ARGM-Time). Bien que ces rôles soient communs pour tous les verbes, ils ne sont pas obligatoirement utilisés.

Les représentations des informations sémantiques de PropBank sont utiles dans les tâches de TAL telles que la réponse aux questions (Shen et Lapata, 2007), le résumé multi-document (Genest et Lapalme, 2011), les systèmes de dialogue (Tur et al., 2005 ; Chen et al., 2013), l'extraction d'informations (Bastianelli et al., 2013) et la traduction automatique (Liu et Gildea, 2010). Dans la section suivante, nous décrivons les formalismes sémantiques à base de graphes.

3.6. Représentations sémantiques à base de graphes

L'analyse sémantique basée sur les graphes (*Graph-based meaning representation*) est une tâche dans presque tous les campagnes SemEval[25] depuis 2014. SemEval est une série d'ateliers de recherche internationaux sur le traitement du langage naturel (TAL) dont l'objectif est de faire progresser les recherches en analyse sémantique. Les représentations sémantiques (RS) à base de graphes représentent trois types de connaissances sur la sémantique : *le sens des mots*, *la structure des arguments de prédicat* (arêtes et étiquettes d'arêtes) et la *coréférence*. Les graphes sémantiques portent sur des éléments qui matérialisent une situation, un événement ou un objet de la représentation. Les relations entre les éléments sont représentées au travers de prédicats. Chacun possède des participants. Chacun reçoit alors un rôle thématique ou une fonction prédéfinie par le prédicat.

Un certain nombre de banques de graphes ont annoté de grands corpus avec des représentations sémantiques basées sur des graphes de divers types. La Conférence CoNLL 2019 (*Computational Natural Language Learning)* a accueilli la tâche *Cross-Framework Meaning Représentation Parsing* (MRP 2019 ; Oepen et al., 2019). L'objectif est de développer un système d'analyse unifié pour traiter cinq banques de graphes sémantiques : (i) *Abstract Meaning Representation* – **AMR** (Banarescu et al., 2013), (ii) *Universal Conceptual Cognitive Annotation* – **UCCA** (Abend and Rappoport, 2013), (iii) *Minimal Recursion Semantics* – **MRS** *(Copestake 2009)* et ses dérivés, (iv) *Elementary*

[25] https://semeval.github.io/

Dependency Structures – **EDS** (Oepen and Lønning, 2006) et enfin (v) *Prague Semantic Dependencies* – **PSD** (Hajič et al., 2012). Etant donné que notre travail se base sur ces types d'analyse, nous présentons dans les sous-sections suivantes chacun de ces formalismes, excepté PSD qui est très peu utilisé.

3.6.1. Abstract Meaning Representation (AMR)

La représentation de sémantique abstraite (*Abstract Meaning Representation* AMR) (Banarescu et al., 2013) est basée sur des graphes orientés à une seule racine avec des nœuds et des arêtes étiquetés. Chaque nœud du graphique représente un concept sémantique et les relations sémantiques sont spécifiées par les arcs. L'un des nœuds définit la racine du graphe. Les concepts peuvent être soit des mots anglais, des patterns PropBank, soit des mots-clés spécifiques permettant d'expliciter certains phénomènes linguistiques. Les prédicats AMR sont annotés sous forme de lemmes et d'étiquettes de sens, mais ils ne constituent qu'un sous-ensemble de concepts. Les auteurs donnent une liste complète d'exemples d'AMR pour des arguments de cadre, relations sémantiques générales, coréférence, relations inverses, modaux et négation, questions, verbes, noms, adjectifs, prépositions, entités nommées, copules et réification. Cette approche a été développée dans le but d'annoter de grandes quantités de données, sur lesquelles des systèmes d'apprentissage automatique peuvent s'entraîner sur cette tâche.

AMR peut être considéré comme un projet beaucoup plus « ambitieux » que FrameNet, puisque l'analyse sémantique de phrases entières peut être représentée (Michalon, 2017). Contrairement à FrameNet, AMR permet de représenter les informations de quantification, de coréférence ou encore de portée de la négation.

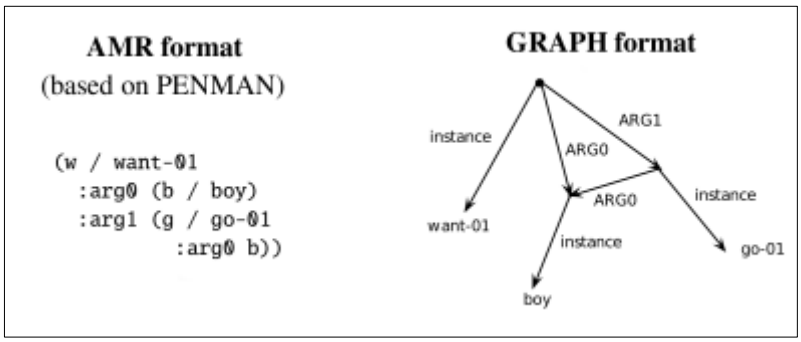


Figure 7 - Graphe AMR de la phrase « *The boy wants to go* ».

Cependant, le formalisme AMR présente certaines limites : il ne spécifie pas les inflexions morphologiques pour le temps et le nombre, l'aspect et omet les articles. La Figure 7 montre un exemple de graphe AMR avec deux cadres "*want-01*" et "*go-01*". Les nœuds sont associés aux concepts *want-01*, *boy* et *go-01*.

3.6.2. Universal Conceptual Cognitive Annotation (UCCA)

Universal Cognitive Conceptual Annotation (UCCA ; Abend et Rappoport, 2013) est une représentation sémantique s'appuyant sur la théorie linguistique de base (*Basic Linguistic Theory BLT*, Dixon, 2009 ; 2010 ; 2012), un ensemble de concepts théoriques utilisés pour la description grammaticale des langues et en typologie linguistique. Il couvre les structures de prédicat-argument et les relations entre eux. Un texte est vu comme une collection de scènes dénotant une situation mentionnée dans la phrase, impliquant généralement un *prédicat*, des *participants* et des *modifieurs*.

UCCA représente la sémantique d'une phrase utilisant un graphe acyclique, où les nœuds non terminaux correspondent à des unités sémantiques représentent et les nœuds terminaux correspondent à des tokens de texte. Ainsi UCCA est toujours ancré dans le texte de surface et tous les mots de surface ont leurs nœuds. Les arêtes sont étiquetées, indiquant le rôle d'un « enfant » vis-à-vis de son « parent ». La description des éléments d'une scène est réalisée en utilisant 12 catégories, tels que : A (*Participant*), P (*Process*), E (*Elaborator*), etc.

Tout comme AMR, UCCA utilise des graphes acycliques dirigés qui sont destinés à s'éloigner des constructions syntaxiques spécifiques afin de représenter les relations sémantiques. UCCA n'est pas lié à une ressource lexicale particulière. Les auteurs mentionnent quelques ressemblances avec le projet FrameNet, où les cadres peuvent être vus comme une abstraction indépendante du contexte des scènes d'UCCA.

UCCA est une représentation sémantique constituée de deux couches. La première couche représente les structures et relations les plus importantes, comme par exemple les verbes, les adjectifs et les noms et structure les énoncés sous la forme de scènes, qui désignent des actions ou des états persistants. Le seconde couche affine ces représentations. La Figure 8 est un exemple d'une représentation UCCA. La relation principale de la scène pointe vers le verbe *kicked*. Il y a deux participants, *John* et *his ball*. Le concept central du second participant est *ball*, associée à *his*.

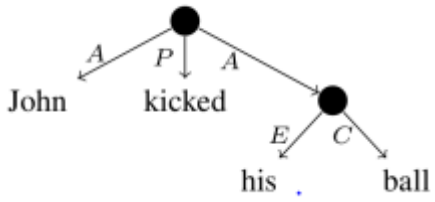


Figure 8 - Graphe UCCA de la phrase « *John kicked his ball* ».

UCCA a démontré son applicabilité en anglais, français et allemand (avec des projets pilotes d'annotation sur le tchèque, le russe et l'hébreu). Il a également été utilisé pour la simplification du texte (Sulem et al., 2018a), ainsi que pour l'évaluation d'un certain nombre de tâches de génération de textes et de traduction automatique (Birch et al.,

2016 ; Choshen et Abend, 2018), car les phrases ayant une même signification ont des graphes UCCA similaires même lorsqu'elles sont traduites dans d'autres langues.

Un point problématique clair est que UCCA ne distingue pas les différents rôles des arguments. Ainsi, dans l'exemple de la Figure 8, *John* et *his ball* sont tous deux des participants (A), bien que dans des théories sémantiques plus traditionnelles, ils seraient respectivement distingués en tant qu'agent et thème. Une telle distinction est utile dans notre travail. Žabokrtský et al., (2020) dressent également quelques points faibles d'UCCA. Ils trouvent que ce formalisme ne modélise pas explicitement la syntaxe ni ne s'appuie sur d'autres couches d'annotation, en supposant que l'annotation sémantique peut être « mappée » directement sur la forme de surface. UCCA est relativement insensible aux variations syntaxiques, donnant des analyses similaires à des scènes syntaxiquement différentes mais sémantiquement proches. Comparons par exemple les annotations de l'anglais *John took a shower* et *John showered* : dans les deux cas, on a une seule scène avec un participant et avec la relation principale dont l'élément central exprime l'action de *showering* : (1) John$_A$ [took$_F$ [a$_E$ shower $_C$]$_C$]$_P$ et (2) John $_A$ showered $_P$.

3.6.3. La famille *MRS dans DELPH-IN

DEep Linguistic Processing with HPSG INitiative (DELPH-IN)[26] est un consortium international de chercheurs dans le domaine de la linguistique informatique. Leur objectif est de développer des outils de traitement linguistique en profondeur du langage humain. Ce projet permet la mise en œuvre des grammaires HPSG disponibles pour une variété de langues, y compris l'anglais avec la grammaire ERG (Flickinger, 2000), une grammaire symbolique de l'anglais à large couverture, développée dans le cadre du DELPH-IN et projet LinGO (Copestake et Flickinger, 2000). Cette mise en œuvre implique l'ajout de représentations sémantiques, la plus basique étant MRS - *Minimal Recursion Semantics* (Copestake et al., 2005).

Les graphes sémantiques de DELPH-IN sont ainsi basés sur deux formalismes : **HPSG** (*Head-driven Phrase Structure Grammar* ; Pollard et Sag, 1988, 1994, Sag et Wasow, 1999) pour la partie syntaxique (section 3.2.1), et **MRS** (*Minimal Recursion Semantics* ; Copestake et al., 2005) pour la partie sémantique. Copestake et al. (2005) définissent comment MRS peut être représentée comme une structure de caractéristiques et utilisée dans ERG comme une représentation sémantique, aux côtés d'autres structures de caractéristiques représentant la syntaxe. Une représentation de structure de caractéristiques HPSG d'une phrase est créée à l'aide d'un analyseur utilisant une unification pour combiner des structures de caractéristiques lexicales et syntaxiques. La représentation MRS d'une phrase est donc construite compositionnellement en parallèle avec l'analyse syntaxique d'une phrase.

MRS et ses représentations associées sont désignées collectivement par *MRS. Dans la suite, nous utiliserons la nomenclature MRS pour faire référence à la représentation telle que définie par Copestake et al. (2005), *MRS pour faire référence à la famille des

[26] http://moin.delph-in.net/wiki/FrontPage

représentations basées sur MRS, et DELPH-IN pour faire référence au cadre sémantique général, c'est-à-dire aux deux modèles d'analyse linguistique : HPSG et MRS.

Goodman (2018) définit DELPH-IN comme une représentation symbolique, structurelle, compositionnelle, sous-spécifiée et abstraite du sens. *Symbolique* parce qu'elle est représentée par des symboles « atomiques » discrets, tels que des prédicats et les rôles plutôt que des valeurs numériques. Elle est *structurelle* parce qu'elle encode les entités impliquées dans un énoncé et leurs relations les unes avec les autres. Elle est *compositionnelle*, où la représentation du sens pour une phrase est construite à partir de la contribution du sens des mots de la phrase et du mode de composition syntaxique de ces mots (Bender et al., 2015). Cela est comparé à AMR qui est à la fois symbolique, structurelle et abstraite, mais non compositionnelle. DELPH-IN est *sous-spécifiée* car une seule instance de représentation englobe une ou plusieurs analyses logiques. Elle est *abstraite* parce qu'elle décrit, comme la syntaxe, un énoncé avec des étiquettes généralisées, non spécialisées pour leur utilisation spécifique à un domaine ; par exemple, requêtes SQL sur une base de données.

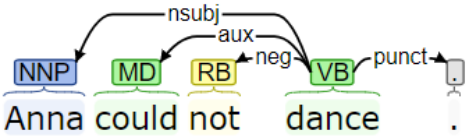
DELPH-IN n'est pas une théorie sémantique en soi, mais « un méta-langage pour décrire les structures sémantiques dans un langage objet sous-jacent »[27] (Copestake et al., 2005). Plusieurs autres représentations sémantiques sont dérivées de MRS, citons notamment *Robust Minimal Recursion Semantics* – RMRS (Copestake, 2004), *Elementary Dependency Structures* – EDS (Oepen et al., 2004 ; Oepen and Lønning, 2006), *DELPH-IN MRS Bilexical Dependencies* – DM (Ivanova et al., 2012) et *Dependency Minimal Recursion Semantics* – DMRS (Copestake, 2009).

Le format DMRS a été conçu pour fournir des représentations alternatives, plus compact, plus lisibles et plus faciles à comparer et à factoriser que les structures MRS et leurs équivalents pour l'analyse robuste, les structures RMRS (Copestake, 2007). Les représentations DMRS encodent la sémantique d'une phrase sous la forme d'un graphe acyclique dirigé au lieu d'une représentation sémantique plate en MRS (Bos, 1995 ; Gardent et Parmentier, 2007). Nous développons ce format dans la section 3.6.4.

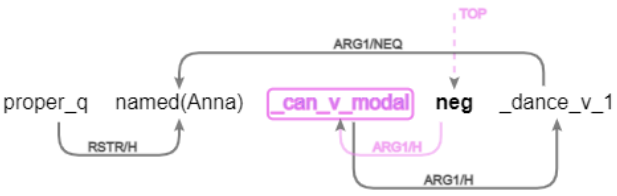
Nous aborderons dans la section 4.1.2 les différences entre les dépendances syntaxiques et sémantiques. Un facteur important que nous discuterons est la représentation des relations scopales entre les prédicats, par exemple, comment divers adverbes interagissent avec d'autres éléments de la phrase. Bien que deux éléments interdépendants dans une analyse syntaxique aient également tendance à interagir sémantiquement, en particulier dans la représentation compositionnelle telle que *MRS, la direction de la dépendance est souvent permutée. Par exemple, considérons les dépendances syntaxiques et sémantiques pour l'exemple 23, illustrées à la Figure 9.

(23) Anna could not dance.

[27] DELPH-IN Semantics is not a semantic theory on its own, but a "meta-level language for describing semantic structures in some underlying object language"



a. Dépendances de Stanford[28] de l'exemple 23



b. DMRS[29] de l'exemple 23

Figure 9 - Dépendances syntaxiques et sémantiques de l'exemple 23

Dans cet exemple, chaque arc syntaxique commence au verbe *dance*, qui a des dépendances de sujet, d'auxiliaire et de négation. Le même verbe dans la DMRS (Figure 9b) n'a qu'une seule dépendance sémantique, notée ARG1/NEQ, correspondant à un rôle d'agent thématique. La négation et le verbe modal modifient tous deux le verbe principal plutôt que d'être ses dépendants, ce qui reflète les relations scopales entre les prédicats. La distinction permet de différencier deux interprétations de la phrase :

(24) a. Anna was really bad at dancing.
b. Anna was capable of not dancing.

La DMRS de la Figure 9b représente la première résolution scopale de l'exemple 23. L'autre variante verrait l'ordre des liens entre la négation et le verbe modal inversé (cf. 4.1.2). Le nœud supérieur est la négation *neg*, tandis que les propriétés grammaticales de la phrase, telles que le temps, sont déterminées par le nœud d'index supérieur *_can_v_modal*. Intuitivement, cependant, le nœud sémantiquement le plus important est le verbe *_dance_v_1*. Ces nœuds clés sont appelés des *nœuds centraux*.

Cette thèse se concentre sur la simplification syntaxique des textes en utilisant la DMRS. Dans cette section, nous introduisons les éléments de base qui sont employés dans les représentations sémantiques DELPH-IN tout en consacrant une section pour décrire dans le détail le formalisme DMRS, le format choisi par les travaux de cette thèse (section 3.6.4).

3.6.3.1. Minimal Recursion Semantics (MRS)

Minimal Recursion Semantics (MRS, Copestake et al., 2005) est la représentation sémantique originale dans DELPH-IN et *l'ancêtre* des représentations du DELPH-IN. MRS a été principalement conçue pour être utilisée avec des grammaires développées dans un formalisme de structure de traits typés, mais ne lui est pas spécifique (Copestake et al., 2001). Les représentations MRS sont fondées sur une logique formelle, telle que le calcul des prédicats. Un exemple de MRS est illustré à la Figure 10, bien que nous ayons omis certaines propriétés de la figure pour plus de clarté.

[28] https://corenlp.run/
[29] http://delph-in.github.io/delphin-viz/demo/

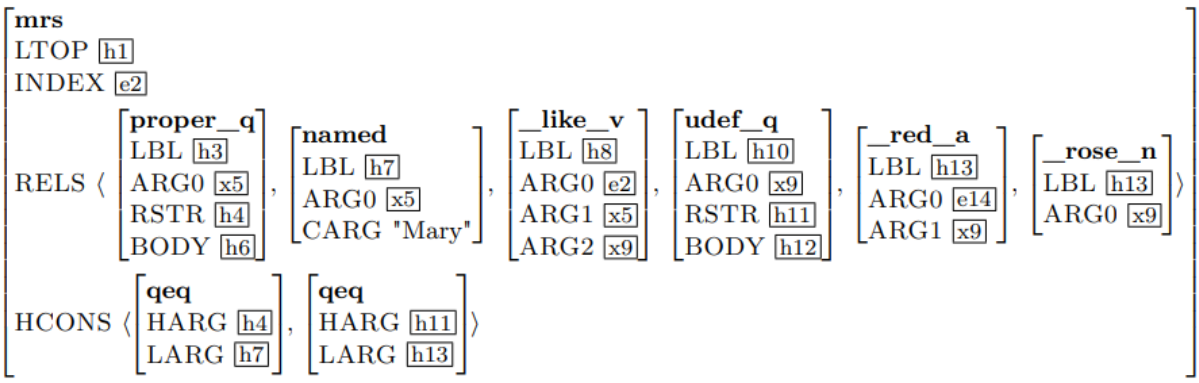


Figure 10 - Représentation MRS de la phrase « *Mary likes red roses* » avec certains attributs linguistiques omis. Les types sont en **gras**. Les caractéristiques (ou attributs) sont en MAJUSCULES. Les valeurs sont dans les □. Les valeurs dans < > sont des valeurs de liste.

La MRS a deux dérivés : le *Robust MRS* (RMRS, Copestake 2004) et le *Dependency MRS* (DMRS, Copestake 2008). La DMRS sera décrit dans une section à part comme c'est le formalisme choisi dans cette thèse (cf. 3.6.4). RMRS sert d'interface entre le traitement profond et superficiel. Il intègre des techniques de traitement superficiel telles que le *POS-tagging* et construit une structure sémantique compatible avec un traitement plus approfondi. Il est robuste dans le sens où les arguments du lexique produits par un traitement superficiel peuvent être omis ou sous-spécifiés. Dans notre travail, nous nous concentrons sur DMRS, mais comme pour l'instant il n'y a aucun moyen de produire des analyses DMRS directement sans conversion à partir de MRS, un certain niveau de compréhension du format MRS d'origine est requis.

Prédications et graphe sémantique

Les éléments les plus riches en MRS sont les prédicats, car ils encodent le lexique. Ils sont répartis en deux grandes catégories :

- ***Les prédicats de surface*** (appelés aussi ***prédicats réels***), correspondant directement aux lemmes de mots sont indiqués par un trait (_) au début, par exemple, *_rose_n_1* pour le mot *rose* ou *roses*, ou *_like_v_1* pour *like*, *likes*, *liked*. Ils correspondent à des lexèmes particuliers et ils ont un lemme, une étiquette de partie du discours et un sens qui leur est associé *: _lemma_pos_sense*, *_example_n_of*.
- ***Les prédicats abstraits*** (appelés les ***prédicats de grammaire*** *gpreds*), qui sont utilisés pour transmettre des informations supplémentaires sur les prédicats existants. Par exemple, *neg* introduit la négation, tandis que *named('Mary')* représente le nom propre *Andy*, *dofw* représente *day of the week*, etc.).

Généralement, les *prédicats de surface* sont définis dans le lexique de la grammaire et sont insérés dans la représentation sémantique lors de l'analyse lorsqu'une entrée lexicale est sélectionnée pour un token d'entrée. Les *prédicats abstraits* sont généralement insérés par des règles de grammaire utilisées dans une analyse, bien qu'ils

puissent également apparaître dans le lexique, comme *much-many_a*, qui est utilisé pour les entrées lexicales pour *much*, *many*, *a lot*, etc.

Une instance d'un prédicat et de ses arguments sortants est appelée une prédication élémentaire (*Elementary Predication* EP). Chaque EP a une étiquette de portée appelée *handle*, un prédicat et une liste d'arguments numérotés ARG1, ARG2, etc. La structure résultante de la composition des prédications d'une phrase est le **graphe sémantique** dans lequel :

- Chaque **EP** est affiché dans une ligne individuelle de la section **RELS**. Chaque EP a une variable caractéristique unique ARG0 et potentiellement une liste d'autres arguments numérotés ARG1, ARG2, etc.
- Local Top (**LTOP**) est le nœud le plus haut (supérieur) de MRS.
- Index (**INDEX**) est une caractéristique importante de HPSG (Pollard et Sag, 1994), qui, au niveau de la phrase, pointe vers la tête syntaxique de la phrase, qui est presque toujours le verbe principal de la phrase. Il commence généralement par un "e" indiquant que MRS représente un événement et que le prédicat principal *(_like_v_rel* dans ce cas) le porte par sa caractéristique arg0, c'est-à-dire sa variable liée.
- **RELS** est un ensemble de prédications élémentaires (EP), dans lequel un seul EP signifie une seule relation avec ses arguments, comme *_like_v_rel* (e2, x5, x9).
- **HCONS** est un ensemble de contraintes qui relient les étiquettes aux « trous » d'arguments dans les quantificateurs et autres prédicats scopaux.

Les **arguments**, ou arêtes, ont des étiquettes appelées **rôles**. Leurs valeurs peuvent être soit des variables **non-scopaux** (par exemple x3) ou **handle scopaux** (par exemple h11). Les variables non-scopaux sont soit des **instances** (marquée par x), pour annoter les indices référentiels, tels que les noms réguliers, les pronoms, les verbes nominalisés ; soit des **éventualités** (*eventuality/event* e), par exemple, les événements et les états, comme des verbes ou des adjectifs. Les quantificateurs n'ont pas de variable propre. En général, les nœuds d'instance ressemblent à des noms avec le genre, le nombre et la personne, tandis que les nœuds d'éventualité ont des propriétés de type verbe.

Les prédicats sémantiques n'encodent que les unités lexicales et certains types de distinctions de sens, et non les variations morphologiques (et les contributions sémantiques associées) présentes dans les tokens. Dans MRS, les propriétés morphosémantiques qui codent, par exemple, le temps du verbe ou le nombre nominal sont appelées propriétés des variables car elles sont encodées sur des variables liées à des prédications, alors que dans DMRS qui est une représentation sans variable, elles sont encodées sur les nœuds du graphe. L'exemple d'informations complètes sur un prédicat donné inclus dans l'analyse est présenté dans le Tableau 3 pour le prédicat *_red_a_*.

Information	Valeur
Prédicat	_red_a_1
Label (LBL)	h13
Variable intrinsèque (ARG0)	e14
Propriétés du nœud	SF: prop TENSE: untensed MOOD: indicative PROG: - PERF: -
Arguments	ARG1: x9

Tableau 3 - Les informations complètes sur le prédicat *_red_a_1*

Dans la Figure 10, les propriétés des nœuds sont supprimées pour plus de clarté. Dans la version originale, le prédicat *red_a_1* est représentée comme suit :

```
[ _red_a_1 LBL: h13 ARG0: e14 [ e SF: prop TENSE: untensed MOOD:
indicative PROG: - PERF: - ] ARG1: x9 ]
```

Lorsqu'un élément lexical ne peut être identifié avec aucun des prédicats de grammaire, l'ERG lui attribue un prédicat personnalisé, avec un lemme composé du token inconnu et de sa balise Penn Treebank POS, par ex. *_balkanized/VBN_u_unknown.*

3.6.3.2. Robust Minimal Recursion Semantics (RMRS)

Robust Minimal Recursion Semantics (RMRS) est une représentation MRS modifiée qui, en plus de la sous-spécification de la portée, permet une sous-spécification des informations relationnelles (Copestake, 2007). A savoir, dans RMRS, les arguments sont représentés comme des éléments distincts et peuvent être omis ou sous-spécifiés.

Le processus de transformation entre MRS et RMRS sépare la plupart des arguments de prédicat élémentaires et utilise des ancres (notées par a) pour relier les EP et les ARG. Par exemple, dans MRS, `l3:sit_v_1(e3,x3)` devient `l3:a3:sit_v_1(e3),l3:a3:ARG1(x31)` dans RMRS. La Figure 11 est la représentation RMRS de la phrase « *The fat cat sat on a mat* » (Copestake, 2007).

Représentation MRS :

$$l0{:}\ \_the\_q(x0, h01, h02), l1{:}\ \_fat\_j(x1), l2{:}\ \_cat\_n(x2), l3{:}\ \_sit\_v\_1(e3, x3), l4{:}\ \_on\_p(e4, e41, x4),$$
$$l5{:}\ \_a\_q(x5, h51, h52), l6{:}\ \_mat\_n\_1(x6),$$
$$h01 =_q l1, h51 =_q l6$$
$$x0 = x1 = x2 = x3, e3 = e41, x4 = x5 = x6, l1 = l2, l3 = l4$$

Représentation RMRS équivalente :

$$l0{:}\ a0{:}\ \_the\_q(x0), l0{:}\ a0{:}\ \mathrm{RSTR}(h01), l0{:}\ a0{:}\ \mathrm{BODY}(h02), l1{:}\ a1{:}\ \_fat\_j(x1), l2{:}\ a2{:}\ \_cat\_n(x2),$$
$$l3{:}\ a3{:}\ \_sit\_v\_1(e3), l3{:}\ a3{:}\ \mathrm{ARG1}(x31), l4{:}\ a4{:}\ \_on\_p(e4, e41, x4), l4{:}\ a4{:}\ \mathrm{ARG1}(e41), l4{:}\ a4{:}\ \mathrm{ARG2}(x4),$$
$$l5{:}\ a5{:}\ \_a\_q(x5), l5{:}\ a5{:}\ \mathrm{RSTR}(h51), l5{:}\ a5{:}\ \mathrm{BODY}(h52), l6{:}\ a6{:}\ \_mat\_n\_1(x6),$$
$$h01 =_q l1, h51 =_q l6$$
$$x0 = x1 = x2 = x3, e3 = e41, x4 = x5 = x6, l1 = l2, l3 = l4$$

Figure 11 - Représentations MRS et RMRS de la phrase « *The fat cat sat on a mat* » (Copestake, 2007).

Un effort pour combiner les points forts des techniques de traitement peu profondes, avec des grammaires profondes a conduit au développement de la sémantique de RMRS, qui permet une sous-spécification des symboles de prédicat, des rôles d'argument et des propriétés morphosémantiques. RMRS représente également les EP dans une notation de style Parsons (Parsons, 1990) où chaque argument de rôle est lié à son EP via une variable d'ancrage unique. Ces variables d'ancrage sont une avancée importante, car elles donnent à chaque EP, y compris les quantificateurs, un identifiant unique, ce que les variables intrinsèques n'ont pas pu accomplir car elles n'étaient pas à l'origine une partie obligatoire du formalisme.

3.6.3.3. Elementary Dependency Structures (EDS)

Oepen et Lønning (2006) ont proposé le formalisme EDS, une conversion de MRS en graphes de dépendance sans variable. La conversion en EDS entraîne des pertes, car elle supprime les informations de portée au profit de la simplicité de représentation. L'EDS n'est pas destiné à être une représentation entièrement expressive, mais à faciliter les tâches comme la recherche d'informations (Kouylekov et Oepen, 2014), la désambiguïsation de l'analyse (Toutanova et al., 2005). La Figure 12 est un exemple de représentation en EDS[30].

```
{e2:
_1:proper_q<0:6>[BV x5]
x5:named<0:6>("Abrams")[ ]
e2:_promise_v_1<7:15>[ARG1 x5, ARG2 x10, ARG3 e16]
_2:_the_q<16:19>[BV x10]
x10:_dog_n_1<20:23>[ ]
e16:_bark_v_1<27:32>[ARG1 x5]}
```

Figure 12 - Représentation EDS de la phrase « *Abrams promised the dog to bark* ».

[30] http://moin.delph-in.net/wiki/EdsTop

3.6.3.4. Bilexical Dependencies (DM)

DELPH-IN MRS Bilexical Dependencies (DM ; Ivanova et al., 2012) réduit encore l'EDS pour projeter le graphe de dépendance directement sur les tokens de surface. La conversion en DM utilise les fonctionnalités d'EDS, telles que la sélection de nœuds représentatifs, mais accorde également des représentations pour les tokens dans la phrase d'origine qui ne sont pas incluses dans d'autres variétés *MRS. DM est un formalisme à la frontière entre la sémantique et la syntaxe. DM possède une annotation des arguments principaux de la phrase similaire à PropBank et NomBank, mais elle est beaucoup « plus dense, car son processus d'annotation conserve des informations syntaxiques telles que les coordinations » (Ribeyre, 2016).

3.6.3.5. Conversion de MRS vers ses dérivés

Plusieurs autres représentations sémantiques sont dérivées de MRS. Les représentations *MRS sont pertinentes pour des applications qui nécessitent à la fois une sémantique structurée et la capacité de produire du texte en langage naturel en sortie. Chacune de ces représentations est conçue pour des utilisations différentes : ainsi MRS pour la traduction automatique, EDS pour l'extraction d'informations, DM pour augmenter l'accessibilité des ressources DELPH-IN à une communauté plus large, et DMRS pour étendre les représentations de dépendances avec des informations de portée. Elles dérivent toutes de la représentation MRS originale, c'est pourquoi elles partagent un grand nombre de propriétés.

La conversion entre MRS et ses autres dérivés se déroule en deux temps de la façon suivante :

- Dans un premier temps, MRS est réduite en EDS qui ne garde que les rôles sémantiques (Arg_0, . . ., Arg_n, etc.) ou les propriétés morphologiques telles que le temps, le nombre, l'aspect, etc. Les structures de dépendances élémentaires EDS sont des graphes sémantiques qui maintiennent la structure prédicat-argument mais **ignorent certaines informations de portée**. Copestake (2009) a donc créé DMRS en tant que représentation de dépendance qui conserve les informations de portée et évite la perte d'informations. Cette expressivité est obtenue en ajoutant des annotations pour indiquer les informations de portée. Chaque prédicat de la forme logique correspond à un nœud dans le graphe de dépendances et les relations entre les arguments et les prédicats se transforment en arcs de dépendances.
- Dans un second temps, l'EDS est transformé en DM, liant prédicat et arguments entre eux. DM exprime la structure prédicat-argument uniquement en termes de tokens de surface, sans introduire de prédicat abstrait, d'où le schéma MRS→DMRS→ EDS→DM.

Dans cette thèse, nous avons fait le choix d'exploiter les structures DMRS. DMRS permet de bien faire ressortir les concepts porteurs de sens et leurs relations. Des phénomènes linguistiques essentiels, tels que les modalités ou les connecteurs logiques, sont pris en compte. Ce format a été conçu pour fournir des représentations alternatives,

plus lisibles et plus faciles à utiliser que les structures MRS et leurs équivalents. Les dépendances syntaxiques et les relations grammaticales coïncident largement avec DMRS. DMRS offre également l'avantage d'être bien outillé. Aussi la section suivante traite en détail ce formalisme, la structure des graphes, ainsi que ses principales caractéristiques.

3.6.4. Dependency Minimal Recursion Semantics (DMRS)

Copestake (2009) a introduit DMRS dans le but d'obtenir un format plus compact de RMRS en supprimant les redondances. Les représentations DMRS et MRS sont inter-convertibles entre elles sans aucune perte d'information. DMRS a été conçu pour inclure toutes les informations sémantiquement liées qui peuvent être dérivées de la syntaxe et de la morphologie.

Chronologiquement, DMRS a été développé après EDS et peut être considérée comme une extension d'EDS. En termes d'informations codées, cependant, il s'agit d'un sur-ensemble d'EDS et d'un sous-ensemble de MRS, donc on pourrait également dire que l'EDS est une forme réduite de DMRS. Outre le codage des informations de portée, DMRS contient presque les mêmes informations qu'EDS. Ce sont deux représentations sans variables, avec des propriétés morphosémantiques codées sur des nœuds plutôt que des variables.

La Figure 13 est la représentation DMRS de la phrase « *Mary likes red roses* ». Ce graphe est obtenu en utilisant Delphin-viz[31], un outil en ligne pour la visualisation de graphe en MRS et DMRS.

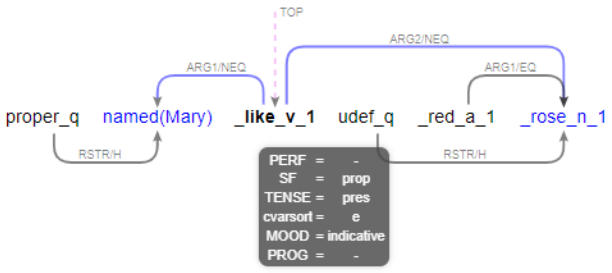


Figure 13 - DMRS de la phrase « *Mary likes red roses* ».

3.6.4.1. Étiquetage des nœuds et des arêtes

Dans une structure DMRS, chaque prédicat a une variable caractéristique distincte et toutes les informations lexicales et grammaticales concernant une prédication élémentaire (**EP**) donnée sont contenues dans un nœud avec un identifiant unique (*node_id*). Chaque nœud d'un graphe DMRS est associé à de nombreuses informations. Par exemple, un nœud représentant un prédicat de nom peut inclure des informations sur le lemme du prédicat, la partie du discours, le sens du prédicat, la personne et le nombre.

[31] http://delph-in.github.io/delphin-viz/demo/

Chaque nœud est connecté à ses arguments via des arêtes, et différents types d'arguments ont des étiquettes d'arêtes différentes. Chaque arête du graphe est étiquetée par une information, qui appartient à l'un des types suivants (Morey, 2011) :

- Les relations prédicat-argument, notées ARGi.
- Les liens dans DMRS combinent des arguments et gèrent des contraintes. Chaque lien a un nœud de début, un nœud de fin et une étiquette sous la forme A/B, composée d'un nom d'argument (A), hérité du nom de la fonctionnalité du MRS d'origine et de l'un des quatre types de lien suivants (B) indique le type d'une relation scopale :
  - **EQ** quand deux prédicats sont directement conjoints et font partie du même nœud ;
  - **H** quand un prédicat est dans la portée d'un autre prédicat, ce dernier étant flottant ; la variable caractéristique du premier prédicat n'est pas obligatoirement un argument direct du deuxième car des quantificateurs peuvent s'intercaler entre ces deux prédicats ;
  - **NEQ** quand la variable caractéristique d'un prédicat est un argument d'un autre prédicat, mais que ces deux prédicats ne sont pas directement liés, de façon structurelle, dans la représentation MRS équivalente. La différence entre EQ et NEQ est que si deux prédications élémentaires partagent la même étiquette, alors elles sont dans une relation EQ, sinon une relation NEQ.

La composition générale des étiquettes pour les nœuds de prédicats réels est : *_lemma_pos_sense***.** La composition générale des étiquettes pour un nœud de prédicat de grammaire est : *carg_gpred*. Exemples d'étiquettes de prédicat de grammaire : *compound* (joignant deux parties d'un composé), *neg* (désignant une négation) et *7_card* (numéro cardinal "7").

Ci-dessous, nous montrons des étiquettes plus granulaires pour certaines classes de nœuds de prédicats réels et de grammaire.

- **`_lemma_pos_sense_pers_num`** pour les prédicats de nom (pos='n'). La valeur par défaut de pers et num si elle n'est pas spécifiée est '3' et 'sg' respectivement. Exemples : `_ chemistry _n_1_3_sg` ('chemistry'), `_instrument_n_of_3_pl` ('instruments').
- **`_lemma_pos_sense_tense_sf_perf_prog`** pour les prédicats verbaux (pos='v'). La valeur par défaut pour sf est 'prop'. Exemples : `_fill_v_1_past` ('fill').
- **`carg_gpred_num`** pour les prédicats de grammaire de type nom. Exemples : `USA_named_n_sg` ('USA'), `person_sg` ('everyone'), `winter_season_sg` ('winter').
- **`gpred_pers_num_gend`** pour les prédicats de grammaire des pronoms. Exemples : `pron_3_sg_n` ('it'), `pron_2` ('you').

Le Tableau 4 représente une description des propriétés morpho-syntaxiques des nœuds.

PROPRIÉTÉ	DESCRIPTION
Prédicat	
Lemma	Lemme du prédicat, par exemple « *play* » pour le verbe « *playing* ».
Pos	Partie du discours, par exemple verbe, nom, adjectif/adverbe, préposition.
Sense	Sens du prédicat, par exemple "for" pour le verbe "standing (for something)".
Gpred	Indique un prédicat de grammaire (G*rammar predicate*).
Carg	Argument constant (*constant argument*), stocke les symboles de prédicat des nombres, des entités nommées, etc., qui ne sont pas stockés dans le lexique. Par exemple, "56" et "Mark".
Caractéristiques nominales	
Gend	Genre, 'm', 'f', et 'n'.
Ind	Individué (booléen).
Num	Nombre, notamment 'sg' et 'pl'.
Pers	Personne.
Temps, aspect et mode	
Tense	Temps du verbe, par exemple présent, passé, futur.
Perf	Aspect perfectif *vs* imperfectif (booléen)
Prog	Aspect progressif *vs* non progressif (booléen)
Mood	Mode et modalité, correspondant aux opinions/attitudes du locuteur, par exemple indicatif et subjonctif
Caractéristiques des phrases	
Sf	Par exemple proposition, question, et impérative.

Tableau 4 - Descriptions des propriétés morpho-syntaxiques des nœuds

3.6.4.2. Sous-graphes sémantiques

La compositionnalité de DMRS signifie que la représentation d'une phrase complète est composée de sous-graphes représentant ses constituants, comme l'indique la Figure 14. Le terme sous-graphe désigne généralement tout graphe formé à partir d'un sous-ensemble de sommets et d'arêtes d'un autre graphe. En se référant à DMRS, il est utile de restreindre la définition et de distinguer les sous-graphes sémantiques en tant que sous-graphes d'un DMRS qui sont eux-mêmes des fragments ou des constituants bien formés de la phrase, comme les propositions. Les sous-graphes sémantiques de l'exemple de la Figure 14 sont marqués par des ellipses colorées. Un sous-graphe sémantique consiste typiquement en un nœud central avec ses arguments et ses modificateurs. Chaque sous-graphe a son propre nœud supérieur (LTOP), qui est la cible des liens des opérateurs scopaux plus haut dans la hiérarchie. En fait, nous pouvons penser que les opérateurs scopaux agissent sur des sous-graphes entiers, plutôt que sur des nœuds individuels. Les sous-graphes centrés autour de nœuds avec des variables d'événement sont des situations, où chaque sous-graphe représente un événement (situation) d'où une proposition, ce qui confirme nos hypothèses concernant la forme des phrases simplifiées où chaque phrase contient un événement dans DMRS. Dans les cas où il s'agit d'une phrase qui contient deux ou plusieurs constructions syntaxiques à traiter (coordination et

apposition par exemple), l'ordre d'application des règles ne modifie pas les sous graphes obtenus. Nous détaillons ce point dans la section 5.4.

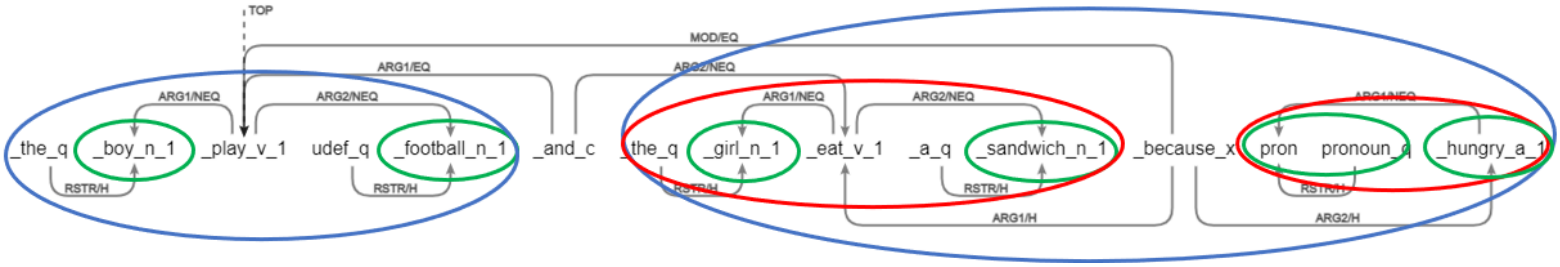


Figure 14 - DMRS de la phrase *"The boy is playing football and the girl is eating a sandwich because she is hungry".* Les ellipses colorées marquent les sous-graphes sémantiques.

3.7. Comparaison entre les formalismes

Les caractéristiques de ces structures sont variées. L'une des principales différences entre ces représentations sémantiques précédemment présentées est leur *niveau d'abstraction* de la phrase. Leurs relations à la syntaxe et leurs liaisons aux mots des phrases s'exercent également à des degrés différents. Nous appelons cette relation *l'ancrage* des nœuds sur des sous-chaînes (*anchoring*). D'autres termes couramment utilisés incluent *l'alignement*, la *correspondance* ou la *lexicalisation.* Ainsi, « l'ancrage aux mots d'une phrase donnée est fort pour les structures en dépendances, tandis qu'il est faible, voire inexistant, pour les structures basées sur la logique. Il en résulte des différences importantes dans leurs capacités à s'abstraire des variations conceptuelles et syntaxiques, et à intégrer des phénomènes linguistiques variés » (Lamercerie, 2021).

En raison des différences dans les principes de conception sous-jacents aux différentes banques de graphes, ces graphes diffèrent considérablement, souvent dans les stratégies fondamentales telles que la façon dont les éléments lexicaux sont ancrés dans les nœuds du graphe (Oepen et al., 2019). DM est un graphe de dépendance bi-lexical, où les nœuds sont des unités lexicales de surface. EDS et UCCA sont des formes générales de graphes sémantiques ancrés, dans lesquels les nœuds sont ancrés à des *spans* arbitraires de la phrase, plusieurs nœuds distincts peuvent avoir des chevauchements ; dans EDS, par exemple, la sémantique d'un adverbe peut être décomposée en quatre nœuds, tous ancrés à la même sous-chaîne. AMR est un graphe dont les nœuds des graphes ne sont pas ancrés, les graphiques AMR connectent des variables, des concepts et des constantes.

Nous suivrons Kuhlmann et Oepen (2016) dans la classification des banques de graphes en fonction de la relation qu'ils supposent entre les tokens de la phrase et les nœuds du graphe, appelé ***ancrage*** des fragments de graphe sur les sous-chaînes d'entrée. Dans la Conférence CoNLL 2019, sont distingués différentes ***flavors*** de graphes sémantiques. Ce terme représente la correspondance et décrit la nature de la relation que les graphes assument entre les mots de la phrase et les nœuds du graphe. Nous

distinguerons trois *flavors* de graphes sémantiques par degré d'ancrage, nous les appellerons de type (0) au type (2) :

- ***Flavor-0 :*** La forme d'ancrage la plus forte est obtenue dans les structures de dépendance bi-lexicale, où les nœuds de graphe correspondent de manière injective aux unités lexicales de surface. Dans de tels graphiques, chaque nœud est directement lié à un token spécifique, et les nœuds héritent de l'ordre linéaire de leurs tokens correspondants. **DM et PSD** représentent ce flavor.

- ***Flavor-1*** comprend une forme plus générale de graphes sémantiques ancrés, caractérisée par une correspondance moins forte que le *flavor* précédent entre les nœuds et les tokens. Les terminaux sont des éléments porteurs de sens, tandis que les arêtes permettent de caractériser les liens entre les unités, chaque jeton (ou un groupe de jetons dans le cas d'entités nommées, telles que des noms propres et des dates) correspond à un nœud dans le graphe. **EDS, DMRS et UCCA** représentent ce flavor. Ces graphiques offrent une plus grande flexibilité dans la représentation du sens apporté, par exemple, par des affixes ou des constructions phrastiques et facilitent la décomposition lexicale (les comparatifs). Les concepts sémantiques sont également ancrés aux phrases.

- ***Flavor-2.*** Certains graphes sémantiques ne considèrent pas la correspondance entre les nœuds et la chaîne de surface comme faisant partie de la représentation du sens. De tels graphes sémantiques ne sont tout simplement pas ancrés. Il n'y a pas de correspondance directe entre les jetons de niveau de surface et les nœuds de graphe : tous les jetons ne sont pas présents en tant que nœuds dans le graphe et tous les nœuds du graphe ne correspondent pas à des jetons. **AMR** représente ce *flavor*, qui est conçue pour abstraire la représentation de sens du token de surface. Les nœuds dans les graphiques AMR sont principalement alignés avec les unités lexicales de surface, bien que l'ancrage explicite soit absent de la représentation AMR. Ainsi, des phrases qui sont différentes en surface, mais qui ont le même sens de base sont représentées par le même AMR. D'où les trois phrases "*The girl made adjustments to the machine*", *"The girl adjusted the machine*" et "*The machine was adjusted by the girl*" possèdent la même representation AMR.

Plus le *flavor* associé est élevé, plus *l'ancrage* est faible, et la représentation sémantique associée est *abstraite*. Pour les cinq formalismes étudiés, on distinguera deux grands types d'ancrage :

- ***Ancrage lexical*** **:** selon les structures de dépendance bi-lexicales de DM et PSD, et l'ancrage de token lexical implicite sur AMR, les nœuds de graphes de **DM, PSD et AMR** sont ancrés aux unités lexicales de surface de manière explicite ou implicite. Les concepts sémantiques dans ces représentations de sens sont ancrés aux unités lexicales individuelles de la phrase. Surtout, ces unités

lexicales ne se chevauchent pas, et la plupart d'entre elles ne sont que des *tokens* uniques, des expressions à plusieurs mots ou des entités nommées. En d'autres termes, lors de l'analyse d'une phrase en graphiques DM, PSD, AMR, les tokens de la phrase d'origine peuvent être fusionnés en recherchant un lexique lors du prétraitement, puis peuvent être considérés comme un seul token pour l'alignement ou l'analyse. Ce type d'ancrage prend en charge l'ancrage explicite et implicite des concepts sémantiques aux *tokens*.

- ***Ancrage phrastique*** (*phrasal anchoring*) : différent de l'ancrage lexical sans chevauchement, les nœuds dans **EDS, UCCA** et **DMRS** ne correspondent pas toujours à des jetons de surface individuels, mais sont ancrés sur des portées plus grandes, chevauchant des ancrages d'autres nœuds, permettant une décomposition lexicale (par exemple des causatifs ou des comparatifs). Les nœuds dans UCCA n'ont pas d'étiquettes de nœud ou de propriétés de nœud, mais tous les nœuds sont ancrés à phrase sous-jacente. En outre, les nœuds dans UCCA sont liés dans une structure hiérarchique, avec des arêtes allant entre les nœuds parents et enfants.

Ces représentations sémantiques sont associées à des *flavors*, c'est-à-dire à des ancrages différents. Leurs propriétés sont résumées dans le Tableau 5, inspiré d'Oepen et al. (2019). DM représente des dépendances bi-lexicales, où les nœuds de graphe correspondent à des unités lexicales de surface. En revanche, EDS et AMR prennent la forme de réseaux sémantiques (ou graphes conceptuels), où les nœuds représentent des concepts et il n'est pas nécessaire d'avoir une correspondance explicite avec les formes linguistiques de surface. DM et PSD sont un graphe de dépendance bi-lexical, où les nœuds sont des unités lexicales de surface (tokens). EDS et UCCA sont des formes générales de graphes sémantiques ancrés, dans lesquels les nœuds sont ancrés à des *spans* arbitraires de la phrase, plusieurs nœuds distincts peuvent avoir des chevauchements ; dans EDS, par exemple, la sémantique d'un adverbe peut être décomposée en quatre nœuds, tous ancrés à la même sous-chaîne. AMR est un graphe dont les nœuds des graphes ne sont pas ancrés, les graphiques AMR connectent des variables, des concepts et des constantes.

	DM	EDS	UCCA	AMR	MRS	DMRS
Flavor	***0***	***1***	***1***	***2***	***Non-graphique***	***1***
Ancrage lexical	✓	X	X	✓		X
Ancrage phrastique	X	✓	✓	X		✓
Nœud supérieur (top)	✓	✓	✓	✓	✓	✓
Étiquettes de nœud	✓	✓	X	✓	✓	✓
Propriétés de nœud	✓	✓	X	✓	✓	X
Ancrage de nœud	✓	✓	✓	X	X	✓
Arêtes dirigés	✓	✓	✓	✓	X	✓
Arêtes étiquetés	✓	✓	✓	✓	X	✓
Attribut des arêtes	X	X	✓	X	X	✓
Propriétés morphosémantiques	X	✓	✓	X	✓	✓
Portée du quantificateur	X	X	✓	X	✓	✓
Temps/aspect	X	X	✓	X	✓	✓

Tableau 5 - Comparaison des propriétés des formalismes

Conclusion

La syntaxe décrit la structure interne des langues. Selon la profondeur du traitement, l'information syntaxique peut se représenter de différentes façons (Gala, 2003). Il existe deux modèles formels généralisés pour appréhender la réalité linguistique d'un énoncé : 1) les syntagmes qui repose sur le principe de constituants immédiats entre les éléments de la phrase (grammaire de constituants) qui se concentre sur les analyses catégorielles et la hiérarchisation des constituants (Bloomfield, 1933 ; Chomsky, 1957) ; 2) les dépendances qui met en évidence les relations établies entre les mots telles que la coordination ou la subordination et comment ces relations sont produites (grammaire de dépendances) (Tesnières, 1959 ; Mel'čuk, 1988 ; Kahane, 2001). Suivant la première approche, Abney (1991) propose une décomposition plus fine des structures la phrase où le découpage se fait en « syntagmes noyau » (*chunks*). On distingue aussi deux niveaux : le niveau profond et le niveau superficiel.

Kahane et Gerdes (2022) distinguent les différences de définition entre les linguistes sur l'usage de la structure profonde et de la structure de surface. Dans le cadre de la Grammaire générative transformationnelle (Chomsky, 1965) : la structure profonde n'est pas réellement un niveau de représentation différent de la structure de surface, mais une structure syntaxique *sous-jacente* à la structure de surface, la structure de surface étant obtenue par l'application de transformations sur la structure profonde. De plus, chez Mel'čuk, la structure syntaxique profonde est un arbre de dépendance non ordonné, l'ordre linéaire n'étant introduit qu'au moment de l'interface entre la syntaxe de surface et le texte, tandis que chez Chomsky, la structure profonde et la structure de surface sont des structures de constituants ordonnées.

En simplification syntaxique, les approches basées sur les arbres syntaxiques introduisent également des limitations spécifiques à la syntaxe, notamment les

dépendances linguistiques, les transformations ad hoc et les formalismes syntaxiques complexes (Alva-Manchego et al., 2020). Les approches basées sur la sémantique se prêtent dans une certaine mesure à ces problèmes. Ils ont un pouvoir plus expressif en utilisant des ressources sémantiques lexicales et en utilisant des générateurs plus flexibles. Plus précisément, les méthodes basées sur la sémantique impliquent à la fois l'analyse syntaxique et la génération à partir de la sémantique, ainsi que la transformation via la sémantique.

Dans ce chapitre, nous avons donné un aperçu de la tâche d'analyse sémantique et de certains formalismes. Les représentations sémantiques diffèrent sous divers aspects, notamment le formalisme de représentation (logique des prédicats, graphes, etc.), l'inventaire des concepts et des relations, et la nature des connaissances externes disponibles sur les concepts et les relations. Nous avons mis l'accent sur les représentations sémantiques à base de graphes, notamment le cadre sémantique DELPH-IN. Nous avons décrit la représentation canonique MRS et les représentations dérivées RMRS, EDS, DM et DMRS.

Dans DELPH-IN, le cadre complet se compose d'un composant syntaxique et d'un composant sémantique, à savoir une dérivation HPSG et une structure MRS ; chaque structure de caractéristique HPSG détermine un MRS. Les représentations basées sur DELPH-IN peuvent un bon choix pour les applications qui nécessitent à la fois une sémantique structurée et la capacité de produire du texte en langage naturel en sortie. Les analyses *MRS sont fortement guidées par la syntaxe.

Dans le cadre de ce travail de thèse, nous avons retenu la représentation sémantique DMRS, associée à un ancrage phrastique. La plupart des représentations sémantiques reposent sur des analyses syntaxiques, de sorte qu'il existe un fort chevauchement entre les constituants sémantiques et syntaxiques. Cela s'applique particulièrement pour DMRS, qui est fortement compositionnel par conception. Sa sémantique est ancrée dans la forme superficielle des phrases et dans les liens syntaxiques entre les constituants. Bien qu'importantes, les relations syntaxiques elles-mêmes ne suffisent pas à déterminer pleinement les événements. Des informations supplémentaires s'avèrent utiles, par exemple, lors de l'examen de la portée et de la nature des modificateurs adverbiaux. DMRS a été développée pour éviter la perte d'informations, principalement grâce à des étiquettes d'arêtes plus riches. De ce fait, DMRS nous parait être adapté pour la simplification de textes, car DMRS est explicitement compositionnel, capturant la contribution à la signification de la phrase de toutes les parties de la forme de surface sans perte d'informations.

Dans le chapitre suivant, nous décrivons les fondements linguistiques de notre méthode de simplification syntaxique des textes, qui exploite des représentations sémantiques des phrases exprimées en DMRS.

Partie II

4. Simplification syntaxique basée sur la sémantique : fondements linguistiques

D'après Huddleston (1984) les phrases sont généralement classées en trois types : *simples*, *composées* et *complexes*, selon les types de propositions qu'elles contiennent. Ainsi, les phrases simples ne contiennent pas de propositions dépendantes (exemple 25a) ; les phrases composées contiennent au moins deux propositions indépendantes (exemple 25b) ; les phrases complexes sont composées d'au moins une proposition dépendante et une proposition indépendante (exemple 25c).

(25) a. He wrote his novel last year.
b. Tom reads novels and Ryan reads comics.
c. Tom, who read novels, rarely read comics.

Dans ce travail de thèse, nous nous intéressons spécifiquement aux transformations syntaxiques qui modifient la macrostructure d'une phrase complexe. Le développement des règles de réécriture permettant les transformations (phrase complexe vers phrase plus simple) dépend de la structure des phrases. Notre objectif est d'étudier chacune des constructions syntaxiques à traiter, c'est-à-dire les différentes formes qu'une construction peut avoir, afin de dresser les stratégies et les traduire en règles de réécriture. Chaque construction syntaxique est définie par une (ou un ensemble de) règle(s) correspondante(s).

L'approche que nous avons retenue dans cette recherche est une approche à base de règles qui n'utilise pas l'apprentissage automatique sur de gros corpus parallèles. Notre méthode à base de stratégies et de règles de transformation explicites, permet d'interpréter, d'expliquer comment les transformations des phrases à simplifier sont réalisées, permettant ainsi d'analyser les erreurs et leur origine dans le but d'améliorer la méthode et son implémentation dans le système par modification de ces règles.

Dans ce chapitre, nous décrivons la méthode de simplification syntaxique des textes en anglais que nous avons choisie et implémentée. Cette méthode est basée sur des transformations de représentations sémantiques des phrases exprimées en notation *Dependency Minimal Recursion Semantics* (DMRS). Ces transformations sont assurées par des règles définies manuellement. Le chapitre présente les grandes étapes de la méthode et les *stratégies* qui permettent de définir les constructions syntaxiques complexes afin de les traiter par des règles spécifiques qui sont présentées. Une mise en œuvre informatique de cette méthode de simplification au travers du système GRASS est présentée dans le chapitre 5, et son évaluation sur des corpus de référence est traitée dans le chapitre 6.

Ce chapitre est organisé en deux sections : la section 4.1 décrit l'apport des informations sémantiques dans la tâche de simplification syntaxique par rapport aux approches existantes, ainsi que les grandes étapes de la méthode et les différents composants de notre système. Les sections de 4.2 à 4.7 développent notre méthode à base de règles pour chacune des constructions traitées, telles que les le découpage des coordinations, des appositions, des subordinations, des relatives et la transformation de la voix passive en voix active.

4.1. Pourquoi une approche à base de la sémantique ?

4.1.1. Constat

Les approches statistiques et neuronales utilisées récemment dans la simplification de textes sont confrontées à plusieurs problèmes. Nous décrivons d'abord quelles sont les points faibles de ces approches basées sur des données (des textes ou des arbres syntaxiques) ; ensuite, nous décrivons comment l'utilisation des informations sémantiques peut s'avérer importante en SAT.

En simplification syntaxique, décider *quand* et *où* diviser une phrase en deux phrases plus simples, associées à des évènements distincts, peut être déterminé par la syntaxe même, par exemple, des phrases avec des subordonnées. Par exemple, les grammaires synchrones (Shieber et Schabes, 1991) ont été utilisées dans des systèmes à base de règles pour générer et classer les réécritures possibles d'une phrase d'entrée (Dras, 1999 ; Woodsend et Lapata, 2011). Cependant, les approches de la simplification syntaxique basées uniquement sur la syntaxe (Zhu et al., 2010 ; Woodsend et Lapata, 2011) n'arrivent pas à reconstruire correctement l'élément partagé ni à identifier correctement le point de découpage. Dans l'exemple (26) ci-dessous, le premier système S1 (Zhu et al., 2010) n'arrive pas à identifier correctement le point de découpage (*and*). L'autre système S2 (Woodsend et Lapata, 2011) n'arrive pas à réécrire correctement l'élément partagé « *the judge* » (dont le pronom personnel correct est « *he* »).

(26) The judge ordered that Chapman should receive psychiatric treatment in prison and sentenced him to twenty years to life.
S1. The judge ordered that Chapman should get psychiatric treatment. *In prison and sentenced him* to twenty years to life. *(Zhu et al., 2010)*
S2. The judge ordered that Chapman should receive psychiatric treatment in prison. *It* sentenced him to twenty years to life. *(Woodsend et Lapata, 2011)*

Candito (2022) affirme que les arbres de dépendances permettent une lecture plus immédiate de la structure argument-prédicat. Mais elle a constaté qu'en pratique, à partir d'un arbre de dépendances, il reste encore beaucoup à expliciter pour obtenir ces structures argumentales. Elle trouve que certains phénomènes entièrement déterminés

par la syntaxe ne sont pas directement encodés dans les représentations "surfaciques" couramment utilisées en TAL (arbres de constituants ou arbres de dépendances). C'est le cas par exemple la coordination de VP (coordination entre deux événements) (cf. section 4.4.2), par exemple dans *Anna terminera son mémoire le 10 et l'enverra au correcteur dans la foulée*, l'information que c'est *Anna* qui joue le rôle du sujet de *enverra* n'est pas immédiatement accessible dans un arbre de dépendances. Dans notre thèse, nous traitons ce type de construction en anglais en divisant la coordination en deux phrases. Ce constat va dans le sens de notre propos : l'exploitation des arbres syntaxiques nous semble insuffisante dans une tache de simplification de textes comme le sujet partagé n'est pas pris en considération dans ces types de représentations.

L'usage de la sémantique dans la simplification syntaxique peut s'avérer nécessaire, notamment dans la *construction de la deuxième phrase* et dans *l'identification de l'élément partagé* avec la première partie à ajouter, comme l'illustre l'exemple précédent. Les représentations sémantiques des phrases donnent une idée claire des événements, de leurs ensembles de rôles associés et des éléments partagés, facilitant ainsi à la fois l'identification des points de découpage possibles et la reconstruction des éléments partagés dans les phrases résultant du découpage. Les opérations de division et de suppression sont conduites par la sémantique : la division est déterminée par les rôles sémantiques qui sont associés à un élément alors que la suppression d'un nœud est déterminée par ses relations sémantiques avec les événements divisés, d'où un modèle de suppression qui doit distinguer entre les arguments et les modifieurs. Dans ce travail de thèse, nous ne traitons pas la question de suppression d'information, nous nous concentrons sur la division de phrases.

Dans de nombreux cas, la division d'une phrase se produit lorsqu'une entité prend part à deux (ou plus) événements distincts décrits dans la phrase. Dans l'exemple (27) ci-dessous, l'élément partagé « *bricks* » est impliqué dans deux événements : "*being resistant to cold"* et "*enabling the construction of permanent buildings*." (Narayan et Gardent, 2014).

(27) a. Being more resistant to cold, bricks enabled the construction of permanent buildings.

b. **Bricks** were more resistant to cold. **Bricks** enabled the construction of permanent buildings.

4.1.2. De l'usage de DMRS pour la simplification syntaxique

Bien qu'importantes pour décrire la structure de la phrase, les relations syntaxiques ne suffisent pas à déterminer correctement les divisions. Dans tous les cas, des informations supplémentaires s'avèrent utiles, par exemple, l'examen de la portée et de la nature des modifieurs adverbiaux. Les représentations syntaxiques ne font pas la distinction entre modifieurs adverbiaux modifiant un verbe et ceux modifiant la proposition entière. L'accent mis sur les constituants sémantiques plutôt que syntaxiques

rend le découpage des phrases plus pertinent pour les tâches qui se soucient de la compréhension du texte, par exemple la traduction et le résumé automatiques (Muszyńska, 2016).

Cependant, l'utilisation des informations sémantiques uniquement n'est pas suffisante non plus pour capturer les constructions syntaxiques complexes. D'où la nécessité de combiner les avantages des deux annotations syntaxique et sémantique. Pour ce faire, nous utiliserons un formalisme qui prend en considération et qui décrit à la fois les informations sémantiques et syntaxiques. Au contraire des méthodes basées sur la syntaxe (Zhu et al., 2010 ; Woodsend et Lapata, 2011), notre modèle fournit une sortie simplifiée qui préserve à la fois la syntaxe et le sens.

Notre objectif est le développement d'une méthode de simplification automatique de phrases qui effectue des opérations de transformation syntaxique **en utilisant des représentations sémantiques**. Notre méthode prend en entrée une représentation sémantique profonde à savoir, la structure DMRS attribuée par Copestake (2009) à la phrase complexe d'entrée. Ce formalisme prend en compte à la fois **les informations sémantiques et syntaxiques**. Cela permet un compte rendu linguistique de l'opération de division en ce que les constructions complexes sont exprimées dans les représentations et les éléments partagés sémantiquement sont pris en compte afin de les réécrire dans la phrase découpée. Ceci entraine une sortie simple qui est à la fois grammaticale (informations syntaxiques du DMRS) et préservant le sens (informations liées à la sémantique dans DMRS).

Plus précisément, nous proposons tout d'abord une méthode de simplification syntaxique automatique à base de règles manuelles en utilisant un formalisme de représentation sémantique à base de graphes, la DMRS, dépassant les limites des méthodes actuelles du même type. De plus, nous implémentons cette méthode en utilisant divers outils informatiques.

Notre approche s'inspire des travaux de Narayan et Gardent (2014, 2015) et de Sulem et al., (2018a) qui ont utilisé des approches à base de représentations sémantiques (cf. 2.2.3). Les différences résident dans les points suivants :

1) Nous utilisons un formalisme qui combine les annotations sémantiques et syntaxiques, à la différence des travaux cités qui utilisent des formalismes sémantiques portant uniquement sur les rôles thématiques : le formalisme DMRS.

2) Nous étendons leur approche afin de traiter des constructions syntaxiques qui n'ont pas été traités, telles que les constructions appositives et les subordinations.

Un des principaux desiderata pour le développement du MRS (cf. 3.6.3.1) est la compatibilité grammaticale : les représentations sémantiques doivent être liées à d'autres types **d'informations grammaticales** (notamment la syntaxe) et l'inclusion de toutes les **informations sémantiquement liées** qui peuvent être dérivées de la **syntaxe et de la morphologie**. Mel'čuk (1988) explique qu'un des points forts de ce formalisme est que, sur le plan linguistique, le format syntaxique repose sur la notion de dépendance syntaxique et sur celle, plus générale, de relation grammaticale : **les dépendances**

syntaxiques et les relations grammaticales coïncident largement avec les dépendances sémantiques. C'est pourquoi une représentation DMRS ressemble à une représentation de dépendances universelles (*Universal Dependencies* UD).

Mel'čuk a contrasté la notion de dépendances syntaxiques avec celles de dépendances sémantiques (Kahane et Gerdes, 2022). Les dépendances sémantiques sont les relations prédicat-argument qui existent entre les signifiés des sémantèmes. Par exemple, un verbe à l'infinitif n'a jamais de sujet, mais il possède toujours un argument sémantique qui peut être restitué par une paraphrase où le verbe est à une forme finie. Si on compare *Marie promet à Pierre de venir* et *Marie permet à Pierre de venir*, dans le premier cas 'Marie' est l'argument de 'venir' (Marie promet qu'elle viendra) et dans le deuxième cas 'Pierre' est l'argument de 'venir' (Marie permet que Pierre vienne). Dans les deux cas, la dépendance sémantique ne correspond pas à une dépendance syntaxique, car ni Marie de venir, ni à Pierre de venir ne sont des unités syntaxiques. Actuellement, le seul format de structures sémantiques sous-spécifiées qui repose sur la notion de dépendance sémantique est la DMRS (Copestake, 2009). La Figure 15 est une représentation DMRS.

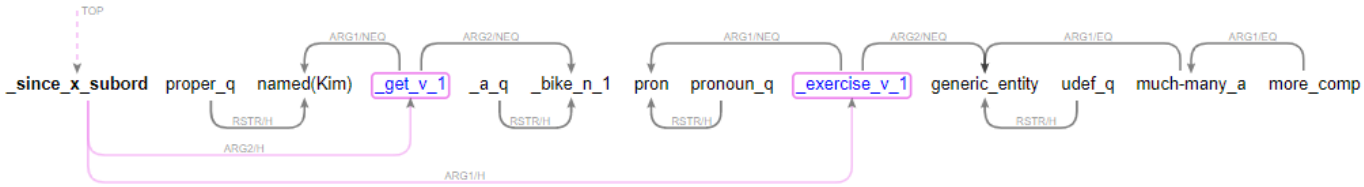


Figure 15 - DMRS de la phrase "*Since Kim got a bike, she exercises more*". Les relations syntaxiques sont annotées (*_since_x_subord*), ainsi que les rôles thématiques (relations prédicat-argument).

La plupart des représentations sémantiques s'appuient sur des analyses syntaxiques, de sorte qu'il existe un fort chevauchement entre les constituants sémantiques et syntaxiques. Cela s'applique particulièrement sur DMRS. **Sa sémantique est ancrée dans la forme superficielle des phrases et dans les liens syntaxiques entre les nœuds**. De plus, elle couvre des informations concernant la coréférence et les informations spatiales, temporelles et discursives.

Dans les taches de génération de textes comme la simplification de textes, l'ordre des mots est important. Tous les nœuds doivent être ancrés à phrase sous-jacente. Le point fort de DMRS est **la facilité d'ancrage des unités sémantiques dans les mots de la phrase**. L'une des principales différences entre les représentations sémantiques existantes est leur *niveau d'abstraction* de la phrase et relation *l'ancrage* des nœuds aux mots d'une phrase donnée. Cette relation est forte pour DMRS. **DMRS intègre des phénomènes linguistiques variés**. Les terminaux sont des éléments porteurs de sens, tandis que les arêtes permettent de caractériser les liens entre les unités, chaque jeton (ou un groupe de jetons dans le cas d'entités nommées, telles que des noms propres et des dates) correspond à un nœud dans le graphe. DMRS offre une plus grande flexibilité dans la représentation du sens, par exemple, par des affixes ou des constructions

phrastiques et facilitent la décomposition lexicale. D'où l'élément-clé qui définit notre choix de ce formalisme est que l**es concepts sémantiques sont ancrés aux phrases**.

Dans DMRS, les adverbes sont de deux classes : 1) les adverbes « scopaux », c'est-à-dire des adverbes dont la portée peut varier lorsque la phrase comporte un quantifieur et/ou une négation, quelle que soit leur position dans la phrase, tels que « *probably* », qui prend les prédicats comme arguments. 2) les adverbes « non-scopaux » tels que « *quickly* », qui prend les événements comme arguments. Prenons la phrase (28) ci-dessous comme exemple. « *Quickly* » dans l'exemple (28a) modifie la manière dont Mary a passé la balle, mais « *Probably* » de l'exemple (28b) a deux interprétations possibles dépendant de sa portée : soit juste le passage de la balle est incertain, soit la chaîne entière des événements.

(28) a. *Quickly*, Mary passed the ball to John and he scored.
b. *Probably*, Mary passed the ball to John and he scored.

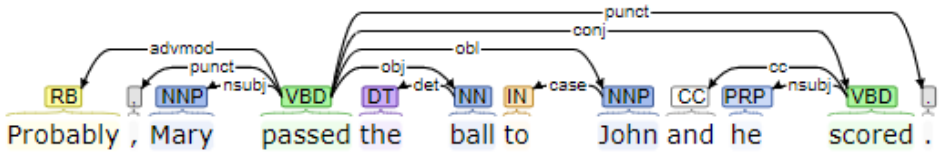
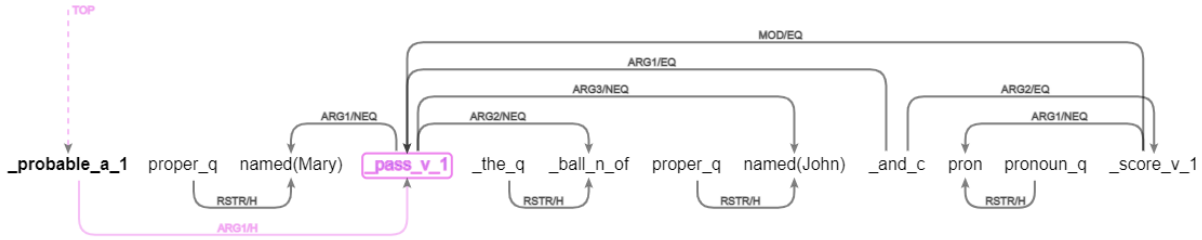
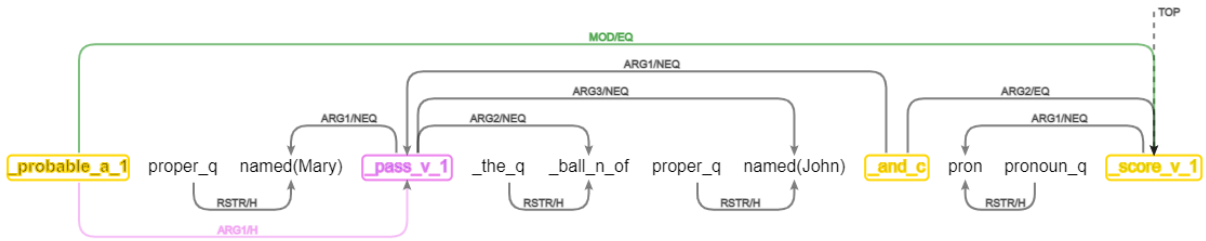


Figure 16 - Structure de dépendance de Stanford[32] de l'exemple 28b.

Contrairement aux structures de dépendance syntaxique (Figure 16), les représentations sémantiques, telles que DMRS, sont capables de distinguer les deux résolutions de portée (Figures 17a et 17b).



(a) DMRS pour l'exemple 28b avec « *probably* » portant sur toute la phrase



(b) DMRS pour l'exemple 28b avec « *probably* » portant sur la coordination gauche

Figure 17 - Deux résolutions de portée de l'exemple 28b[33].

[32] https://corenlp.run/
[33] http://delph-in.github.io/delphin-viz/demo/

La distinction est pertinente pour la division de phrases car dans cette dernière interprétation, les deux propositions ne sont pas indépendantes et toute tentative de division doit préserver les informations sur le modifieur appliqué aux deux propositions.

4.2. Présentation générale de la méthode

Pour proposer notre méthode de simplification syntaxique à base de sémantique utilisant le formalisme de représentation DMRS, nous nous sommes basés sur les quarante premiers textes du corpus Newsela (cf. 1.5.1.2). Il s'agit d'un jeu de données pour la simplification de textes qui a comme avantage le fait de proposer différentes versions de simplification à partir d'un texte original. Dans cette section, nous décrivons notre méthode de simplification syntaxique.

4.2.1. Les grandes étapes de la méthode

La méthode de simplification syntaxique que nous proposons est structurée en trois étapes principales, comme illustré dans la Figure 18. La première étape vise à représenter la phrase complexe par une représentation du sens basée sur un graphe DMRS. La deuxième étape consiste à transformer ce graphe DMRS en un ou plusieurs graphes DMRS en appliquant un ensemble de règles de transformation définies manuellement (règles de simplification). La troisième étape consiste à générer les phrases simplifiées à partir de ces graphes DMRS transformés.

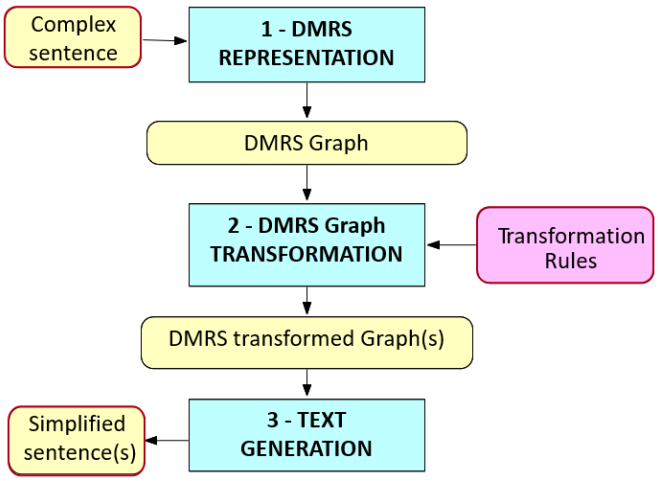


Figure 18 - Etapes de la méthode proposée.

Nous avons développé nos règles de simplification en examinant des phrases dans des textes bruts et en transformant des modèles structurels en graphes DMRS. Actuellement, notre système permet la simplification syntaxique de 5 constructions grammaticales : la coordination (exemple 29), la subordination (exemple 30), les appositives (exemple 31), les relatives (exemple 32), les formes passives (exemple 33).

(29) The wave traveled across the Atlantic, and organized into a tropical depression off the northern coast of Haiti on September 13.
(30) He settled in London, devoting himself chiefly to practical teaching.
(31) Finally, in 1482, the Order dispatched him to Florence, the city of his destiny.
(32) It is located on an old portage trail which led west through the mountains to Unalakleet.
(33) Most of the songs were written by Richard M. Sherman and Robert B. Sherman.

Pour simplifier ces constructions, nous extrayons des indicateurs déclencheurs (les liens, les arguments des conjonctions ou des prépositions). Le développement des règles dépend de la structure des phrases en anglais. Nous avons établi une étude linguistique pour chacune des constructions syntaxiques à traiter, les *patterns* de construction pour élaborer les *stratégies* et les traduire en règles manuelles.

4.2.2. Règles de réécriture de graphes DMRS

Notre méthode de simplification syntaxique utilise cinq paquets de règles de transformation de représentation DMRS des phrases à simplifier. Quatre paquets de règles permettent de réaliser la division de phrases complexes présentant des structures syntaxiques spécifiques, comme la coordination, l'apposition, la subordination, les relatives, et un cinquième qui réalise le passage de la voix passive à la voix active de la phrase considérée.

Pour le développement de règles, nous avons considéré la présence d'un verbe comme un indicateur de l'introduction d'une nouvelle situation ou évènement. *English Ressource Grammar* (ERG) (cf. 3.2.2) suit une analyse dans laquelle les éventualités (les événements) correspondent non seulement aux verbes, mais aussi aux autres parties du discours, aussi bien dans les constructions prédicatives qu'attributives. Le sous-graphe résultant est ensuite exploré à son tour à la recherche de nœuds de déclenchement inférieurs, et ainsi de suite, jusqu'à ce que nous atteignions un sous-graphe sans déclencheurs. Les sous-sections suivantes parcourent les modules de règles d'un point de vue linguistique.

À partir d'une représentation sémantique, les paires de division candidates sont sélectionnées à partir d'événements qui partagent un agent commun. La division est déterminée par les rôles sémantiques associés à un candidat, tout en préservant un ordre SVO (sujet-verbe-objet), considéré comme une structure agent-verbe-patient. Il s'agit d'appliquer des règles de transformation de graphes DMRS en distinguant les indices et indicateurs déclencheurs. Pour chaque segmentation, nous identifions les nœuds et les liens entre eux et un point de découpage qui joue le rôle de déclencheur, dont sa présence indique la possibilité de segmentation.

Comme nous l'avons indiqué dans la section 2.1, dans le cadre de cette thèse, nous nous intéressons essentiellement aux deux opérations suivantes : le découpage des phrases et le passage de la voix passive en voix active. Dans la suite de cette section, nous présentons

les structures syntaxiques que nous avons choisi de traiter. A chaque traitement d'une structure syntaxique correspond un paquet de règle spécifique.

4.3. Division dans le cas des appositions

4.3.1. L'apposition

Une apposition est un syntagme nominal qui suit un autre syntagme nominal et a pour fonction d'identifier le nom précédent ou de fournir des informations supplémentaires. L'apposition est utilisée pour rendre le message clair pour le décodeur en évitant l'ambiguïté (Bitea, 1977). Il s''agit d'une relation dans laquelle la seconde unité de l'apposition ajoute une spécificité à l'interprétation de la première (Meyer, 1992).

Blanche-Benveniste (2013b) affirme qu'à la lecture, on peut confondre les virgules des appositions avec celles qui marquent des énumérations. Tous ces éléments détachés, qui rompent la liaison de l'énoncé, forment des « îlots de prédilection » pour y mettre des implicites, des sous-entendus et des présupposés que le lecteur doit reconstruire. Un enchâssement, c'est-à-dire une subordination, entraîne une augmentation, des assemblages d'éléments lexicaux dans la phrase. Si une phrase contient une apposition, un certain nombre de mots vont s'additionner à la proposition initiale et par la suite, une augmentation du nombre de mots dans une phrase, ce qui entraine une difficulté syntaxique.

Rioul (1983) distingue quatre acceptions différentes du terme « apposition » :

a. « L'apposition a parfois été définie comme le rapport de contiguïté linéaire immédiate entre deux noms réunis dans une unité de rang supérieur, quelle que soit leur relation sémantique.
b. L'apposition est considérée comme la relation de subordination qu'entretient un nom avec un autre nom ayant même référence extralinguistique que lui, les syntagmes auxquels ils appartiennent ne pouvant être séparés l'un de l'autre.
c. On a également souvent nommé apposition la relation à distance, par-delà au moins une pause, entre un syntagme nominal ou un syntagme adjectival et le SN auquel il se rattache.
d. Enfin, le terme d'apposition, achevant la dérive qui le soustrait à son origine purement nominale, s'est même trouvé être réservé à la relation entre un syntagme adjectival mobile et le SN auquel il se rattache ».

Dans notre travail, nous nous sommes limités à l'apposition non-restrictive à antécédent nom propre. Un exemple typique est donné en (34).

(34) *Paul*, my friend, is a professor.

Dans cet exemple, *Paul* et *my friend* font référence à la même personne. Ces deux syntagmes nominaux sont dans une relation d'apposition. Le syntagme nominal *Paul* est considéré comme l'antécédent et *my friend* comme l'appositive.

Cette juxtaposition de deux noms est considérée comme un exemple paradigmatique de l'apposition et on ne traite que ce cas dans notre travail. Cependant, d'autres exemples pourraient être également considérés comme des cas d'apposition. Les caractéristiques définitoires de la construction comprennent, entre d'autres, le caractère restrictif ou non-restrictif de l'apposition, la catégorie syntaxique de l'antécédent et de l'appositive (Sopher, 1971), la coréférence de l'antécédent et de l'appositive, la possibilité de supprimer un des deux termes (Bitea, 1977 ; Hollenbach, 1983), la possibilité d'inverser l'antécédent et l'appositive (Burton-Roberts, 1975 ; McCawley, 1995) et la présence de marqueurs d'apposition.

Quirk et al. (1985) définissent des sous-catégories de constructions appositives, pleine/partielle, stricte/faible. Dans une apposition pleine (exemple 35a), chacun des termes peut être supprimé sans nuire à l'acceptabilité de la proposition. Chacun des termes remplit la même fonction syntaxique dans les propositions qui résultent de la suppression. Le sens de la phrase qui résulte de la suppression ne diffère pas du sens de la proposition d'origine. Si une de ces conditions n'est pas remplie, l'apposition est une apposition partielle (35b, c, d).

(35) a. Apposition pleine : A neighbour, *Fred Brick*, is on the telephone. *Cf*. *Fred Brick*, a neighbour, is on the telephone.
b. Apposition partielle : An unusual present was given to him on his birthday, *a book on ethics*. *Cf.* *Was given to him on his birthday, *a book on ethics*.
c. Apposition partielle : Norman Jones, *at one time a law student*, wrote several bestsellers. *Cf*. **At one time a law student* wrote several best-sellers.
d. Apposition partielle : The reason he gave, *that he didn't notice the car till too late*, is unsatisfactory. *Cf*. *That he didn't notice the car until too late* is unsatisfactory.

Une appositive peut être restrictive ou non-restrictive. Quirk et al. (1985) définissent ces deux types de modification de la façon suivante. Nous traduisons :

- Modification *restrictive* : une modification est restrictive lorsque la référence de la tête est un membre d'une classe qui ne peut être identifiée que par la modification qui a été fournie. [...] La restrictivité indique alors une limitation de la référence possible de la tête.
  Exemple : *Mr Campbell the lawyer was here last night.* [M. Campbell l'avocat par opposition à tout autre M. Campbell que nous pourrions connaître.]

- Modification *non restrictive* : Alternativement, le référent d'un syntagme nominal peut être considéré comme unique ou comme membre d'une classe qui a été identifiée indépendamment (par exemple dans le contexte précédent). Toute modification apportée à une telle tête est une information supplémentaire non indispensable à l'identification.
  Exemple : *Mr Campbell, a lawyer, was here last night*.

Huddleston et Pullum (2002) font remarquer que le critère de restriction n'est pas toujours utile. Ils donnent l'exemple (36). Ils notent qu'il n'y a aucune implication que la première personne ait plus d'un mari : la construction restrictive intégrée fournit simplement une manière succincte de dire que la personne concernée est mon mari et s'appelle George[34].

(36) This is *my husband* George.

La différence concernant ces deux relations sémantiques entre les constituants en apposition se manifeste également au niveau de l'intonation et de la ponctuation (Quirk et al., 1985 ; Huddleston et Pullum, 2002). Une appositive restrictive n'est pas séparée de sa proposition hôte par des virgules ou par une démarcation intonative, tandis qu'une appositive non-restrictive est entourée de virgules ou de tirets dans le langage écrit et reçoit une intonation typique des incises dans le langage oral.

Cette virgule peut être une source d'ambigüité dans le traitement de la phrase chez certains individus. Considérons la dédicace en exemple (37). Avec la virgule, le lecteur pourrait comprendre la dédicace comme signifiant soit que le livre est dédié aux trois parties (la mère d'une part, Laura d'autre part et à Dieu), soit que le livre est dédié à la mère de l'écrivain, qui se trouve être Laura Stephan, et à Dieu.

(37) À ma mère [,] Laura Stephan [,] et à Dieu.

Quirk et al., (1985) établissent une échelle sémantique à plusieurs niveaux qui décrit la relation entre l'antécédent et l'appositive :

- **Equivalence**
  - *Appellation* : les deux termes de l'apposition sont définis et le second est typiquement un nom propre et le second constituant de la relation appositive est généralement plus spécifique que l'antécédent. Exemple : *The company commander, that is to say Captain Madison, assembled his men and announced their mission.*
  - *Identification* : l'antécédent est typiquement indéfini ; le second constituant de la relation appositive est plus spécifique et il identifie le référent de l'antécédent. Exemple : *A company commander, namely Captain Madison, assembled his men and announced their mission.*
  - *Désignation* : c'est l'inverse des relations d'identification et d'appellation : l'antécédent est plus spécifique que le l'appositive et les deux termes sont typiquement définis. Exemple : *Captain Madison, that is to say the company commander, assembled his men and announced their mission.*

[34] There is no entailment or implicature that I have more than one husband: the integrated [restrictive] construction simply provides a succinct way of saying that the person concerned is my husband and is named George.

- *Reformulation* : le second constituant de la relation appositive reprend le contenu lexical de l'antécédent. Exemple : *Sound units of the language, technically phonemes, are usually surrounded by slant lines* : /p/.

- **Attribution**
  Implique la prédication. Les marqueurs d'apposition ne sont pas admis dans cette catégorie. Le second constituant de la relation appositive est souvent indéfini dans ces exemples. Exemple : *Captain Madison, a company commander, assembled his men and announced their mission.*

- **Inclusion**
  - Exemplification : le second constituant de la relation appositive fournit des exemples de l'ensemble désigné par l'antécédent. Exemple : *They visited several cities, for example Rome and Athens.*
  - Particularisation : est un cas d'exemplification où le second constituant est un exemple important ou saillant de l'ensemble désigné par l'antécédent. Exemple : *The children liked the animals, particularly the monkeys.*

4.3.2. Division de l'apposition

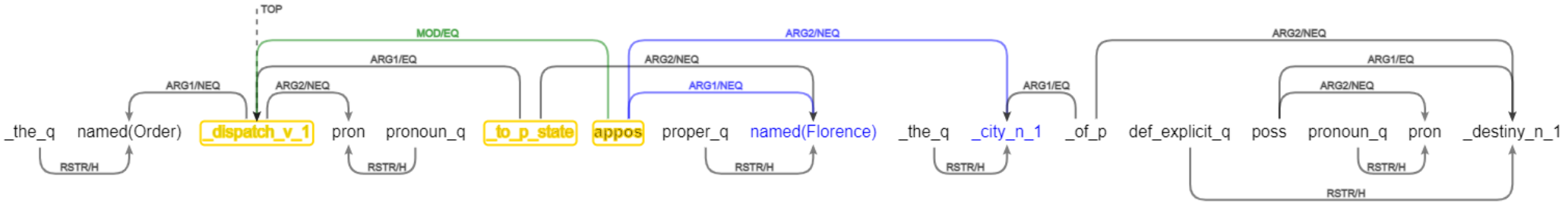
Pour notre étude, nous nous concentrons exclusivement sur les appositives non-restrictives. Dans DMRS, l'apposition est identifiée par une relation de type *appos* qui prend les deux noms adjacents comme arguments. Dans notre travail de thèse, nous avons traité les trois types d'apposition : apposition en début de phrase (exemple 38), en milieu de phrase (exemple 39) et en fin de phrase (exemple 40). Au contraire des approches à base de la syntaxe qui demandent de traiter ces trois cas séparément, notre méthode traite l'apposition en utilisant une seule base (paquet) de règles. Dans ce qui suit, nous développons comment nous avons procédé pour la transformation des appositives. Nous prenons la phrase (40) comme exemple, mais la même méthode s'applique sur les autres phrases. Le but est de transformer la phrase a en phrase b.

(38) a. Florence, *the capital city of Tuscany*, is one of the most beautiful and famous towns of the world.
b. Florence is the capital city of Tuscany. Florence is one of the most beautiful and famous towns of the world.

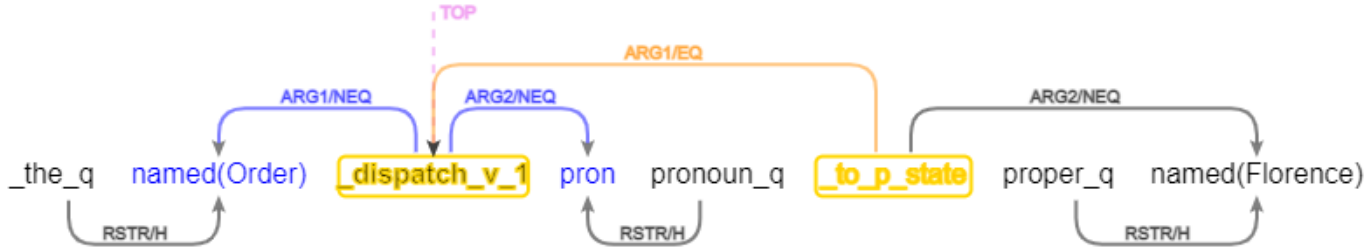
(39) a. You can't visit Florence, *one of the most picturesque cities in Europe*, without going to see the Duomo.
b. You can't visit Florence without going to see the Duomo. *Florence* is one of the most picturesque cities in Europe.

(40) a. The Order dispatched him to Florence, *the city of his destiny*.
b. The Order dispatched him to Florence. Florence is the city of his destiny.

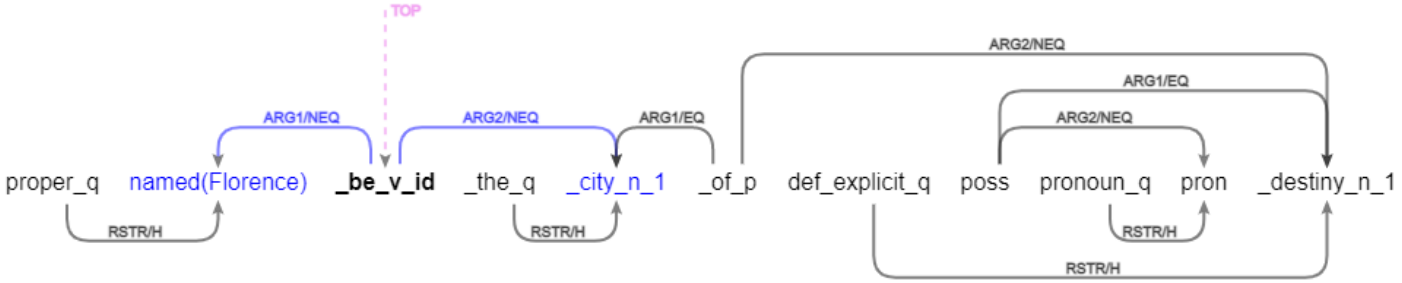
La Figure 19a est la représentation DMRS de la phrase originale d'entrée (40a). Les Figures 19b et 19c sont les représentations DMRS des deux phrases simplifiées de sortie en (40b). L'idée de base est de réécrire l'antécédant de l'apposition avant le verbe et de transformer l'apposition en phrase indépendante afin de construire deux nouvelles DMRS. Ainsi, d'une manière plus générale, la règle du découpage d'apposition supprime d'abord le nœud *appos* ainsi que le nœud de son ARG2 (l'apposition '*city*') pour former une première DMRS, puis il construit l'autre DMRS en réécrivant l'ARG1 (l'antécédant *Florence*) d'*appos* avant ARG2 ('*city*'). Pour passer du graphe (Figure 19a) aux deux sous-graphes (Figures 19b et 19c), un ensemble d'étapes de transformation de graphes est appliqué. Il s'agit d'abord de définir le graphe sur lequel il faut appliquer l'ensemble de règles. D'où, la stratégie suivante : nous cherchons une relation d'apposition « *appos* », dont l'ARG2 est toujours un nom (pos=n) et jamais un nom propre avec une relation ARG1/NEQ entre le nœud *appos* et l'antécédant d'une part et une relation ARG2/NEQ entre le nœud *appos* et l'apposition. Une fois ce graphe trouvé, il s'agit d'appliquer un ensemble de règles de transformation. Nous traitons les appositions non-restrictives dont l'ARG2 n'est pas un nom propre. Nous ne traitons pas les cas tels que la phrases « *The professor Campbell was here last night* ». Les résultats d'application de ces règles sont représentés dans les Figures 20a à 20e.



a. DMRS de la phrase "*The Order dispatched him to Florence, the city of his destiny*".



b. DMRS de la phrase "*The Order dispatched him to Florence*".



c. DMRS de la phrase "*Florence is the city of his destiny*".

Figure 19 - DMRS de la phrase 43 originale et sa simplification

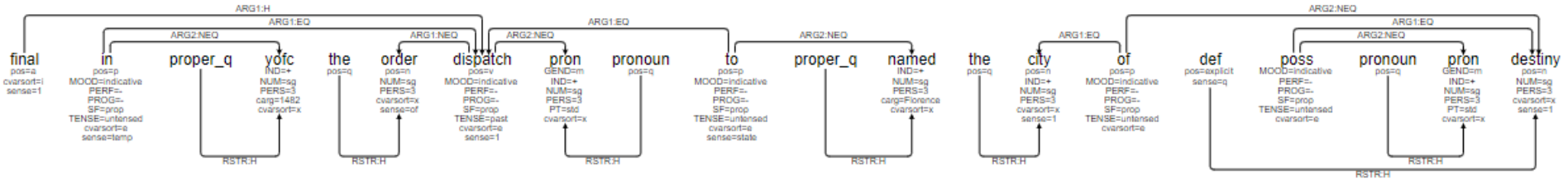
La stratégie de simplification des différents cas d'appositions est la suivante :

1) Copier la composante connexe [l'antécédant (N) de la relation d'apposition+le nœud *proper_q*] avant le nœud *proper_q*, d'une façon à avoir deux composantes connexes qui se suivent. Il s'agit de copier les nœuds avec tous les traits morpho-syntaxiques, ainsi que les liens entre les éléments de cette composante connexe ;

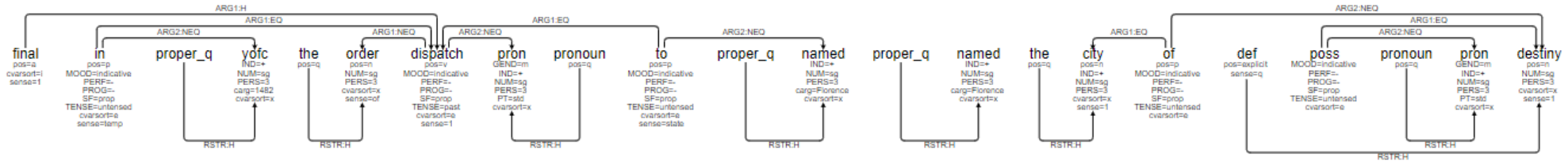
2) Rajouter le verbe être V après le nœud N1 (la copie du nœud N *Florence*). Pour le verbe être, nous avons décidé qu'il soit toujours au présent de l'indicatif comme il s'agit d'une définition (vérité générale) dans la plupart des cas. Il s'agit de définir les traits morpho-syntaxiques du verbe être (le temps, l'aspect, le mode, etc.) ;

3) Rajouter les arêtes entre V (le verbe être) et le nœud N1 et M (l'apposition au départ). L'ARG1 de V est N (*Florence*) et l'ARG2 est M (*city*) ;

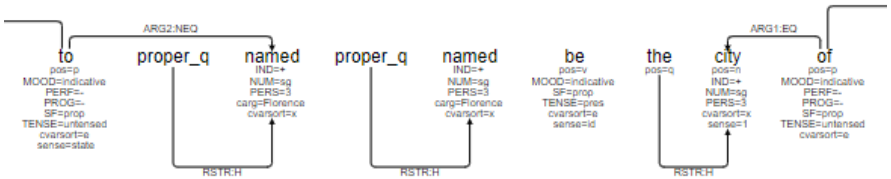
4) Supprimer le nœud de la relation d'apposition.



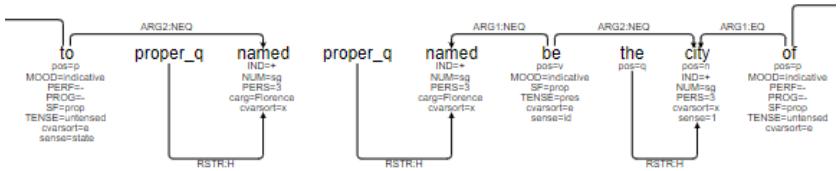
a. DMRS obtenu après application de l'étape 1 : suppression du nœud *appos* qui marque une relation d'apposition (deux sous-graphes).



b. Graphe DMRS obtenu après application de l'étape 2 : copie de la composante connexe de l'apposition (nœuds + arêtes).



c. Partie du graphe DMRS obtenu après application de l'étape 3 : ajout du verbe « *to be* » après le Nœud N2 de l'antécédant *Florence* copié dans l'étape 2



d. Partie du graphe DMRS obtenu après application de l'étape 4 : Ajout des arêtes entre les arguments du verbe *« to be »* rajouté à l'étape 3

Figure 20 - Etapes de transformation de graphes DMRS de la phrase 43a en 43b

4.4. Division dans le cas des coordinations

4.4.1. La coordination

La coordination correspond aux cas où deux propositions sont liées par une conjonction de coordination ou un adverbe conjonctif. Chaque proposition peut fonctionner par elle-même comme une phrase portante une idée. Lorsqu'on fusionne deux phrases en rajoutant une conjonction entre elles, on fusionne deux propositions. Cet empaquetage de propositions entraine l'accumulation du nombre d'idées par phrase.

Une des ambiguïtés qu'un lecteur peut rencontrer dans un texte est la coordination, par exemple, la phrase « *Put the butter in the bowl and the pan on the towel »*. Le syntagme nominal *the pan* est ambigu car il peut être soit un emplacement, soit un objet à déplacer (Engelhardt et Ferreira, 2010). L'ambiguïté de coordination a été étudiée dans plusieurs études en lecture (Frazier, 1987 ; Hoeks et al., 2006 ; Hoeks et al., 2002 ; Staub et Clifton, 2006). Elle est illustrée dans les exemples (41a) et (41b). Notant que l'ambiguïté dépend du fait que le syntagme nominal suivant la conjonction fait partie d'un objet complexe comme dans l'exemple (41a), ou le sujet d'une phrase conjointe comme dans l'exemple (41b).

(41) a. Mary bumped into [$_{NP1}$the busboy] and [$_{NP2}$the waiter] last Saturday night.
b. [$_{S1}$ Mary bumped into the busboy] and [$_{S2}$ the waiter told her to be careful].

Une des premières études portant sur le traitement des structures de coordination ambiguës a été menée par Frazier (1987). Elle a trouvé des temps de lecture plus lents en néerlandais lorsqu'il s'agit d'une S-coordination (*Sentential coordination,* exemple 42c). Hoeks (1999) affirme que lorsqu'une phrase ambiguë telle que celle en (42a) est lue, il n'est pas immédiatement clair pour le lecteur si *photographer* fait partie d'un objet direct du verbe principal *embraced* (42b) et dans ce cas on parle de *Noun-Phrase coordination* ou, alternativement, s'il fait l'objet d'une nouvelle phrase conjointe (comme en 42c) où il s'agit d'une *Sentential coordination.*

(42) a. The model embraced the designer and the photographer laughed.
b. The model embraced [the designer and the photographer] because she was very happy. *(NP-coordination)*
c. [The model embraced the designer] and [the photographer laughed]. *(S-coordination)*
d. [The model embraced the designer], and [the photographer laughed].

Plusieurs travaux (MacDonald et al., 1994 ; Tanenhaus et Trueswell, 1995 ; Frazier et Clifton, 1996) montrent que les lecteurs préféreront la NP-coordination (exemple 42b) à la S-coordination (exemple 42c). En raison de cette préférence, les lecteurs rencontreront des difficultés de traitement lorsqu'il s'agit d'une S-coordination. Hoeks et al. (2002) montrent que les lecteurs préfèrent la NP-coordination, non pas pour des raisons de

simplicité syntaxique, mais parce que la NP-coordination est plus simple en termes de *structure thématique*[35]. Dans les NP-coordinations, il n'y a qu'un seul thème, tandis que les S-coordinations contiennent un thème supplémentaire. Selon Hoeks et al. (2002), avoir plus d'un thème est très inattendu et entraînera des difficultés de traitement, car les lecteurs devront intégrer la deuxième entité, non introduite, comme thème dans leur modèle mental du discours (Crain et Steedman, 1985 ; Lambrecht, 1996).

Afin d'évaluer d'éventuelles difficultés de traitement, des phrases ambiguës ont été comparées à des phrases non ambiguës (exemple 42d). La virgule attachée au *designer* désambiguïse la phrase vers la S-coordination.

4.4.2. Division des coordinations

Dans notre travail, nous avons essayé de passer d'une S-coordination à deux phrases indépendantes. Dans DMRS, la coordination est identifiée par toute relation qui a un suffixe ***_c_***, comme *_and_c_* et *_or_c_*.
Les structures de liens des graphiques reflètent les différences de type de coordination (exemple 43). Les graphiques DMRS de ces exemples sont présentés à la Figure 21.

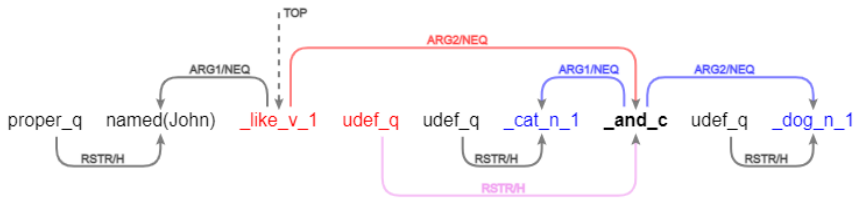
(43) a. John likes cats and dogs. (Coordination entre deux noms).
b. John likes cats but hates dogs. (Coordination entre deux propositions partageant le même sujet).
c. John likes cats and Mary likes dogs. (Coordination entre deux propositions qui ne partagent pas le même sujet).
d. * John likes cats. John likes dogs.

Dans notre travail, nous avons traité le cas de coordinations entre deux propositions, celles qui partagent un sujet commun (43b) et d'autres qui ne le partagent pas (43c). Le découpage des coordinations entre deux noms entraine un contre-effet (43d).

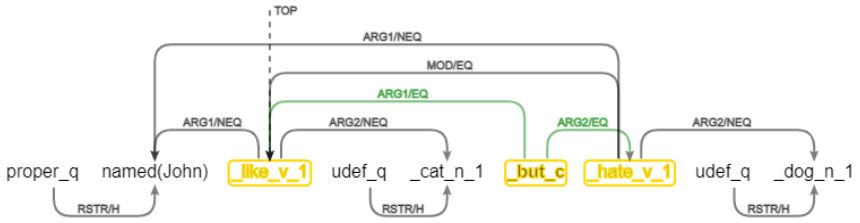
Pour la coordination entre deux noms (Figure 21a), la conjonction prend comme arguments les noms en coordination. Pour la coordination entre deux propositions, la conjonction prend comme arguments les deux verbes.

Pour distinguer les coordinations entre les propositions qui partagent le même sujet avec celles qui ne le partagent pas, il faut signaler dans la règle que les deux verbes doivent avoir un sujet commun avec un lien ARG1/NEQ commun entre les deux verbes et le sujet. La différence réside dans le nombre d'événements dans la phrase *_like_v_1* et *_hate_v_1* font référence au même sujet *proper_q (named (John))*. Le sujet *John* est régi par les deux verbes de la phrase, *like* et *hate*.

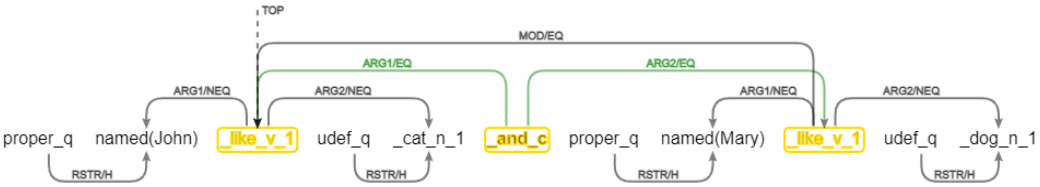
[35] La structure thématique décrit la relation entre le thème d'une phrase – c'est-à-dire l'élément faisant référence à une entité sur laquelle des informations sont données – et l'information exprimée par une phrase (Hoeks et al., 2006 ; Lambrecht, 1996).



a. DMRS de la phrase 43a : John likes cats and dogs.



b. DMRS de la phrase 43b : John likes cats but hates dogs.



c. DMRS de la phrase 43c : John likes cats and Mary likes dogs.

Figure 21 - Différents types de coordination : (a) une coordination entre deux noms ; (b) une coordination entre deux propositions partageant un sujet partagé ; (c) coordination entre deux propositions qui ne partagent pas un sujet partagé.

Dans la représentation davidsonienne de la sémantique, tous les verbes introduisent des événements. Ainsi les deux verbes introduisent deux événements différents. Davidson (1967) suppose qu'un verbe d'action transitif exprime une relation à trois places entre les deux arguments nominaux et un argument (« davidsonien »).

Dans le cas où il s'agit d'une coordination mais avec absence d'une conjonction (le cas d'asyndète), il y a dans l'ERG 2018 un prédicat de grammaire unique *implicite_conj* qui suit principalement les mêmes règles aux fins de la transformation DMRS. Un schéma particulier apparaît dans la coordination multiple (Figure 22), où elle prend en ARG1 le premier verbe et la coordination la plus à droite est l'ARG2 de la conjonction implicite. Le graphe est modifié dans le but de lisibilité.

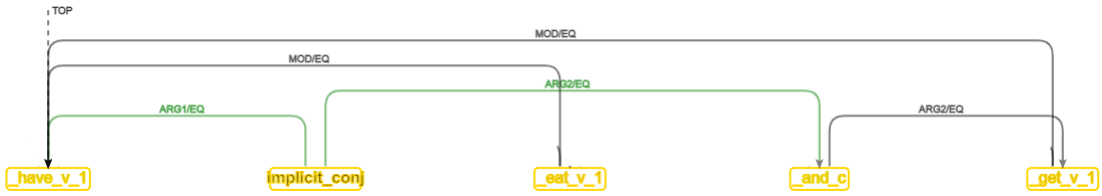


Figure 22 - Triple coordination dans "*The boy had chips, the girl ate a cookie, and the dog got nothing*".

4.4.3. Coordination entre deux événements partageant le même sujet

Pour la coordination entre deux propositions qui partagent le même sujet (exemple 44), il s'agit d'abord de définir le graphe sur lequel il faut appliquer l'ensemble de règles.

(44) a. The book, Political Economy, was published in 1985, but had limited classroom adoption.

b. The book, Political Economy, was published in 1985. But the book had limited classroom adoption.

D'où, la stratégie est la suivante : nous cherchons un nœud de la conjonction (_and_c, _or_c, etc.), dont l'ARG1 et l'ARG2 sont des verbes. Il faut que les deux verbes aient le même sujet, d'où l'argument ARG1 soit pareil pour les deux verbes. Une fois ce graphe est recherché, il s'agit d'appliquer un ensemble de règles de transformation. La seule conjonction qui sera supprimée lors de la simplification est « *and* ». Toutes les autres conjonctions seront conservées. Il s'agit de supprimer toutes les arêtes entre les deux événements et reconstruire le sujet partagé, afin de construire deux DMRS indépendantes et par la suite deux phrases. Les résultats d'application de ces règles sont représentés dans les Figures 23a à 23d.

La stratégie de simplification des coordinations entre deux événements qui partagent le même sujet est la suivante :

- Si la conjonction est « *and* » :

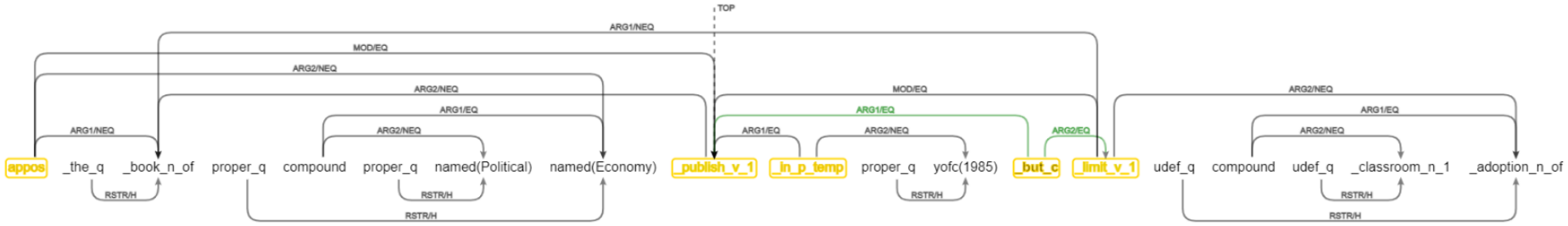
1) Supprimer le nœud de la conjonction ;

2) Copier la composante connexe du sujet partagé avant le deuxième verbe V2. Il s'agit de copier les nœuds avec tous les traits morpho-syntaxiques, ainsi que les liens entre les éléments de cette composante connexe.

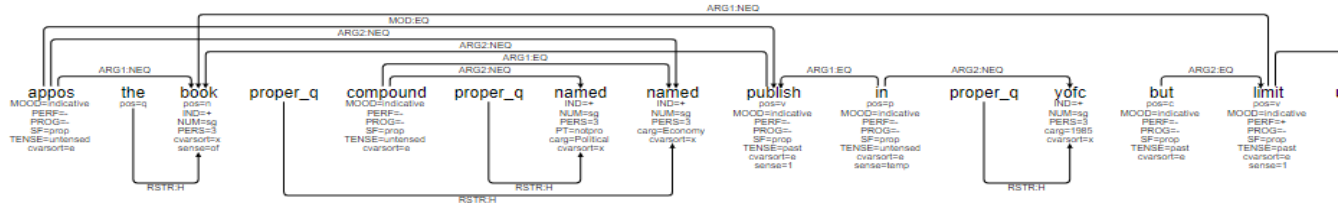
3) Supprimer tous les liens entre V2 et V1 et entre V2 et le sujet partagé.

4) Rajouter les arêtes (ARG1) entre V2 et le sujet partagé reconstruit.

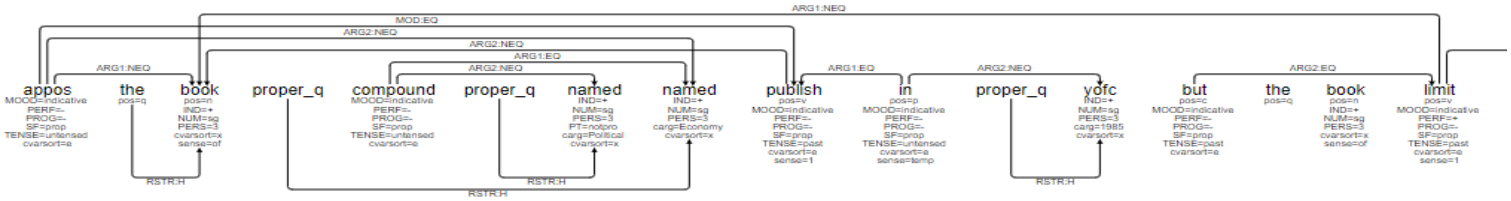
- Si la conjonction est une conjonction autre que « *and* », il s'agit de suivre les mêmes étapes précédentes mais sans supprimer le nœud de la conjonction.



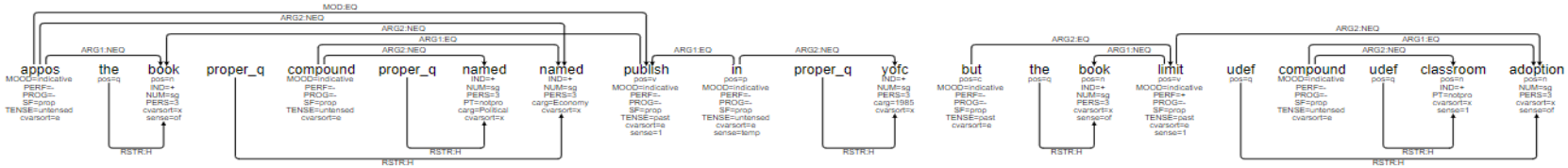
a. DMRS de la phrase “The book, Political Economy, was published in 1985, but had limited classroom adoption”.



b. Supprimer l’arrête entre la conjonction et V1.



c. Copier la composante connexe du sujet partagé avant le deuxième verbe.



d. Supprimer tous les liens entre V2 et V1 et entre V2 et N1 et rajouter les arêtes (ARG1) entre V2 et le sujet partagé reconstruit.

Figure 23 - DMRS de la phrase d’entrée 44a et sa simplification 44b

4.4.4. Coordination entre deux événements ne partageant pas le même sujet

Pour la coordination entre deux propositions qui ne partagent pas le même sujet (exemple 45), il s'agit d'abord de définir le graphe sur lequel il faut appliquer l'ensemble de règles. La stratégie est la suivante : nous cherchons un sous-graphe avec un nœud de la conjonction (_and_c, _or_c, etc.), dont l'ARG1 et l'ARG2 sont des verbes. Une fois ce sous-graphe trouvé, il s'agit d'appliquer un ensemble de règles de transformation. Comme pour la coordination entre deux événements qui partagent un sujet, la seule conjonction qui sera supprimée lors de la simplification est « *and* » et toutes les autres conjonctions seront conservées.

(45) His real date of birth was never recorded, but it is believed to be a date between 1935 and 1939.

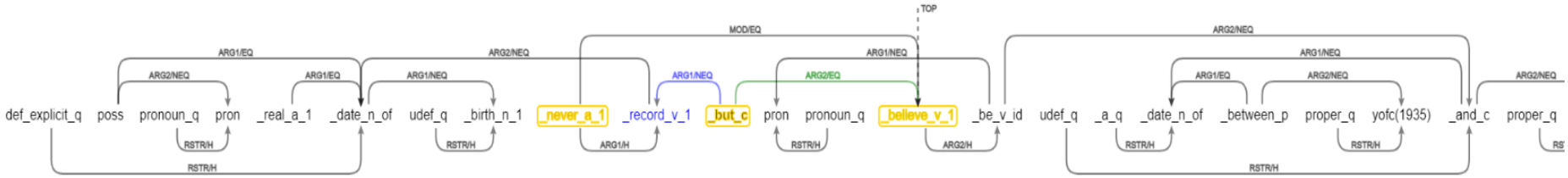
b. His real date of birth was never recorded. But it is believed to be a date between 1935 and 1939.

La stratégie de simplification des coordinations entre deux événements qui ne partagent pas le même sujet est la suivante :

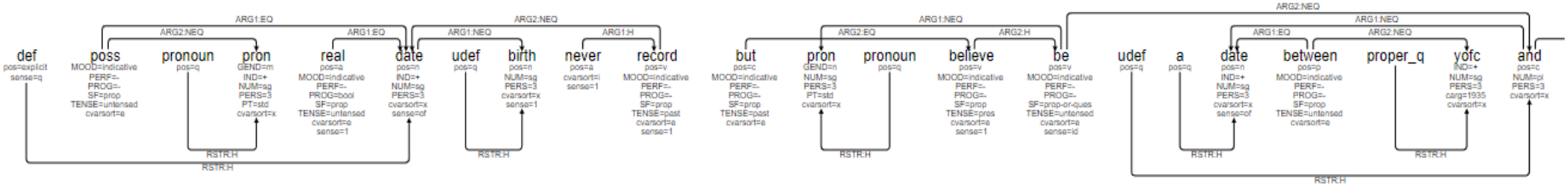
- Si la conjonction est « *and* », il suffit de supprimer le nœud de cette conjonction et le lien MOD/EQ entre les deux verbes.

- Si la conjonction est une conjonction autre que « *and* », il s'agit de supprimer le lien entre la conjonction et le premier verbe et le lien entre les deux verbes.

Les résultats d'application de ces règles sont représentés dans les figures 24a et 24b.



a. DMRS de la phrase: “His real date of birth was never recorded, but it is believed to be a date between 1935 and 1939”.



b. Graphe DMRS obtenu après suppression du lien entre la conjonction et le premier verbe. Nous avons déjà deux sous-graphes

---

Figure 24 - Etapes de transformation de graphes DMRS de la phrase 45a en 45b

4.5. Division dans le cas des subordinations

4.5.1. Les subordinations

Les effets de la structure des phrases sur la compréhension des enfants sont un sujet souvent débattu par les chercheurs. Une préoccupation spécifique a été de savoir si l'utilisation de subordination dans les phrases affecte les capacités de compréhension chez les jeunes lecteurs. L'utilisation et la longueur des subordonnées sont considérées parmi les mesures dans le calcul du « score de densité syntaxique ». Il s'agit d'une « description qualitative et quantitative des structures linguistiques » qui peut avoir un impact dans la lecture (Golub et Kidder, 1974).

Pearson (1974) donne un autre point de vue sur cette question. Il conclut après une série d'expériences que l'utilisation de propositions subordonnées n'augmente pas le niveau de difficulté ni n'affecte la lisibilité. Il conclut que les données n'appuient pas la recommandation selon laquelle la difficulté du discours écrit peut être réduite en éliminant les constructions subordonnées ou en réduisant la longueur des phrases. Lorsque la relation sémantique est maintenue constante et que la question du test est pertinente pour la relation dont la forme varie, soit la compréhension est également efficace entre les formes, soit les formes de phrases plus subordonnées et plus longues suscitent une meilleure compréhension.

La coordination entre deux propositions, avec *and, but, or, for, nor, yet, et so,* concerne la liaison de structures parallèles tandis que la subordination aide à ajouter du contenu supplémentaire à une proposition indépendante dans des phrases complexes. Les subordonnées sont considérées comme le dispositif formel de subordination le plus important, en particulier pour les propositions finies (Quirk et al., 1985). Selon Edwards (1980), les subordonnées sont notoirement plus difficiles à traiter que leurs formes indépendantes. Il y a deux explications possibles à cela, toutes deux de nature sémantique. Selon la première explication, la nature dépendante du contexte de la subordination pourrait bien impliquer un traitement sémantique plus « sophistiqué » que les propositions indépendantes correspondantes. Une explication syntaxique alternative est que les relations syntaxiques introduites par la subordination sont plus complexes que les relations entraînées par la proposition indépendante (Morgan, 1978). Les recherches sur la subordination ont mis en évidence les dépendances complexes créées par ce type de proposition. Turner et Greene (1977) affirment que la difficulté relative de la subordination est particulièrement marquée pour les jeunes enfants. Par rapport à des conjonctions de subordination telles que *although, even though, unless or when*, elles déclarent que même lorsque les enfants peuvent lire ces structures sans difficulté, ils fournissent la preuve, lorsqu'ils discutent ou racontent ce qu'ils ont lu, qu'ils ne comprennent pas les relations sémantiques. Les auteurs prétendent que cela est dû au fait que les conjonctions de coordination sont structurellement plus flexibles et sémantiquement moins contraintes que les conjonctions de subordination. Selon eux, les

structures coordonnées concernent principalement des relations temporelles consécutives ou concurrentes. Avec les structures subordonnées, en revanche, nous introduisons les facteurs sémantiques plus subtils de « cause » et « effet ».

Les propositions subordonnées (propositions dépendantes) ont plusieurs fonctions possibles dans une phrase. Comme le montrent les exemples ci-dessous, ils peuvent fonctionner comme des propositions nominales, adverbiales, relatives ou comparatives (exemples tirés de Quirk et al., 1985).

- Nominale : I noticed (that) he spoke English with an Australian accent.
- Adverbiale : We left after the speeches ended.
- Relative : I took what they offered me.
- Comparative : Caroline is less perceptive than Rosemary.

Les propositions subordonnées peuvent être introduites par une conjonction dans une proposition finie comme dans les exemples ci-dessus, ou bien sous une forme réduite avec le subordonnant complètement omis, une proposition sans verbe, un infinitif, à l'infinitif, ou une proposition avec –*ed*, ou –*ing* (Quirk et al., 1985). Les propositions subordonnées fonctionnent comme des « ponts » reliant les informations supplémentaires à la proposition principale.

Sur la base de leur fonction potentielle, nous distinguons plusieurs grandes catégories fonctionnelles de propositions subordonnées nominales, adverbiales, relatives et comparatives. On peut présenter graphiquement ces différentes distinctions en termes de formes et de fonctions comme en Figure 25 (Quirk et al., 1985).

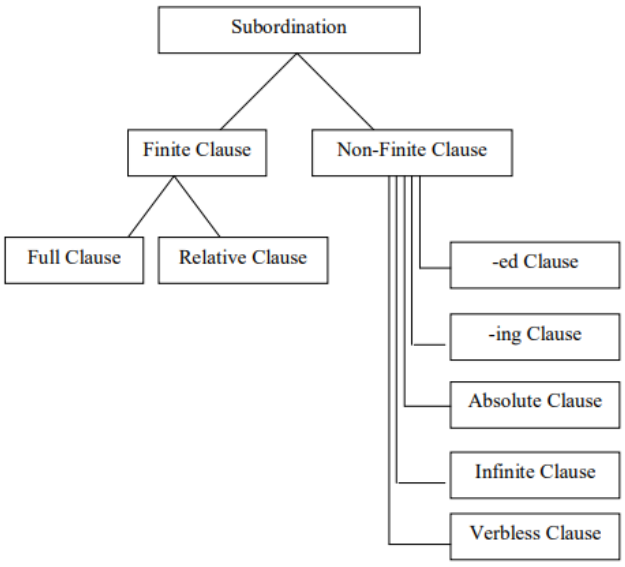


Figure 25 - Catégories fonctionnelles de propositions subordonnées (Quirk et al., 1985)

Dans notre système, une règle de découpage des subordinations traite les propositions ***adverbiales*** (*Needing money to pay my rent, I forced myself to beg my parents)* et les

participiales (*The participants will be presented with a book, edited by the town council*) (cf. Tableau 1 de Stajner (2016)).

English Ressource Grammar (ERG) encode plus d'informations que d'autres techniques de traitement superficielles, telles qu'une Grammaire hors-contexte probabiliste (*Probabilistic context-free grammar* PCFG). Cependant, une grammaire PCFG donne une lecture faible pour la phrase 46 (Figure 26).

(46) Given that he has no money, I can see no way he will pay you.

Premièrement, il découpe la conjonction « *Given that* » en deux constituants. Chacun se voit attribuer une balise POS différente (Given ← VBN, that ← IN). Deuxièmement, la subordonnée « *he has no money* » n'est marquée que par un S mais n'est pas explicitement spécifiée comme une proposition subordonnée. Ainsi, cette balise POS peut indiquer une relative ou une subordonnée, mais un traitement superficiel n'indique pas la distinction.

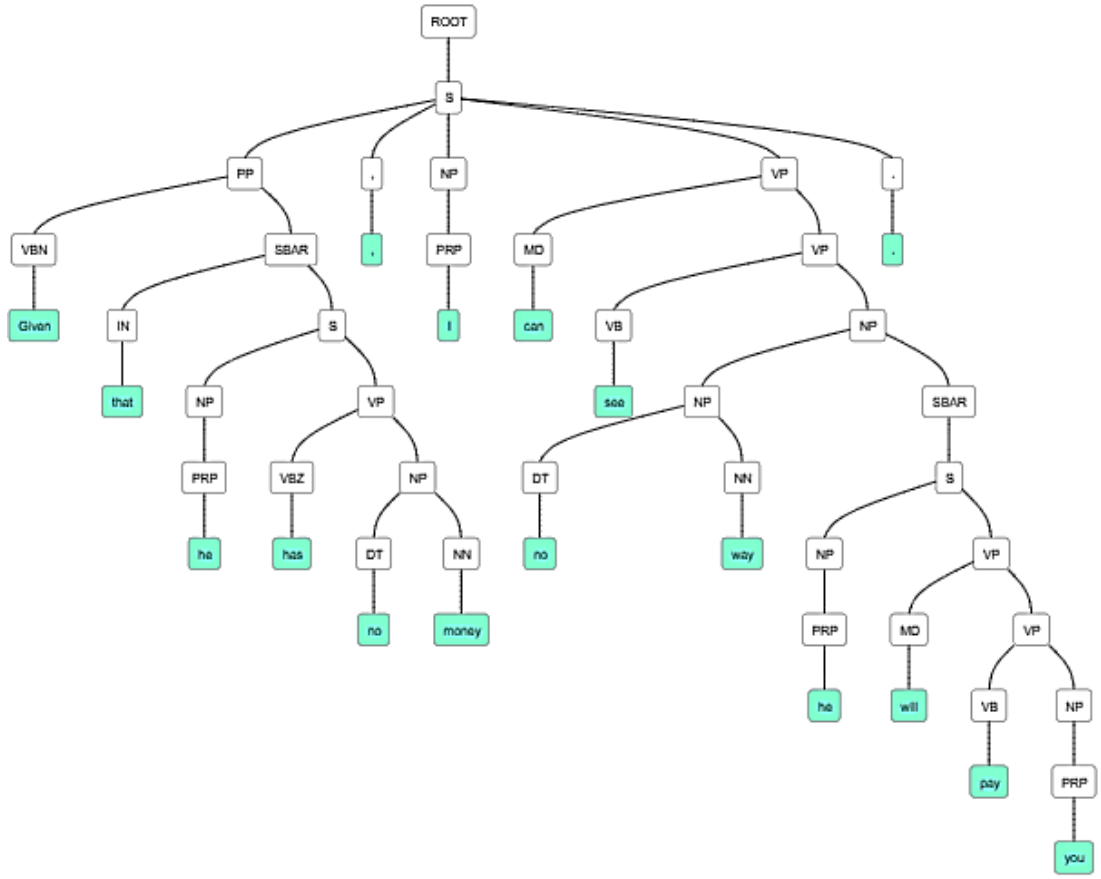


Figure 26 - PCFG de la phrase « *Given that he has no money, I can see no way he will pay you* »[36]. La conjonction « *Given that* » n'est pas correctement marquée. La subordonnée « *he has no money* » n'est pas indiquée explicitement.

La partie avant la virgule est la proposition dépendante, qui commence par une conjonction de subordination « *Given that* » et suivie d'une proposition indépendante. Dans ERG, la conjonction de subordination est principalement identifiée par un suffixe

[36] http://nlpviz.bpodgursky.com/

*_subord_*, tel que *_given+that_x_subord_*, *_once_x_subord_*, ou parfois un motif *_x _*, tel que *_although_x_*, *_unless_x_*.

4.5.2. Division de la subordination

Dans les approches à base des arbres syntaxiques, pour simplifier des participiales et les adverbiales, il s'agit de développer deux (ou plusieurs règles). Cependant, notre méthode les traite dans une seule règles comme il s'agit d'un *motif* commun : une relation de subordination avec une relation ARG*/H entre deux verbes.

Nous avons développé une règle de découpage qui ressemble à la règle de découpage de la coordination. L'ARG1 de la subordonnée fait référence à la principale tandis que l'ARG2 fait référence à la subordonnée. Ainsi, la règle de découpage extrait tous les nœuds liés à ARG1/2 séparément et construit deux nouvelles DMRS. Chaque décision de segmentation identifie également deux propositions. Le but est de de transformer une subordonnée en principale et de réécrire le sujet partagé. Les conjonctions de subordination dans l'ERG sont des opérateurs de portée à deux arguments, et les liens de portée agissent comme des liens déclencheurs pour les propositions découpées. Les liens ARG1/H et ARG2/H mènent respectivement à la proposition principale et à la proposition subordonnée. La phrase peut être fragmentée comme le montre la Figure 27.

L'idée de base est de supprimer la relation de subordination et de transformer la forme verbale non-personnelle (en *-ed* et *-ing*), en réécrivant l'antécédant avant le verbe, afin de transformer la subordination en phrase indépendante. Le but est de reconstruire deux nouvelles structures DMRS. Ainsi, d'une manière plus générale, la règle du découpage des subordinations supprime d'abord le nœud *subord* pour former une première DMRS pour que la principale forme le premier DMRS, puis il construit l'autre DMRS en réécrivant le sujet partagé (ARG1 de verbe de la principale) avant le verbe de la subordonnée et en modifiant les traits morpho-syntaxiques de ce verbe. La Figure 27a est la représentation DMRS de la phrases 47a (phrase originale d'entrée) et les Figures 27b et 27c sont les DMRS des phrases de sortie (47b).

(47) a. He stayed in Italy, devoting himself to nursing.
b. He stayed in Italy. He devoted himself to nursing.

Pour passer du graphe (Figure 27a) aux deux sous-graphes (Figures 27b et 27cb), un ensemble d'étapes de transformation de graphes est appliquée.

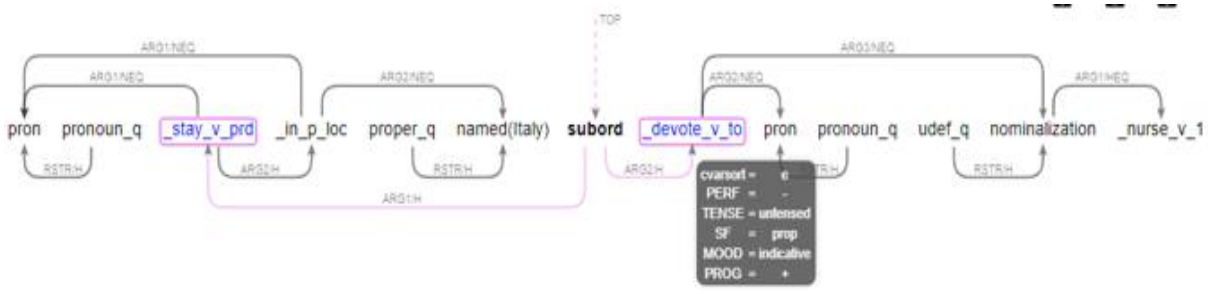
La stratégie de simplification des subordinations est la suivante :

1) Supprimer le nœud de la relation de subordination ;
2) Copier la composante connexe de l'antécédant (ARG1 du verbe de la principale) avant le verbe de la subordonnée.
3) Modifier le temps du verbe en subordonnées en lui affectant les mêmes traits morpho-syntaxiques du verbe de la principale.
4) Rajouter les arêtes entre le verbe de la subordination et le sujet copié.

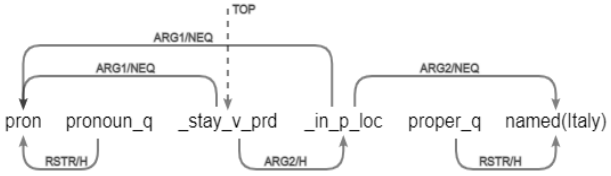
Il s'agit d'abord de définir le graphe sur lequel il faut appliquer l'ensemble de règles. La stratégie est alors la suivante :

- Nous cherchons une relation de subordination « *subord* » dont l'ARG1 et l'ARG2 sont deux verbes.
- Une fois ce graphe est recherché, il s'agit d'appliquer un ensemble de règles de transformation.

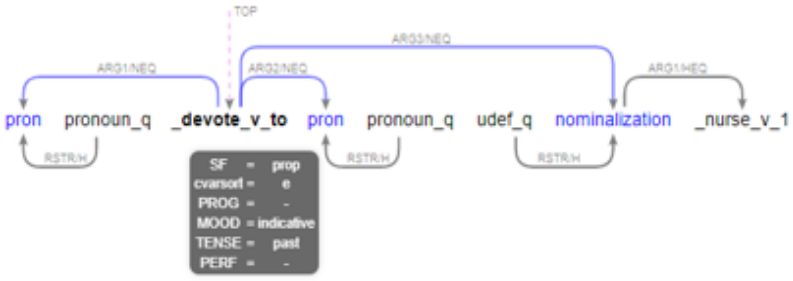
Ceci est illustré à la Figure 27.



a. DMRS de la phrase *He stayed in Italy, devoting himself to nursing*.



b. DMRS de la phrase *He stayed in Italy*



c. DMRS de la phrase *He devoted himself to nursing*

Figure 27 - DMRS de la phrase 47a et 47b

4.6. Division dans le cas des relatives

4.6.1. Les propositions relatives

Crain et Steedman (1985) ont inspiré un grand nombre d'études portant sur l'effet du contexte référentiel sur le traitement syntaxique. Prenons la phrase (48) :

(48) The psychologist told the woman that he was having trouble with to visit him again.

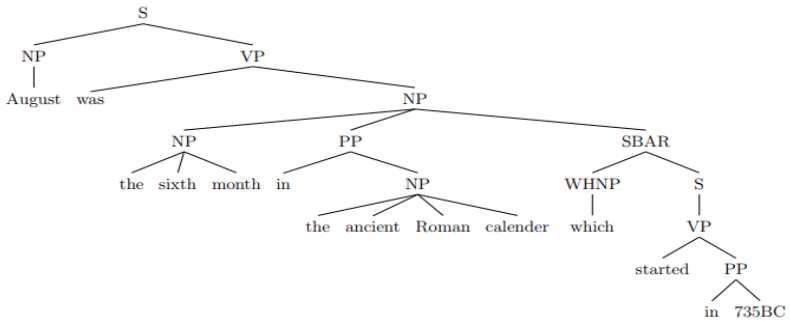
La relative dans cette phrase (*that he was having trouble with*) est ambiguë, en ce sens qu'elle pourrait être soit un complément du verbe « *told* », soit un modificateur du SN « *the woman* ».

Coleman (1965) a constaté que les propositions relatives écrites sous des formes fortement enchâssées comme (49a) étaient plus difficiles à retenir chez les adultes que celles écrites sous des formes moins fortement enchâssées, comme (49b).

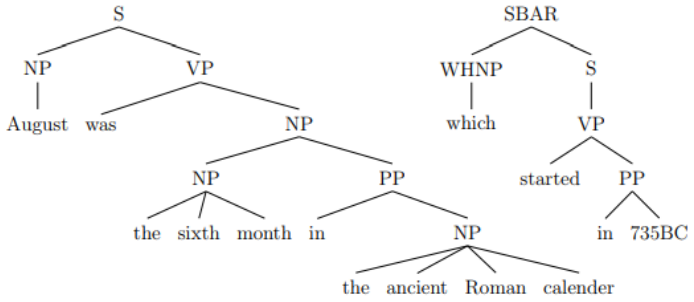
(49) a. The rat that the cat killed ate the malt.
b. The cat killed the rat that ate the malt.

Zhu et al. (2010) ont proposé un modèle de simplification inspiré de la traduction automatique basée sur les arbres syntaxiques. Étant donné une phrase complexe à simplifier, ils analysent la phrase complexe et appliquent sur l'arbre d'analyse pour simplifier quatre opérations : division, suppression, réorganisation et substitution. Nous montrons un exemple (50) tiré de Zhu et al., (2010) pour expliquer comment se déroule la simplification.

(50) a. August was the sixth month in the ancient Roman calender which started in 735BC.
b. August was the sixth month in the old calender. The old calender started in 735BC.



a. Arbre d'analyse de la phrase complexe 50a



b. Découpage de l'arbre d'analyse

Figure 28 - Simplification de la phrase 50 en utilisant les arbres syntaxiques (Zhu et al., 2010).

Dans cet exemple, la division se déroule en deux étapes : la segmentation et la complétion. L'étape de segmentation décide si on peut diviser l'arbre d'analyse en jugeant le constituant syntaxique du point de découpage et la longueur de la phrase complexe. Etant donné l'arbre d'analyse de la phrase complexe (50a) illustrée à la Figure 28a, il s'agit de décider s'il est possible de diviser le point de découpage « *which* » avec le constituant syntaxique « SBAR ». La Figure 28b montre l'arbre d'analyse segmenté dans le cas d'une division réussie. L'opération de division ne s'arrête pas là et passe par une étape de complétion. L'étape de complétion reconstruit les phrases après découpage ; il décide de laisser tomber ou de conserver le point de découpage en utilisant deux caractéristiques : le point de découpage et ses constituants directs ; et il copie les parties nécessaires (sujet partagé) pour compléter les nouvelles phrases. La dernière partie est jugée par deux traits : les relations de dépendance et le constituant. La dernière étape consiste à supprimer le mot limite « *which* » et copier la phrase NP entière « *ancient Roman calendar* » dans la phrase de droite.

4.6.2. Division de la proposition relative

Bien que les pronoms relatifs indiquent des propositions relatives, dans une structure DMRS, ces pronoms relatifs ne sont pas explicitement représentés. Avec ou sans pronom relatif, l'ERG donne une interprétation DMRS identique aux deux phrases. Prenons la phrase (51) comme exemple :

(51) a. It is located on the trail which led to the mountain.
b. It is located on the trail. The trail led to the mountain.

La Figure 29a représente la phrase originale et les Figures 29b et 29c représentent les DMRS des phrases simplifiées. Dans la figure 29a, il n'y a pas un nœud pour le pronom relatif « *that* ». Cependant, le verbe *_lead_v_1* régit son sujet par une relation /EQ. Cela indique que *_lead_v_1* et *_trail_n_1* partagent la même étiquette et ont la même portée. Après découpage de la phrase, cette contrainte de même portée doit être résolue.

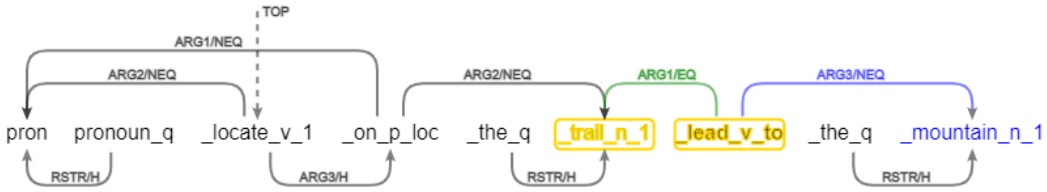
Il s'agit d'abord de définir le graphe sur lequel il faut appliquer l'ensemble de règles. La stratégie est la suivante : nous cherchons un nœud qui doit être un nom, suivi d'un verbe et une relation ARG1/EQ partant du verbe vers le nom. Nous avons choisi de ne pas prendre toute la composante connexe mais juste le nœud de l'antécédant et le nœud attaché, par exemple, pour le cas d'un nom propre, il s'agit de prendre le nœud de ce dernier avec le nœud attaché formant avec lui un nœud RSTR/H. Une fois ce graphe est recherché, il s'agit d'appliquer un ensemble de règles de transformation.

La stratégie de simplification des relatives est la suivante :

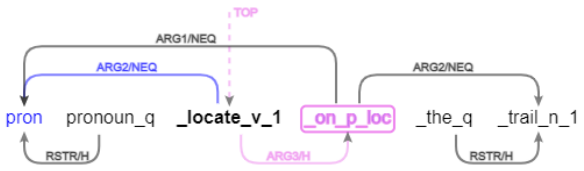
1) Supprimer l'arrête entre le verbe et l'antécédant.

2) Copier la composante connexe de l'antécédant avant le verbe de la relative d'une façon à avoir deux antécédant qui se suivent puis le verbe.

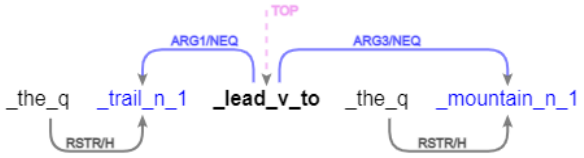
3) Rajouter une arête ARG1/NEQ entre l'antécédant copié et le verbe.



a. DMRS de la phrase *It is located on the trail which led to the mountain.*



b. DMRS de la phrase It is located on the trail.



c. DMRS de la phrase The trail led to the mountain.

Figure 29 - DMRS de la phrase 51a et sa simplification 51b et 51c

4.7. Passage de la voix passive à la voix active

4.7.1. La voix passive

La voix passive est la construction grammaticale dans laquelle un nom principal fonctionnant comme sujet d'une phrase, d'une proposition ou d'un verbe est affecté par l'action d'un verbe ou est agi par le verbe. Le nom fonctionnant comme sujet grammatical est généralement le destinataire de l'action désignée par le verbe plutôt que par l'agent, et peut être utilisé pour éviter d'attribuer la « responsabilité » à l'auteur (Choomthong, 2011 ; Crystal, 2011). Ainsi, dans une phrase passive en anglais, le sujet logique (l'agent) sort de la position de sujet grammatical et est relégué à une phrase dérivée (Brinton et Brinton, 2010). La construction de la voix passive implique donc l'inversion des positions du sujet du syntagme nominal (SN) et de l'objet du SN. C'est un mouvement syntaxique.

L'hypothèse de Chomsky (1965), selon laquelle les phrases grammaticalement complexes telles que le passif doivent être transformées dans leur structure de base afin d'arriver à leur représentation sémantique, est compatible avec un large éventail de preuves ; les phrases passives et négatives sont en effet plus difficiles à générer, apprendre, comprendre et retenir (Mehler, 1963 ; Miller, 1962 ; Savin et Perchonock, 1965 ; Sachs, 1967). Dans une phrase à la forme passive, l'ordre des mots est modifié, l'objet est maintenant au début de la phrase, suivi de l'action et du sujet qui la pose. Ce changement demande à des individus ayant des difficultés de modifier sa façon habituelle de traiter l'information.

Deux types de voix passives existent en anglais, à savoir les passives agentives (*agentive passives*) et les passives sans agent (*agentless passives*). Le passif agentif prend toujours un nom d'agent ; c'est-à-dire l'exécutant de l'action doit être mentionné. Par conséquent, le marqueur *by* est obligatoire. Par exemple : *Purple Hibiscus is written by Chimamanda Adiche*. Au contraire, les passifs sans agent ne prennent pas *by* parce qu'il n'y en a pas besoin puisque l'accent est mis sur l'action et non sur le sujet de l'action. Par exemple : *Many roads were constructed*.

Des études sur l'enseignement, l'apprentissage et l'utilisation de la voix passive en anglais dans des contextes natifs et non natifs ont montré qu'elle pose des difficultés aux apprenants. Pullum (2014), Alvin (2014) et Moreb (2016) ont attribué les difficultés aux « attitudes négatives associées à son utilisation », car le plus souvent, les enseignants mettent en garde leurs élèves contre son utilisation et recommandent plutôt l'utilisation de la voix active. Hinkel (2001 ; 2004) et Murcia et Freeman (1999) cités dans Neilson (2016) ont également noté qu'enseigner, apprendre quand et comment utiliser la voix passive présente la plus grande difficulté pour les enseignants et les apprenants de l'anglais langue seconde et l'anglais langue étrangère (ESL/EFL).

Dans les Figures 30a et 30b, nous montrons un exemple de règles manuelles basées sur des structures de dépendances typées pour la simplification syntaxique. Nous avons emprunté cet exemple à Siddharthan et Mandya (2014). La Figure 30a montre les règles de conversion passive en active et de simplification de proposition relative. La Figure 30b (gauche) montre la structure de dépendance de la phrase complexe (52a). La règle pour

la proposition relative supprime la relation d'intégration « rcmod », lorsqu'il y a un sujet disponible pour le verbe dans la proposition relative. La règle de passif à actif supprime les relations « nsubjpass », « auxpass » et « agent » et insère deux nouvelles relations « nsubj » et « dobj » rendant la phrase active. Les opérations de nœud modifient les informations de temps et de nombre des mots en conséquence. Après l'application de ces deux règles sur 30b (gauche) nous nous retrouvons avec la structure de dépendance montrée en 30b (droite) qui génère la phrase simplifiée (52b).

(52) a. The cat was chased by a dog that was barking.
b. A dog chased the cat. A dog was barking.

```
Rule: Passive2Active
1) Delete
   a) nsubjpass(?X0, ?X1)
   b) auxpass(?x0, ?X2)
   c) agent(?X0, ?X3)

2) Insert
   a) nsubj(?X0, ?X3)
   b) dobj(?X0, ?X1)

3) Node Operations
   a) AGR-TENSE: ?X0 <-- ?X2
   b) AGR-NUMBER: ?X0 <-- ?X3
```

```
Rule: RelativeClause
1) Delete
   a) rcmod(?X0, ?X1)
   b) nsubj(?X1, ?X0)

2) Insert
   a) nsubj(?X1, ?X0)
```

a. Règles développées manuellement pour la transformation de la voix passive en voix active (à gauche) et la relative (à droite)

```
det(cat-2, The-1)
nsubjpass(chased-4, cat-2)
auxpass(chased-4, was-3)
det(dog-7, a-6)
agent(chased-4, dog-7)
nsubj(barking-10, dog-7)
aux(barking-10, was-9)
rcmod(dog-7, barking-10)
```

```
det(cat-2, The-1)
dobj(chased-4, cat-2)
det(dog-7, a-6)
nsubj(chased-4, dog-7)
aux(barking-10, was-9)
nsubj(barking-10, dog-7)
```

b. Structures de dépendance typées des phrases 54a et 54b

Figure 30 - Exemple de règles manuelles basées sur des structures de dépendances typées pour la simplification syntaxique (Siddharthan et Mandya, 2014).

4.7.2. Passage de la voix passive à la voix active

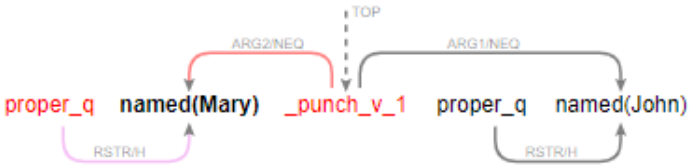
Une phrase sous sa forme active ou passive possède deux analyses syntaxiques, alors que ces deux phrases possèdent la même représentation sémantique. La phrase (53) est un exemple. Les Figures 31a et 31b représentent les DMRS de la même phrase sous sa forme passive et active respectivement. Nous cherchons le sous-graphe dont la stratégie est la suivante : un verbe dont son ARG2 (l'objet) précède son ARG1 (le sujet). Il s'agit de

prendre la composante connexe du sujet et le permuter avec la composante connexe de l'objet.

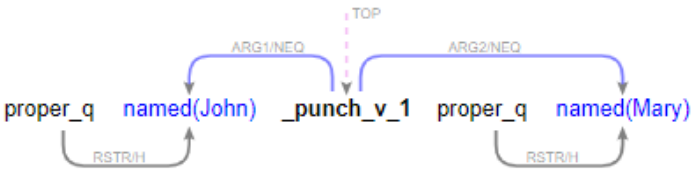
La stratégie de transformation des voix passives permet de permuter les deux arguments du verbe.

Cette stratégie est illustrée à la Figure 31 avec la phrase (53).

(53) a. Mary was punched by John.
b. Mary punched John.



a. DMRS de la phrase *Mary was punched by John*



b. DMRS de la phrase *John punched Mary*

Figure 31 - DMRS de la même phrase en voix active et en voix passive.

Conclusion

Dans ce chapitre, nous avons décrit les fondements théoriques de la méthode de simplification syntaxique de textes basée sur des transformations de représentation sémantique exprimée avec la notation DMRS. Pour cela nous avons considéré différentes constructions grammaticales en anglais en ce qui concerne les types de propositions qui y participent, en se basant sur une étude linguistique du corpus pour identifier les différentes stratégies linguistiques et en se basant sur les travaux de Huddleston et Pullum (2002). Nous avons identifié celles qui peuvent être détectées sans ambiguïté dans les graphes DMRS (celles qui peuvent avoir plusieurs formes selon la phrase). Pour chaque construction syntaxique à simplifier, nous avons défini une stratégie de transformation des graphes DMRS. Nous avons d'abord présenté les grandes étapes de la méthode et les différents composants de notre système. La méthode est structurée en trois étapes principales : (1) représenter la phrase complexe par une représentation du

sens basée sur un graphe DMRS ; (2) transformer ce graphe DMRS en un ou plusieurs graphes DMRS en appliquant un ensemble de règles de transformation définies manuellement (règles de simplification) ; (3) la dernière étape consiste à générer les phrases simplifiées à partir de ces graphes DMRS transformés. Pour finir, nous avons développé notre approche à base de règles pour chacune des constructions traitées, telles que le découpage des coordinations, des appositions, des subordinations, des relatives et la transformation de la voix passive en voix active.

Dans le chapitre suivant, nous décrivons GRASS (*GRAph-based semantic representation for Syntactic Simplification*), notre système de simplification implémentant notre méthode de simplification syntaxique proposée dans ce chapitre.

5. GRASS : un outil de simplification syntaxique basé sur la représentation sémantique DMRS

Dans ce chapitre, nous décrivons GRASS (*GRAph-based semantic representation for Syntactic Simplification*), le système de simplification mettant en œuvre notre méthode de simplification syntaxique proposée dans le chapitre 4. Nous expliquons l'architecture logicielle du système, commençant par la phase de l'analyse sémantique en graphes des phrases d'entrée, et pour chaque construction à simplifier, la transformation de ces graphes selon la stratégie de simplification définie dans le chapitre 4 pour cette construction, jusqu'à la phase de génération de graphes transformés.

Ce chapitre est organisé comme suit : la section 5.1 développe l'architecture logicielle du système et ses différents composants informatiques mis en œuvre et exploités dans notre travail, telles que le système de réécriture de graphes, GREW, qui est utilisé pour implémenter chaque stratégie sous forme d'une base de règles. Dans la section 5.2, nous mettons l'accent sur le composant « SIMPLIFICATION » de l'architecture du système GRASS en détaillant les règles développées pour chacune des constructions à traiter avec des exemples de transformation. La section 5.3 concerne le composant « TEXT GENERATION » associé à la génération de la phrase simplifiée à partir d'un graphe DMRS transformé par le composant « SIMPLIFICATION ». La section 5.4 définie la stratégie d'application des règles dans le système GRASS, et discute de la question de l'ordre dans laquelle sont appliquées les bases de règles GREW. La section 5.5 présente des écrans d'exécution de notre système GRASS. Nous clôturons ce chapitre par une conclusion.

5.1. Architecture logicielle du système

GRASS, notre système de simplification de textes, est un cadre général pour la simplification syntaxique en utilisant DMRS, y compris le prétraitement, l'analyse des phrases, la manipulation des graphes DMRS pour obtenir d'autres graphes et la génération des graphes obtenus après transformation. La Figure 32 montre l'architecture de GRASS. Les sous-sections suivantes décrivent chacun des composants.

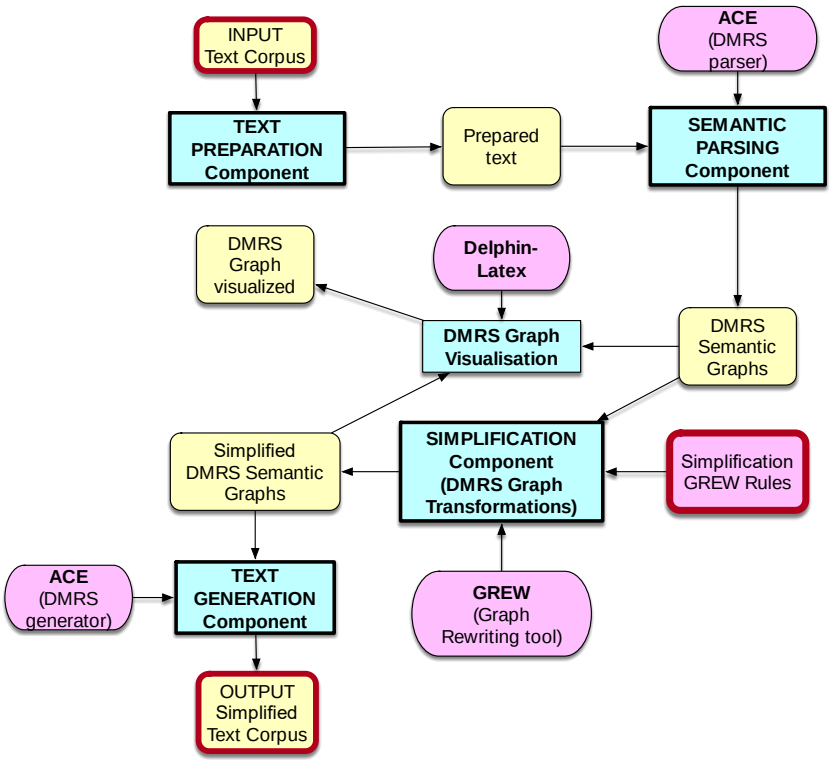


Figure 32 - Architecture logicielle du système GRASS avec ses principaux composants

5.1.1. Pré-traitement

Le composant « *TEXT PREPARATION* » prépare le corpus pour la simplification. La première opération est de le mettre dans un format interprétable pour le composant "Analyse sémantique" (il transforme chaque phrase du corpus en un graphe sémantique DMRS). En particulier, le corpus à traiter doit être découpé en phrases. Il est important de conserver la position des phrases dans le corpus original pour pouvoir les générer au bon endroit. Le texte d'origine est sous-forme de phrases brutes, chacune sur une ligne.

5.1.2. Analyse sémantique

L'analyse sémantique est effectuée par le composant « *SEMANTIC PARSING* » qui utilise l'outil ACE[37] (*Answer Constraint Engine*), développé par le Consortium DELPH-IN. ACE est un processeur pour les grammaires DELPH-IN : ACE permet à la fois de traduire une phrase en un graphe DMRS (analyseur ACE) puis de générer une phrase à partir d'un graphe DMRS (générateur ACE). Une phrase est prise en entrée et la sortie est un fichier au format MRS associé décrivant les informations sémantiques. Le format MRS ne peut pas être géré par les outils que nous avons choisis d'utiliser pour la visualisation et la transformation. Par conséquent, nous devons le transformer en un graphique DMRS à l'aide d'un utilitaire DELPH-IN.

[37] http://sweaglesw.org/linguistics/ace/

Notons que les graphes DMRS peuvent être manipulés à l'aide de deux bibliothèques Python existantes :

- La bibliothèque PyDelphin[38] est une bibliothèque dédiée à MRS que nous utiliserons pour convertir les structures MRS produites par ACE en graphiques DMRS. PyDelphin est une suite de bibliothèques Python pour le traitement des données et l'interaction avec les outils du DELPH-IN. L'objectif de PyDelphin est de réduire les obstacles à l'utilisation des ressources DELPH-IN pour aider les utilisateurs à créer rapidement des applications ou à réaliser des expériences, et il a été utilisé avec succès pour la recherche sur la traduction automatique (Goodman, 2018), l'analyse sémantique neuronale (Buys et Blunsom, 2017) et la génération des textes (Hajdik et al., 2019).
- La bibliothèque PyDMRS[39] (Copestake et al., 2016) est uniquement dédiée à la manipulation des graphes DMRS. Actuellement, les données DMRS sont stockées au format XML dans des fichiers texte.

Cependant, ces outils sont programmés en Python et l'expression d'un *motif négatif* qui impose des conditions négatives sur un graphe cherché et qui permet la non-application d'une règle est impossible. C'est pourquoi, nous avons choisi, pour la transformation des graphes, un système de réécriture de graphes indépendant de la famille DELPH-IN qui prend en considération d'exprimer des conditions négatives.

5.1.3. Transformation des graphes DMRS

Le composant « *SIMPLIFICATION* » réalise proprement dit la simplification des phrases par transformation des graphes DMRS selon les constructions linguistiques rencontrées. Ces transformations suivent les stratégies définies pour chacune des constructions étudiées dans le Chapitre 4 et sont réalisées par des bases de règles spécifiques en utilisant le système de réécriture GREW[40] (*Graph REWriting system*) (Guillaume et al, 2012 ; Bonfante et al., 2018 ; Guillaume, 2021), un outil de réécriture de graphes pour les applications en TAL qui peut manipuler des représentations syntaxiques et sémantiques. Un système de réécriture de graphes est un ensemble de règles qui décrivent des transformations élémentaires et qui sont appliquées successivement pour réaliser une transformation plus globale. Le principe de la réécriture de graphes est de modifier un sous-graphe dans l'arborescence générale pour qu'il corresponde à la structure finale voulue. Ces modifications peuvent s'appliquer au niveau des nœuds, des arcs, des attributs et des types des relations au niveau des arcs. La transformation de graphes permet donc beaucoup de liberté sur les types de transformations souhaitées.

Dans notre travail, ce composant permet de transformer le corpus, phrase par phrase, au niveau des graphes DMRS associés à chaque phrase du corpus. Bien qu'il ait été utilisé sur la séquence étiquetée POS, la syntaxe de dépendance de surface, la syntaxe de

[38] https://pydelphin.readthedocs.io/en/latest/index.html#
[39] https://github.com/delph-in/pydmrs
[40] https://grew.fr/

dépendance profonde, la représentation sémantique, GREW peut être utilisé pour représenter n'importe quelle structure basée sur des graphes. Chaque opération de simplification transformant un graphe DMRS est associée à un ensemble de règles GREW (ces règles sont présentées dans la section 5.2). GREW permet de transformer la représentation sémantique basée sur les graphes en DMRS selon un ensemble de règles, structurées en trois parties :

- ***Pattern*** (un patron positif) : décrit la partie du graphe à faire correspondre avec une autre partie. Il permet la sélection de nœuds ou d'arêtes grâce à leurs caractéristiques, relations ou positions dans le graphe.
- ***Without*** (un patron négatif) : filtre les occurrences à exclure des éléments d'une sélection précédente. C'est un ensemble de patrons négatifs qui sont des extensions du patron positif et qui imposent des conditions négatives sur le graphe cherché.
- ***Commands*** (ensemble de commandes) : permet d'appliquer des transformations structurelles sur le graphe, telles que la suppression, la création ou la réorganisation des nœuds et des arêtes ainsi que la modification de leurs caractéristiques dans le graphe.

GREW est structuré en modules. Il a été développé pour mettre en œuvre les opérations suivantes :

- **Recherche d'un motif** donné dans un ensemble de graphes à la fois pour corriger l'annotation d'un corpus et pour exploiter une annotation donnée à des fins linguistiques.
- **Transformation d'un graphe** en un autre graphe afin de convertir les annotations entre les différents formats.

Ces deux opérations sont en pratique très liées car les transformations utilisent des règles formées d'une paire (motif, transformation élémentaire) et l'application d'une telle règle nécessite la recherche du motif correspondant. La recherche de motifs est donc commune aux deux opérations.

GREW a notamment été utilisé pour diverses applications linguistiques :

- La transformation d'analyses syntaxiques en dépendances vers des graphes sémantiques (transformation en DMRS et AMR).
- L'ajout d'informations en syntaxe profonde sur des analyses syntaxiques de surface.
- La pré-annotation du corpus DeepSequoia (Candito et al., 2014 ; Perrier et al., 2014). Le projet DeepSequoia a porté sur la définition de « graphes syntaxiques profonds », conçus comme niveau de représentation intermédiaire entre syntaxe et sémantique (Candito, 2022).

Pour faciliter la lisibilité et la maintenance de la cohérence globale, les règles sont regroupées en modules, chaque module comprend un ensemble de règles. L'ordre

d'application des règles au sein d'un module est libre alors que l'ordre entre modules peut être contraint. Le système actuel comprend 11 règles organisées en quatre modules.

5.1.4. La visualisation de graphe DMRS

Le composant « DMRS *Graph Visualisation* » qui utilise l'outil Delphin-Latex[41] prend en entrée une représentation exprimée en DMRS et visualise le graphe DMRS associé. Cet outil est utile pour le développement des stratégies de simplification puis des règles de simplification GREW les implémentant. Il permet de visualiser la représentation DMRS avant et après simplification. Notons qu'il existe une démo Web DELPH-IN, Delphin-viz Demo[42] produisant des graphiques MRS et DMRS directement à partir de la sortie ACE.

5.1.5. La génération de texte à partir de graphes DMRS transformés

Le composant « *TEXT GENERATION* », qui utilise l'outil ACE[43] (*Answer Constraint Engine*) développée par le Consortium DELPH-IN, permet à partir d'un graphe DMRS transformé par des règles GREW de générer une phrase en format texte. Cette génération sera présentée dans une prochaine section spécifique.
Dans ce qui suit nous mettrons plus particulièrement l'accent sur le composant « *SIMPLIFICATION* » et sur les bases de règles GREW implémentant les stratégies associées à des constructions spécifiques.

5.2. Les règles de transformation de graphes DMRS

Deux types de relations sont les plus utilisés comme liens déclencheurs dans le développement de nos règles. Il s'agit des étiquettes ARG*/EQ et ARG*/NEQ :

- ARG*/EQ : un cas spécial qui indique l'identité de la portée.
  - Les adjectifs gouvernent les noms, comme _little_a_1 → _boy_n_1 de « *the little boy* » dans la Figure 33.

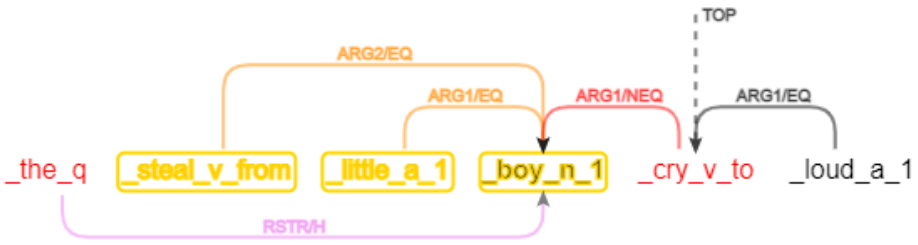


Figure 33 - DMRS de la phrase « The stolen little boy cries loudly ».

[41] https://github.com/delph-in/delphin-latex
[42] http://delph-in.github.io/delphin-viz/demo/
[43] http://sweaglesw.org/linguistics/ace/

- Les adverbes régissent les verbes, tels que _loud_a_1 → _cry_v_1 de « […] cries loudly ».
- Les prépositions gouvernent et s'attachent aux verbes, comme _with_p → _live_v_1 de « *Mary is the girl John lives with* ».
- Les verbes passifs régissent et modifient les phrases, telles que _steal_v_1 → _boy_n_1 de « *the stolen boy* » de la Figure 33.
- Dans une proposition relative :
    - Les verbes régissent et modifient les phrases qui sont représentées par des pronoms relatifs, comme _chase_v_1 → _cat_n_1 de « *the cat that chase…* » dans la Figure 28a.
    - Les prépositions gouvernent les phrases qui sont représentées par des pronoms relatifs, comme _with_p → _girl_n_1 de « *the girl John lives with »*.

- ARG*/NEQ : le cas le plus général qui sous-spécifie les portées.

- Les verbes régissent les noms (sujets, objets, objets indirects, etc.), tels que _like_v_1 → named ("John") et _like_v_1 → _cat_n_1 de « *John likes cats and Mary likes dogs* » dans la Figure 21c.
- Certaines relations de construction grammaticales régissent leurs arguments, comme appos → named(“Florence”) et appos → _city_n_1 de « Florence, the city » dans la Figure 19a.

Dans la suite de cette section, nous présentons les bases de règles de transformation de graphes DMRS en GREW. Dans ce chapitre nous ne détaillons que la base de règles GREW pour le traitement de la simplification de la construction présentant des appositions. Les bases de règles associées aux autres constructions se trouvent dans l’Annexe B.

5.2.1. Règles GREW pour les appositions

Comme mentionné dans le chapitre précédent (section 4.3.2), dans DMRS, l’apposition est identifiée par une relation de type *appos* qui prend les deux noms adjacents comme arguments. Pour tous les cas d’apposition (en début, milieu et fin de phrase), une seule base de règles est développée. La stratégie ou procédure est la suivante : il s’agit de réécrire l’antécédent de l’apposition avant le verbe et de transformer l’apposition en phrase indépendante afin de construire deux nouvelles DMRS et par suite deux nouvelles phrases indépendantes. Ainsi, d'une manière plus générale, la règle du découpage d'apposition supprime d'abord le nœud *appos* ainsi que le son ARG2 pour former une première DMRS, puis il construit l'autre DMRS en remplaçant l'ARG1 d’*appos* par son ARG2 et en supprimant ensuite *appos* et son ARG1.

Comme vu au chapitre 4, cette stratégie de simplification d’appositions est la suivante (cf. 4.3.2) :

La stratégie de simplification des appositions est la suivante :

1) Supprimer le nœud R de la relation d'apposition ;

2) Copier la composante connexe [l'antécédant (N) de la relation d'apposition+le nœud *proper_q*] avant le nœud *proper_q*, d'une façon à avoir deux composantes connexes qui se suivent. Il s'agit de copier les nœuds avec tous les traits morpho-syntaxiques, ainsi que les liens entre les éléments de cette composante connexe.

3) Rajouter le verbe être V après le nœud N1 (la copie du nœud N *Florence*). Pour le verbe être, nous avons décidé qu'il soit toujours au présent de l'indicatif comme il s'agit d'une définition (vérité générale) dans la plupart des cas. Il s'agit de définir les traits morpho-syntaxiques du verbe être (le temps, l'aspect, le mode, etc.).

4) Rajouter les arêtes entre V (le verbe être) et le nœud N1 et M (l'apposition au départ). L'ARG1 de V est N (*Florence*) et l'ARG2 est M (*city*).

De façon plus précise, nous cherchons un graphe qui contient une relation d'apposition « *appos* », dont l'ARG2 est un nom commun, avec une relation ARG1/NEQ entre le nœud R (*appos*) et l'antécédant N d'une part et une relation ARG2/NEQ entre le nœud R et l'apposition. L'ARG2 ne doit pas être un nom propre. La Figure 34 décrit la formalisation de la stratégie de simplification d'une relation d'apposition, pour les différents cas, en pseudo-code.

```
if
pattern {
  R avec R is appos
  N is composante connexe ARG1
  M is ARG2 of appos }

commands {
  Del_node R ;
  N1 is copy of N; Copy N1 :> N conserving edge and node label properties ;
  Add V > N1 with V is verbe to in present tense ;
  Add edges between the added V verb to be and N1 and V and N2 }
```

Figure 34 - Formalisation de la stratégie de simplification des appositions

Cette stratégie est implémentée par la base de règles GREW (Figure 35).

```
%% On cherche d'abord le graphe qui contient une apposition

rule select_sujet {

  pattern {
    R[gpred = "appos"];
    m: R - [ARG1: NEQ] - > N;
    p: R - [ARG2: NEQ] - > M;
    M[pos = "n"] }

  commands {
    N.select = yes;
    del_edge m;
    N.subject = yes }}


%% Ensuite, il faut déterminer la composante connexe de l'antécédant dans le but de la
copier

package propagate {
  rule down {
      pattern { M[select]; Z1[!select]; M -> Z1 }
      commands { Z1.select = yes }}

  rule up {
      pattern { M[select]; Z2[!select]; Z2 -> M; Z2<M }
      commands { Z2.select = yes }} }


rule right_limit {
      pattern {LR_selected[select=yes];LR[!select];LR > LR_selected;}
      without {X[select=yes]; LR_selected < X;}
      commands {LR.last=yes;}}

%% Copier les nœuds de la composante connexe

rule copy_node{
      pattern {S[select=yes];LR[last=yes];}
      without {X[select=yes];S > X;}
      commands{del_feat S.select;add_node S_copy:<LR;append_feats S ==> S_copy; del_feat
S.subject;add_edge S -[copy:yes]-> S_copy}}

%% Copier les arêtes entre les nœuds de la composante connexe

rule copy_edge{
      pattern {A1 -[copy:yes]-> A2;B1 -[copy:yes]-> B2;e: A1 -> B1}
      without {f: A2 -> B2; e.label = f.label}
      commands {add_edge f: A2 -> B2; f.label = e.label;}}


rule del_copy_edge{
      pattern { c: A1 -[copy:yes]-> A2 }
      commands { del_edge c }}


%% Rajouter le verbe être au présent de l'indicatif
rule add_verb{
      pattern { S [subject=yes];R [gpred=appos];m: R-[ARG2:NEQ]-> M; F[last=yes]}
      without {S2[subject=yes];S > S2}
```

```
      commands{ add_node V :> S; V.lemma=be; V.pos="v"; V.sense=id; V.SF=prop;
V.TENSE=pres; V.MOOD=indicative; V.cvarsort=e; add_edge V -[ARG1:NEQ]-> S; del_feat
S.subject; del_feat F.last; add_edge V -[ARG2:NEQ]-> M;}}


%% Supprimer le noeud de l'apposition

rule del_appos{
    pattern { R [gpred=appos];}
    commands { del_node R; }}

%% Appliquer une stratégie : l'ordre d'application des règles

strat appos
{Seq (select_sujet,Onf(propagate),
right_limit,
Onf(copy_node),
Onf(copy_edge),
Onf(del_copy_edge),
add_verb,
del_appos,
)}
```

Figure 35 - Base de règles GREW pour le cas des appositions

5.2.2. Règles GREW pour les coordinations

Pour les coordinations, la stratégie de simplification des coordinations entre deux événements qui partagent le même sujet est la suivante (cf. 4.4.3) :

La stratégie de simplification des coordinations entre deux événements qui partagent le même sujet est la suivante :

– Si la conjonction est « *and* » :

1) Supprimer le nœud de la conjonction ;

2) Copier la composante connexe du sujet partagé avant le deuxième verbe V2. Il s'agit de copier les nœuds avec tous les traits morpho-syntaxiques, ainsi que les liens entre les éléments de cette composante connexe.

3) Supprimer tous les liens entre V2 et V1 et entre V2 et N1.

4) Rajouter les arêtes (ARG1) entre V2 et le sujet partagé reconstruit.

5) Si la conjonction est une conjonction autre que « *and* », il s'agit de suivre les mêmes étapes précédentes mais sans supprimer le nœud de la conjonction.

La stratégie de simplification des coordinations entre deux événements qui ne partagent pas le même sujet est la suivante :

– Si la conjonction est « *and* », il suffit de supprimer le nœud de cette conjonction et le lien MOD/EQ entre les deux verbes.
– Si la conjonction est une conjonction autre que « *and* », il s'agit de supprimer le lien entre la conjonction et le premier verbe et le lien entre les deux verbes.

La figure 36 est la formalisation de la stratégie de simplification des coordinations en pseudo-code.

```
if
    pattern {A ="and"; V1 ="v"; V2 ="v"; s2: V2-> S; A-> V1;
            A-> V2; V1-> S; V2-> S;}
    commands { add_node S1:>A; del_node A;
                        add_edge d1: D1->S1;
                        add_edge q2: V2->S1; del_edge s2     }}

else if
    pattern {A ="and"; V1 ="v"; V2 ="v"; s2: V2-> S; A-> V1;
            A-> V2; V1-> S; V2-> S2; e:V2-[MOD:EQ]->V1 }
    commands { del_edge e; del_node A}}

else if
    pattern { A ="but|so|nor|yet"; V1 ="v"; V2 ="v"; s2: V2-> S;
            A-> V1; A-> V2; V1-> S; V2-> S2; e:V2-[MOD:EQ]->V1 }
    commands { del_edge e }}
```

Figure 36 – Formalisation de la stratégie de simplification des coordinations

5.2.3. Règles GREW pour les subordinations

Nous cherchons dans un graphe une relation de subordination « *subord* » dont l'ARG1 et l'ARG2 sont deux verbes. Une fois ce graphe est recherché, il s'agit d'appliquer un ensemble de règles de transformation.

La stratégie de simplification des subordinations est la suivante :

1) Supprimer le nœud de la relation de subordination ;
2) Copier la composante connexe de l'antécédant (ARG1 du verbe de la principale) avant le verbe de la subordonnée.
3) Modifier le temps du verbe en subordonnée en lui affectant les mêmes traits morpho-syntaxiques du verbe de la principale.
4) Rajouter les arêtes entre le verbe de la subordination et le sujet copié.

La figure 37 est la formalisation de la stratégie de simplification des subordinations en pseudo-code. La règle GREW est présentée dans l'annexe B.

```
pattern {
      A = "subord";
      V1 = v;
      V2 = v;
      n1: A-[ARG1]->V1;
      n2: A-[ARG2]->V2;
      s1: V1-[ARG1:NEQ]->S; }

commands {
      V2.PROG=V1.PROG; V2.TENSE=V1.TENSE;
      del_node A ;
      add_node S2:<V2; conserving edge and node label properties;
      add_edge V2-[ARG1:NEQ]->S2    }
```

Figure 37 - Formalisation de la stratégie de simplification des subordinations

5.2.4. Règles GREW pour les relatives

Il s'agit d'abord de définir le graphe sur lequel il faut appliquer l'ensemble de règles. D'où, la stratégie est la suivante : nous cherchons un nœud qui doit être un nom, suivi d'un verbe et une relation ARG1/EQ partant du verbe vers le nom.
Une fois ce graphe est recherché, il s'agit d'appliquer un ensemble de règles de transformation.

La stratégie de simplification des relatives est la suivante :

1) Supprimer l'arête entre le verbe et l'antécédant.

2) Copier la composante connexe de l'antécédant avant le verbe de la relative d'une façon à avoir deux antécédant qui se suivent puis le verbe.

3) Rajouter une arête ARG1/NEQ entre l'antécédant copié et le verbe.

La figure 38 est la formalisation de la stratégie de simplification des relatives. La règle GREW est présentée dans l'annexe B.

```
pattern {

      V1; V2[pos=v]; V1-[ARG2:NEQ]->S; a: V2-[ARG1:EQ]->S ; V1<<S; V2>>S; }

commands {
      del_edge a
      add_node S2:<V2; conserving edge and node label properties
      add_edge V2-[ARG1:NEQ]->S2; }
```

Figure 38 - Formalisation de la stratégie de simplification des relatives

5.2.5. Règles GREW pour la voix passive

Nous cherchons le graphe suivant : un verbe dont son ARG2 (l'objet) précède son ARG1 (le sujet).

La stratégie de transformation des voix passives permet de permuter les deux arguments du verbe.

La figure 39 est la formalisation de la stratégie de simplification des relatives. La règle GREW est présentée dans l'annexe B.

```
pattern {
        V [pos="v"];
        n1: V-[ARG2:NEQ]-> N1;
        n2: V-[ARG1:NEQ]->N2;
        N1<<V;
        N2>>V}

commands {
        unorder N1; insert N1:>V; unorder N2 ; insert N2:<V }
```

Figure 39 - Formalisation de la stratégie de transformation de la voix passive en voix active

5.3. Génération

En fonction du type de représentation en entrée, la génération de phrases peut être classée en deux catégories : 1) la génération de texte à partir de représentation de sens, et 2) la génération à partir de texte.

Génération à partir de représentation de sens vise à générer des phrases à partir de représentations syntaxiques, sémantiques, etc. Certaines approches de cette catégorie s'appuient ainsi sur des bases de données ou de connaissances (des ontologies par exemple) en entrée, tandis que d'autres exploitent des représentations sémantiques plus ou moins profondes (par exemple, les F-structures, les formules de logique du premier ordre, les formules de récurrence sémantique minimales, des arbres de dépendance).
En revanche, la tâche de *génération à partir de texte* transforme un texte en un autre texte en langage naturel.

Dans cette thèse, nous avons utilisé ACE, un processeur pour les grammaires DELPH-IN, prenant en charge à la fois *l'analyse syntaxique* (production du graphe DMRS d'une phrase) et la *génération* (production d'une phrase à partir d'un graphe DMRS). A partir des représentations DMRS des phrases du corpus transformées par les règles GREW, le générateur génère le texte associé à chaque phrase et place chaque phrase générée dans l'ordre du corpus d'origine.

Les processeurs ERG tels qu'ACE utilisent l'algorithme de génération de diagrammes exploitant la nature bidirectionnelle de la grammaire. La grammaire est une composante essentielle de la génération, indépendamment du fait qu'elle soit statistique ou symbolique. Les approches statistiques, combinent l'utilisation de règles écrites à la main avec un modèle de langage (Langkilde and Knight, 1998 ; Langkilde, 2002), utilisent des grammaires probabilistes comme TAG (Bangalore and Rambow, 2000), CCG (White and Rajkumar, 2009), CFG (Konstas and Lapata, 2012) et HPSG (Nakanishi et al., 2005). Une autre méthode consiste à apprendre une grammaire de génération automatiquement en s'aidant d'un corpus (Zhong and Stent, 2005). De même, les approches symboliques utilisent des grammaires symboliques orientées génération (Elhadad and Robin, 1996 ; Lavoie and Rambow, 1997) ou s'appuient sur des grammaires génériques réversibles comme TAG (Koller and Striegnitz, 2002), CCG (White, 2004), HPSG (Carroll and Oepen, 2005).

Dans ACE, la première étape consiste à rechercher des prédicats dans le lexique de la grammaire, ainsi que toutes les arêtes potentielles auxquelles ils peuvent participer, sur la base des règles de grammaire. Le générateur tente alors de créer une structure syntaxique dont la MRS correspondant correspond au MRS d'entrée. Si l'entrée ne peut pas être liée aux règles de grammaire existantes, la génération échoue. Ce composant est basé sur l'outil ACE que nous utilisions déjà pour l'analyse sémantique. Ainsi, le pipeline d'analyse-transformation-génération est totalement automatique.

Cependant, ce processus automatique produit généralement trop de sorties. Les raisons sont doubles : premièrement, ACE et ERG « *over-parse* ». C'est-à-dire que parfois, même une phrase simple a des dizaines de structures MRS analysées en raison de règles d'unification substantielles. Deuxièmement, ACE et ERG surgénèrent. Parfois, même une simple structure MRS a des dizaines de réalisations également en raison de règles d'unification substantielles.

Ainsi, même si les 5 meilleures structures MRS sont extraites d'une phrase analysée et que les 5 meilleures réalisations sont extraites de chaque structure MRS, une seule phrase peut toujours avoir 25 réalisations (y compris les doublons). Par exemple, pour la phrase « *This is the decision that John took* », le parseur donne 5 résultats. Pour un résultat d'analyse, on obtient 5 choix :

a) This is the decision that, John took.
b) This is the decision which, John took.
c) This is the decision who, John took.
d) This is the decision, which, John took.
e) This is the decision which John took.

Pour faire face à ce problème, le premier résultat d'analyse et de génération est choisi. Le générateur ACE ne dispose pas actuellement d'un mécanisme pour faire face aux prédicats inconnus, mais ils peuvent être analysés et on peut leur attribuer une étiquette de partie du discours.

Dans les expériences de départ, chaque prédicat inconnu était remplacé par un prédicat connu avec la même étiquette de partie du discours. Ensuite, la forme de surface

connue du prédicat de substitution est remplacée dans la chaîne de réalisation par la forme de surface du prédicat inconnu d'origine. Par exemple, tout prédicat de nom singulier inconnu est converti en un prédicat de substitution *_chocolate_n_1*. Ce choix n'était pas idéal pour nous, surtout pour les verbes comme il y a toujours des différences de modes et d'aspect. C'est pourquoi nous sommes intervenus dans la base lexicale d'ACE dans le but de l'enrichir.

La représentation du sens considérée dans cette thèse est la DMRS. DMRS est une représentation sémantique fondée sur le calcul des prédicats. Cependant, puisque la représentation est obtenue de manière compositionnelle via l'analyse HPSG, elle reste quelque peu liée à la syntaxe de la phrase. Pour cette raison, une représentation DMRS prédétermine ses réalisations en langage naturel plus que, par exemple, AMR. En raison de sa dépendance à la grammaire, l'approche établie tend à produire des sorties grammaticales et fluides qui sont garanties de préserver le sens par rapport à la représentation du sens.

5.4. Stratégie d'application des règles

Pour les 5 opérations de simplification syntaxiques traitées dans le cadre de cette thèse, nous disposons pour chacune d'une base de règles GREW. Disposant de 5 bases ou paquets de règles, il faut décider dans quel ordre les appliquer afin de prendre en compte la question de la cohésion du texte. Si une phrase contient deux ou plusieurs phénomènes syntaxiques complexes, il faut définir lequel doit être simplifié en premier. L'ordre d'application des règles peut minimiser les erreurs d'analyseur et influence ainsi la qualité de la phrase simplifiée finale (Gasperin et al., 2010).

Nous nous sommes basés sur le travail de Gasperin et al., (2010) pour définir cet ordre. Le but de leur travail est de rechercher quel ordre peut causer le moins de problèmes à l'analyseur (syntaxique dans leur cas), augmentant ainsi les chances que ses analyses résultantes soient correctes, principalement en ce qui concerne la segmentation des phrases et des propositions. Les auteurs présentent deux ordres dans lesquels appliquer les règles :

- Le premier ordre, appelé « *empirique* », a été défini empiriquement en fonction de la qualité de la sortie de l'analyseur pour le seul phénomène impliqué : plus la sortie de l'analyseur est fiable, plus une règle est avancée dans la séquence des opérations de simplification. L'ordre empirique est : 1. Voix passive ; 2. Apposition ; 3. subordonnées ; 4. relatives non-restrictives ; 5. relatives restrictives ; 6. Coordinations ; 7. Non SVO.
- Le deuxième ordre, appelé « *hiérarchique* », suit l'ordre dans lequel les phénomènes syntaxiques apparaissent dans l'arbre résultant de l'analyse syntaxique de la phrase. Plus le phénomène est proche de la racine de l'arbre, plus tôt il va être simplifié. C'est l'ordre suivi par Siddharthan (2006). Les auteurs montrent que l'ordre « *hiérarchique* », facilite l'analyse des segments suivants, améliorant la qualité de la simplification.

En effet, lors du développement du système au départ, nous avons défini un ordre d'application des règles semblable à l'ordre empirique ci-dessus. Cependant, nous avons trouvé que l'ordre d'application des règles n'a aucun effet sur les résultats, ni sur la cohésion des phrases. GREW effectue des transformations de graphes dans le but d'avoir deux (ou plusieurs) sous-graphes. Si le système traite les coordinations avant les subordinations ou vice versa, dans les deux cas, on aura les mêmes sous-graphes.
Considérons l'exemple (54) suivant contenant une apposition, une coordination et une subordination. La figure 40 est la représentation DMRS de cette phrase.

(54) He visited me in Paris, the city of love, and he sat on the carpet, contemplating me.

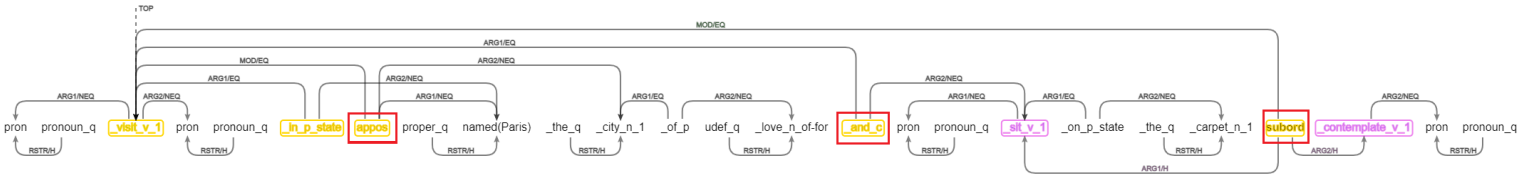


Figure 40 - DMRS de la phrase 54

Si on supprime les nœuds « *appos* », « *and* » et « *subord* » (encadrés en rouge), nous obtiendrons 3 sous-graphes sémantiques. Quel que soit l'ordre d'application, on aura les mêmes sous-graphes de sortie relatifs aux fragments suivants (Figure 41) : He visited me in Paris | The city of love | he sat on the carpet | contemplating me.

Après complétion de phrases (réécriture du sujet partagé), chaque sous-graphe, correspondant à une phrase, est écrit dans un fichier séparé afin de le générer.

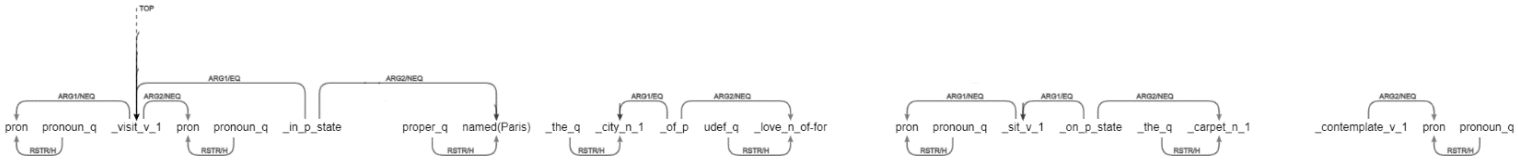


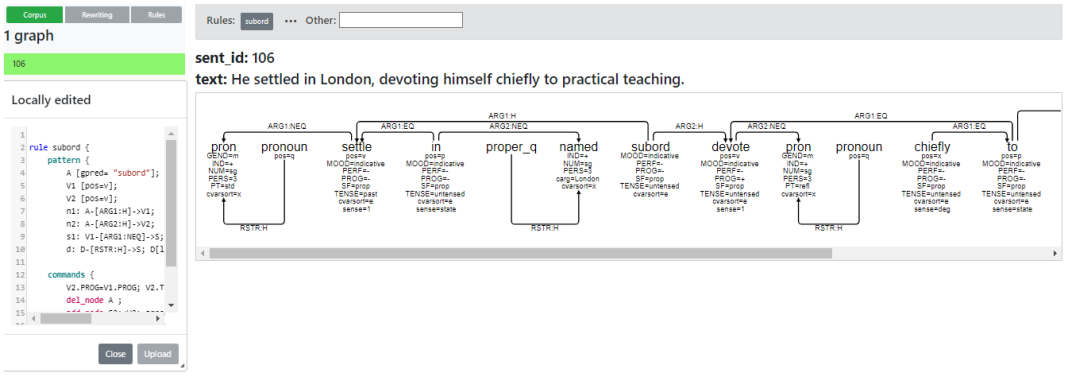
Figure 41 - DMRS de la phrase 54 en supprimant les nœuds déclencheurs et montrant 4 sous-graphes sémantiques.

5.5. Trace d'exécution de GRASS

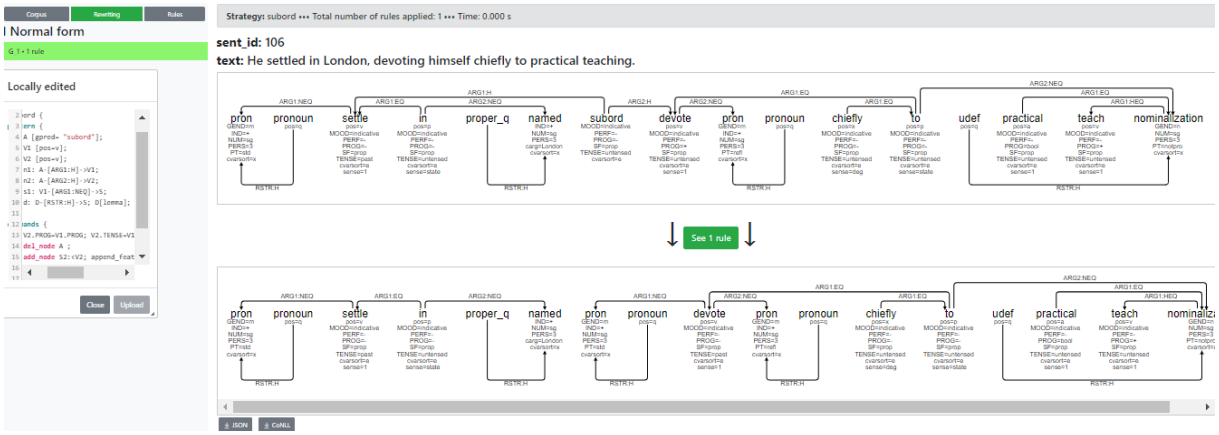
Considérons la phrase (55) ci-dessous. La Figure 42 est une capture d'écran de la plateforme GREW-Web[44]. Cette plateforme est interactive, elle permet ainsi de visualiser l'exécution des règles appliquées au fur et à mesure par notre système GRASS, ainsi que les transformations qui en résultent au niveau des graphes DMRS.

(55) He settled in London, devoting himself chiefly to practical teaching.

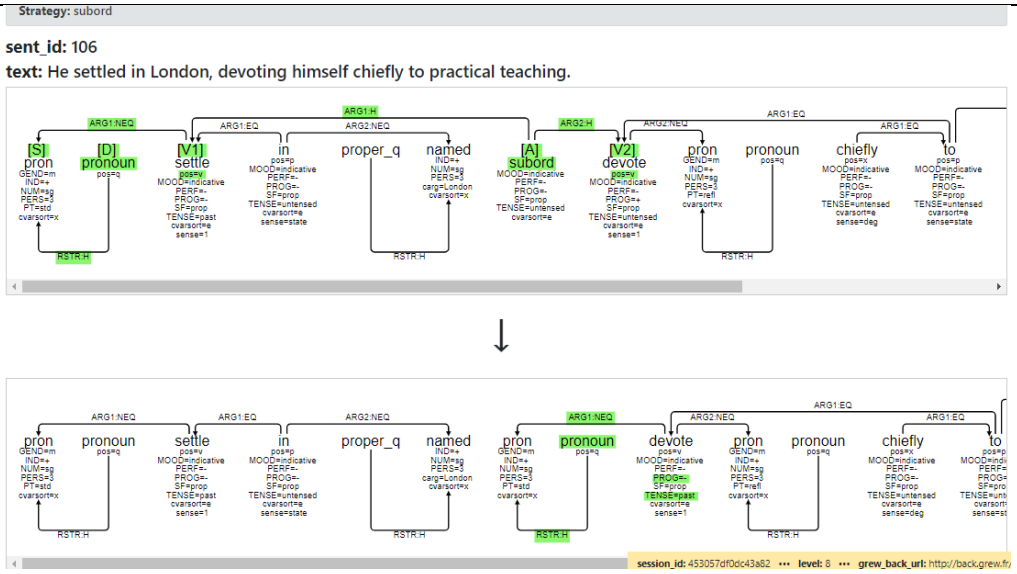
[44] http://transform.grew.fr/



a. Graphe de la phrase d'entrée avec la règle de transformation



b. Une règle est appliquée. Le deuxième graphe est le graphe transformé



c. Les nœuds et les arêtes surlignés en verts sont les ceux cherchés comme motifs (indicateurs déclencheurs) dans les règles

Figure 42 - Capture d'écran de la plateforme GREW

Ainsi, par cette trace d'exécution de GRASS, nous pouvons expliquer pas à pas comment les transformations des phrases à simplifier sont réalisées ce qui permet d'analyser les erreurs et leur origine dans le but d'améliorer le système par modification des règles de transformation.

Conclusion

Dans ce chapitre, nous avons présenté le système GRASS implémentant notre méthode de simplification syntaxique développé dans le Chapitre 4 précédent. Nous avons montré comment une représentation sémantique complexe comme DMRS peut être divisée en plusieurs sous-graphes, en utilisant un modèle de transformation de graphes basé sur un ensemble de règles élaborées manuellement dans le système de réécriture GREW. Les règles définissent des nœuds déclencheurs et des liens qui indiquent la présence de l'une des constructions grammaticales sélectionnées : *coordination, apposition, subordination, relative* et la *voix passive*. Chaque décision de segmentation est associée à un sous-graphe fonctionnel, qui stocke des informations sur les connexions entre les segments.

Comme nous l'avons déjà évoqué dans le chapitre précédent, l'approche que nous avons retenue dans cette recherche est une approche à base de règles qui n'utilise pas l'apprentissage automatique principalement basées sur l'apprentissage profond (*Deep Learning*) et conduisant à des systèmes de type « boites noires » incapables d'expliquer les transformations qu'ils réalisent sur les phrases à simplifier. Au contraire, notre méthode à base de règles, permet d'interpréter, d'expliquer comment les transformations des phrases à simplifier sont réalisées par le système, permettant ainsi d'analyser les erreurs et leur origine dans le but d'améliorer la méthode et son implémentation dans le système par modification des règles de transformation.

Dans le chapitre suivant, nous décrirons les expérimentations menées pour évaluer notre système GRASS, le corpus, les métriques d'évaluation automatique, le protocole d'évaluation humaine et les résultats par rapport aux autres systèmes existants.

6. Évaluation du système GRASS

Une étude critique sur l'évaluation des textes simplifiés a été menée par Grabar et Saggion (2022). Ils constatent que l'évaluation de la simplification reste peu étudiée : contrairement à d'autres tâches de TAL, comme la recherche et l'extraction d'informations ou les systèmes questions-réponses, qui attendent des sorties *factuelles* et *consensuelles* des systèmes, il est difficile de définir une sortie standard de simplification : (1) elle n'est pas *factuelle* car elle repose sur des séries de transformations plus ou moins gérées par des simplificateurs humains et des systèmes automatiques, et (2) elle n'est pas *consensuelle* car elle est fortement basée sur les connaissances propres à des personnes et parce que chacun a une opinion sur le résultat de la simplification. Par conséquent, plusieurs facteurs interviennent dans le processus de simplification et son évaluation comme le rôle des utilisateurs finaux, les données de référence et leur domaine, ainsi que les mesures d'évaluation utilisées.

Ce chapitre décrit les expérimentations pour évaluer GRASS (*GRAph-based semantic representation for Syntactic Simplification*), notre système de simplification syntaxique basé sur les représentations sémantiques en DMRS. Nous présentons dans la section 6.1 les corpus utilisés pour le développement de nos règles de simplification de textes, le corpus d'évaluation de GRASS, ainsi que le protocole d'évaluation. Dans la section 6.2, nous montrons les résultats de l'évaluation automatique tout en détaillant les métriques d'évaluation. La section 6.3 décrit le protocole mis en place pour l'évaluation manuelle des sorties du système, ainsi que les résultats. Une analyse d'erreurs est décrite dans la section 6.4. Nous clôturons cette section par une discussion des résultats obtenus, les avantages de l'approche utilisée, les limitations de notre système de SAT par rapport aux systèmes existants basés sur la sémantique et ceux à base de la syntaxe. Nous développons, enfin, quel est l'apport d'une méthode à base de représentations sémantiques pour un système de simplification syntaxique des textes.

6.1. Corpus et protocole d'évaluation

Afin d'évaluer les simplifications générées automatiquement, des travaux antérieurs ont comparé la simplification générée avec des simplifications de référence à l'aide de métriques automatiques. Dans notre travail, nous utilisons l'outil EASSE[45] (*Easier Automatic Sentence Simplification Evaluation*) (Alva-Manchego et al., 2019), un package Python qui donne accès à des métriques automatiques pour l'évaluation de la simplification des phrases et à des ensembles de données publiques via une interface de ligne de commande. Avec cet outil, les auteurs (1) fournissent des métriques automatiques dans un seul progiciel, (2) complètent ces métriques avec une analyse de

[45] https://github.com/feralvam/easse

transformation au niveau des mots et des fonctionnalités d'estimation de la qualité (*Quality Estimation* QE) sans corpus de référence, (3) fournissent un accès direct aux outils couramment utilisés ensembles de données d'évaluation, et (4) gênèrent un rapport HTML complet pour l'évaluation quantitative et qualitative d'un système de simplification des phrases.

EASSE donne accès à trois ensembles de données accessibles au public pour l'évaluation automatique de la simplification : PWKP (Zhu et al., 2010), TurkCorpus, appelé aussi WikiLarge (Xu et al., 2016) et HSplit (Sulem et al., 2018b). Tous sont constitués d'ensembles de données originales : des phrases extraites d'articles de Wikipédia en anglais (EW). EASSE peut également évaluer les sorties du système dans d'autres ensembles de données personnalisés fournis par l'utilisateur.

6.1.1. Corpus de développement

Nous avons développé nos règles de simplification en nous basant sur le corpus Newsela (Xu et al., 2015), une collection d'articles de presse avec des simplifications professionnelles en quatre niveaux. Zhang et Lapata (2017) ont aligné les paires de phrases correspondant aux niveaux 4-3, 3-2, 2-1 et 1-0 avec 94 208 paires de phrases d'entraînement, 1 129 de validation et 1 077 de test. Nous avons développé notre système de règles en utilisant les paires de phrases d'entraînement.

6.1.2. Corpus d'évaluation

Xu et al., (2016) ont proposé TurkCorpus[46], un jeu de données composé de 2 359 phrases complexes (2 000 de validation et 359 de test) extraites de Wikipédia où, pour chaque phrase complexe, 8 simplifications de référence ont été collectées à l'aide d'Amazon Mechanical Turk. La plupart des phrases simplifiées sont cependant très similaires à la phrase complexe avec seulement quelques simplifications lexicales ou suppressions de mots, c'est-à-dire qu'elles ne sont pas adaptées à l'évaluation de systèmes de simplification de phrases à part entière effectuant des divisions de phrases et des opérations de réécriture plus complexes.

Axé uniquement sur la division de phrases, l'ensemble d'évaluation HSplit[47] (Sulem et al., 2018b) a été créé en utilisant les mêmes 2 359 phrases complexes que TurkCorpus et fournit 4 références humaines par phrase source. Chaque référence a été créée en opérant uniquement la division de la phrase sur la phrase complexe d'origine. Il s'agit donc d'un ensemble de données pour l'évaluation du découpage des phrases, mais il ne se généralise pas à la simplification des phrases en général. Nous avons utilisé les 2 000 phrases de validation pour l'amélioration de nos règles de simplification et les 359 phrases pour l'évaluation de notre système.

[46] https://github.com/cocoxu/simplification/tree/master/data/turkcorpus
[47] https://github.com/eliorsulem/HSplit-corpus

6.1.3. Protocole d'évaluation

L'hypothèse de départ était d'étudier la capacité de DMRS dans la simplification syntaxique. Nous avons choisi de traiter essentiellement le **découpage** de phrases**.** Dans l'évaluation, le corpus total d'évaluation contient 359 phrases. Ce corpus contient des phrases syntaxiquement complexes où on trouve des constructions syntaxiques à découper, mais aussi il contient des phrases **syntaxiquement simples** (c'est-à-dire des phrases où l'ordre SVO est respecté, il n'y a pas d'appositions ou des propositions enchâssées, etc., par exemple : *It is not actually a true louse*). Ces phrases sont nombreuses dans le corpus. Ainsi, parmi les 359 phrases, on trouve **193 phrases qui ne contiennent aucune construction syntaxique complexe à traiter et qui sont simples syntaxiquement**. Une statistique de l'ensemble des phrases dans le corpus d'évaluation est présentée dans le Tableau 6 et par un arbre de décision (Figure 43).

Simples	Ne sont pas traitées	Non analysées par ACE	Non transformées par nos règles	Transformées	**Total**
193	42	26	7	91	**359**

Tableau 6 - Statistiques sur l'ensemble du corpus d'évaluation

- 193 phrases qui sont simples syntaxiquement
- 26 phrases qui n'ont pas été analysées par l'analyseur automatique ACE.
- 42 phrases qui sont complexes syntaxiquement et qui peuvent être découpées mais pour lesquelles nous n'avons pas dévelopé des règles scpécifiques pour les traiter. Il s'agit par exemple des cas de d'exemplifications.
  Exemple : *Nevertheless, Tagore emulated numerous styles, including craftwork from northern New Ireland, Haida carvings from the west coast of Canada (British Columbia), and woodcuts by Max Pechstein*.
- 7 phrases n'ont pas été traitées par notre système et pour lesquelles nous avons développé des règles.

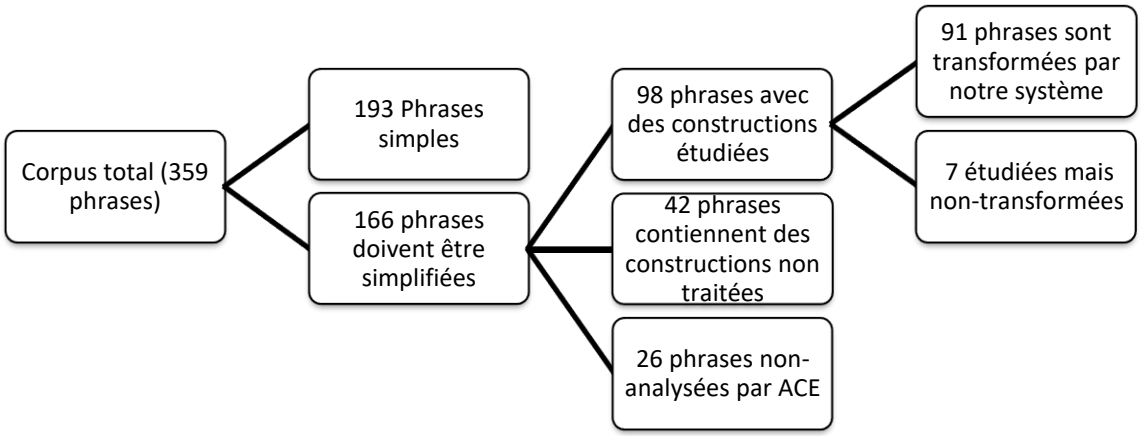


Figure 43 - Arbre de décision pour le passage de l'ensemble de 359 phrases au sous-ensemble de 91 phrases

Notre but était d'évaluer la **qualité** de ces 91 phrases, pour cela nous avons mené une évaluation automatique et humaine.

6.2. Évaluation automatique

Il existe quelques jeux de données de simplification avec des références (*gold standard)* qui peuvent être utilisées pour une évaluation. Il est courant d'utiliser des métriques de traduction automatique comme BLEU (Papineni et al., 2002), des métriques de simplicité comme SARI (Xu et al., 2016) et des métriques de lisibilité comme FKGL (Kincaid et al., 1975). La plupart de ces métriques sont disponibles dans des référentiels de code individuels, avec des exigences logicielles particulières qui diffèrent parfois même dans le langage de programmation (par exemple, SARI au niveau du corpus est implémenté en Java, tandis que SARI au niveau de la phrase est disponible à la fois en Java et en Python). D'autres métriques comme SAMSA (Sulem et al., 2018b) souffrent d'une documentation insuffisante ou nécessitent l'exécution de plusieurs scripts avec des codes difficiles, ce qui empêche les chercheurs de les utiliser (Martin, 2021).

L'outil EASSE fournit des métriques automatiques telles que BLEU, SARI, SAMSA, FKGL, ainsi que des scores de précision au niveau des mots et des fonctionnalités d'estimation de qualité telles que le taux de compression et le nombre moyen de mots ajoutés/supprimés. Dans la suite, nous présentons chacune des métriques automatiques et des fonctionnalités d'estimation de qualité.

6.2.1. Métriques automatiques

6.2.1.1. BLEU

Les systèmes de simplification sont souvent inspirés des modèles de traduction automatique, comme discuté dans les sections 2.2.3 et 2.3.2. La métrique automatique la plus couramment utilisée est *BiLingue Evaluation Understudy*, ou BLEU (Papineni et al., 2002). Au niveau de la phrase, BLEU compare la simplification générée avec des simplifications humaines de référence et peut être utilisé dans un cadre multi-référence. Il calcule d'abord les précisions de n-grammes du texte généré par rapport aux références de base, pour des longueurs de n-grammes de 1 à 4. Ensuite, ces précisions de n-grammes sont combinées en un score unique en utilisant une moyenne géométrique.

BLEU correspond généralement bien aux jugements humains sur la qualité de la traduction (Papineni et al., 2002), bien que cela ne soit pas vrai lorsqu'il existe de nombreuses traductions de qualité similaire ou de permutation de phrases (Callison-Burch et al., 2006). Il a toutefois été démontré que BLEU présente une faible corrélation avec les jugements humains sur la simplicité (Xu et al., 2016), mais aussi sur la préservation du sens et la fluidité, en particulier lorsque des opérations de réécriture telles que la division des phrases sont impliquées, conduisant à la notation incorrecte des phrases plus courtes et plus simples (Sulem et al., 2018b). Un autre problème de l'utilisation de BLEU en SAT réside dans le fait que cette métrique compte le nombre

de n-grammes communs entre le texte original et simplifié : par conséquent, plus le texte original est proche, meilleur est le score. BLEU donne des scores élevés aux phrases qui sont proches ou même identiques à l'entrée. Ainsi, le résultat élevé du BLEU est une conséquence du faible nombre de modifications apportées au texte (puisque les simplifications syntaxique et lexicale entraînent des modifications plus importantes du texte).

6.2.1.2. SARI

Xu et al. (2016) ont proposé SARI, une métrique d'évaluation pour la simplification de textes qui présente une meilleure corrélation avec les évaluations humaines. SARI s'appuie sur le fait que la simplification de textes est une tâche de réécriture monolingue et au lieu de comparer la simplification automatique uniquement par rapport aux références, il utilise également la phrase source pour une meilleure analyse de la réécriture effectuée. SARI compare la simplification prédite avec les références source et cible. Il s'agit d'une moyenne des scores F1 pour trois opérations sur les n-grammes : ajouts, conservations et suppressions. Pour chaque opération, ces scores sont ensuite moyennés pour tous les ordres de n-grammes (de 1 à 4) pour obtenir le score F1 global.
SARI peut être calculé comme suit :

$$SARI = (F1_{ADD} + F1_{KEEP} + P_{DEL}) / 3$$

Où
$F1_{ADD}$ est le score F1 n-gramme pour les opérations de rajout ;
$F1_{KEEP}$ est le score F1 n-gramme pour les opérations de conservation ;
P_{DEL} est le score de précision n-gramme pour les opérations de suppression.

Cardon et al. (2022) montrent que les sous-composants de SARI peuvent informer sur les opérations linguistiques présentes dans les références qui apparaissent dans une sortie du système, alors que cela est perdu lors de la moyenne des trois sous-composants en un seul score. En observant la corrélation entre la présence d'opérations spécifiques et les scores SARI globaux, on trouve peu de corrélation entre les deux. Cela signifie qu'aucune opération spécifique ne semble correspondre au score SARI global.

6.2.1.3. SAMSA

Sans s'appuyer sur des références, SAMSA (Sulem et al., 2018b) évalue la simplicité structurelle d'une simplification. Il fait des hypothèses fortes sur la façon dont les phrases doivent être simplifiées : (1) chaque phrase de sortie doit contenir un seul événement (ou appelé « scène » dans le formalisme sémantique UCCA ; (2) tous les événements sémantiques doivent être conservés entre la source et la simplification. Un système qui privilégie les découpages de phrases, obtiendra les scores les plus élevés. SAMSA crée d'abord une analyse sémantique de la phrase source et de la simplification en utilisant la représentation UCCA. Il aligne ensuite les tokens entre eux et note la simplification pour pénaliser les phrases qui contiennent plusieurs scènes UCCA, qui ont supprimé des scènes, ou dans lesquelles une seule scène est incorrectement divisée en plusieurs

phrases. SAMSA n'a pas été beaucoup utilisé en pratique depuis son introduction, probablement pour deux raisons (Martin, 2021). Premièrement, les hypothèses fortes de SAMSA sur la tâche de simplification le rendent incapable d'évaluer les simplifications lexicales. Deuxièmement, l'approche et l'implémentation de SAMSA le rendent très lent et lourd à utiliser, ce qui peut entraver son utilisation pratique. EASSE refactorise l'implémentation originale de SAMSA de façon à obtenir un score SAMSA avec une seule ligne de commande au lieu d'exécuter une série de scripts.

Les mesures de lisibilité, telles que *Flesch-Kincaid Grade Level* (FKGL), sont généralement signalées comme des mesures de simplicité. Cependant, elles ne s'appuient que sur la longueur moyenne des phrases et le nombre de syllabes par mot, de sorte que les phrases courtes obtiendraient de bons scores même si elles ne sont pas grammaticales ou ne préservent pas le sens (Wubben et al., 2012).

6.2.2. Analyse au niveau du mot et fonctionnalités d'estimation de la qualité

Les méthodes d'estimation de la qualité (*Quality Estimation* QE) ont été introduites pour la première fois dans le domaine de la traduction automatique pour mesurer la qualité d'un texte traduit automatiquement sans avoir besoin de traductions de référence (Bojar et al., 2017 ; Martins et al., 2017). Récemment, une approche QE a été proposée pour l'évaluation de résumé des textes (Xenouleas et al., 2019).

6.2.2.1. Analyse de transformations au niveau du mot

EASSE inclut des algorithmes pour déterminer quelles transformations un système de simplification effectue plus efficacement. Ceci est fait sur la base de l'alignement et de l'analyse au niveau du mot. Il s'agit d'utiliser des alignements de mots pour identifier les suppressions, les déplacements, les remplacements et les copies. Les auteurs génèrent deux ensembles d'annotations automatiques au niveau des mots : (1) entre les phrases originales et leurs simplifications de référence, et (2) entre les phrases originales et leurs simplifications automatiques produites par un système de simplification de phrases. En considérant (1) comme étiquettes de référence, le score F1 est calculé de chaque transformation en (2) pour estimer leur exactitude. Lorsqu'il existe plusieurs simplifications de référence, les scores F1 sont calculés par transformation de la sortie par rapport à chaque référence, puis le score le plus élevé est gardé comme score au niveau de la phrase. Les scores au niveau du corpus sont la moyenne des scores au niveau de la phrase.

6.2.2.2. Fonctionnalités d'estimation de la qualité

Les métriques automatiques traditionnelles utilisées pour la simplification des phrases reposent sur l'existence et la qualité des références, et ne sont souvent pas suffisantes pour analyser le processus complexe de simplification. L'estimation de la qualité exploite à la fois la phrase source et la simplification de sortie pour fournir des informations supplémentaires sur les transformations subies par un système de simplification et qui ne sont pas calculées par les métriques telles que SARI. Les fonctionnalités QE disponibles

sont : le *taux de compression* de la simplification par rapport à sa phrase source, sa *similarité Levenshtein*, le *nombre moyen de découpages* de phrases effectués par le système, la *proportion de correspondances exactes* (c'est-à-dire les phrases originales laissées intactes), la *proportion moyenne des mots ajoutés*, des mots *supprimés* et du score de *complexité lexicale*. Notons que la distance de Levenshtein est une mesure de la similarité du texte, définie en termes d'opérations de chaîne et liée au nombre minimum de modifications d'un seul caractère nécessaires pour faire correspondre deux chaînes. Il est égal à zéro lorsque les chaînes sont égales et ne peut être inférieur à la différence de leurs longueurs.

6.2.3. Résultats

Nous présentons les scores automatiques de notre système *GRAph-based semantic representation for Syntactic Simplification* (GRASS) dans le Tableau 7 sur les 91 phrases transformées. Nous nous comparons également à d'autres systèmes de l'état de l'art suivant : Hybrid et DSS sont deux systèmes qui utilisent les représentations sémantiques pour la simplification syntaxique. PBMT-R est un système à base de traduction automatique. TSM et RevILP sont deux systèmes qui utilisent les arbres syntaxiques. DRESS-LS et UNTS sont deux modèles neuronaux. Nous avons récupéré les résultats des systèmes TSM et RevILP en contactant les auteurs. Les résultats du DSS[48] sont récupérés à partir du site web de l'outil. Les résultats des autres systèmes sont récupérés à partir du travail d'Alva-Manchego et al. (2019)[49], où les auteurs présentent un outil visant à faciliter et à standardiser l'évaluation et la comparaison automatiques des systèmes de simplification de phrases. Ils présentent les résultats de différents systèmes sur plusieurs corpus, y compris TurkCorpus.

- **DSS** (Sulem et al., 2018a) : un système de simplification combinant des structures sémantiques et une traduction automatique neuronale. La division de phrases s'effectue avec *Direct Semantic Splitting* (DSS), un algorithme basé sur un analyseur sémantique de division de phrase en ses constituants sémantiques principaux en utilisant le formalisme sémantique *Universal Conceptual Cognitive Annotation* (UCCA).
- **Hybrid** (Narayan et Gardent, 2014) : une approche de la simplification de phrases qui associe une sémantique profonde et une traduction automatique monolingue. Cette approche est basée sur la sémantique, en ce sens qu'il prend en entrée une représentation sémantique profonde, *Discourse Representation Structure* (DRS). Elle combine un modèle de simplification pour le découpage et la suppression avec un modèle de traduction monolingue pour la substitution et la réorganisation de phrases.

[48] https://github.com/eliorsulem/simplification-acl2018

[49] https://github.com/feralvam/easse/tree/master/easse/resources/data/system_outputs/turkcorpus/test

- **PBMT-R** (Wubben et al., 2012) : modèle de simplification de phrases basé sur l'utilisation de la traduction automatique à partir des syntagmes (*Phrase-Based Machine Translation*).
- **TSM** (Zhu et al., 2010) : le modèle de simplification de phrases par transformation d'arbres syntaxiques TSM (*Tree-based Simplification Model*) s'appuie sur des techniques de traduction automatique statistique. Le modèle couvre intégralement la division, la suppression, la réorganisation et la substitution de phrases/mots.
- **RevILP** (Woodsend et Lapata, 2011) : un modèle de simplification de phrases capable de gérer les inadéquations structurelles et les opérations de réécriture complexes. L'approche est basée sur la grammaire quasi-synchrone (*Quasi-Synchronous Grammar QG*) (Smith et Eisner, 2006), un formalisme adapté à la réécriture de texte.
- **DRESS-LS** (Zhang et Lapata, 2017) : un modèle d'encodeur-décodeur couplé à un cadre d'apprentissage par renforcement profond.
- **UNTS** (Surya et al., 2018) : modèle neuronal non supervisé. Le noyau de la structure d'auto-encodage (non supervisée) est un encodeur partagé et une paire de décodeurs basés sur l'attention.

Deux exemples sont présentés dans le Tableau 8 avec les sorties de systèmes de comparaison.

Métriques	**GRASS**	**DSS**	**HYBRID**	**PBMT-R**	**TSM**	**RevILP**	**DRESS-LS**	**UNTS**
SAMSA	**51.44**	48.13	30.86	33.54	30.52	36.27	25.45	26.69
BLEU	***63.85***	*62.49*	*25.65*	*60.23*	*51.34*	*48.21*	*43.06*	*48.0*
SARI	***48.81***	*48.03*	*25.04*	*36.24*	*37.2*	*35.33*	*38.10*	*32.4*
Nb découpages	2.01	**2.53**	0.98	1.04	1.02	0.97	0.99	1.01
Exact copies	**0.0**	0.01	0.04	0.08	0.05	0.15	0.13	0.12

Tableau 7 - Résultats de l'évaluation sur les phrases du corpus de test d'HSplit ayant été transformées par GRASS (91 phrases)

EXEMPLE 1	
Phrase originale	*A matchbook is a small cardboard folder (matchcover) enclosing a quantity of matches and having a coarse striking surface on the exterior.*
GRASS (Hijazi et al 2022)	**A matchbook is a small cardboard folder (matchcover). A matchbook encloses a quantity of matches. A matchbook has a coarse striking surface on the exterior.**
HYBRID (Narayan et Gardent, 2014)	A matchbook is a cardboard folder (matchcover) enclosing a quantity of matches and having a coarse surface.
DSS (Sulem et al., 2018a)	a matchbook is a small cardboard folder matchcover. enclosing a quantity of matches. having a coarse striking surface on the exterior.
TSM (Zhu et al., 2010)	a matchbook is a small cardboard folder (matchcover) enclosing a quantity of matches, having a coarse striking surface on the exterior.
RevILP (Woodsend et Lapata, 2011)	a matchbook is a small cardboard folder (matchcover) enclosing a quantity of matches. having a coarse striking surface on the exterior.
PBMT-R (Wubben et al., 2012)	A matchbook is a small cardboard folder (matchcover) with a lot of games and has a coarse striking area on the outside.
DRESS-LS (Zhang et Lapata, 2017)	A matchbook is a small cardboard container.
UNTS (Surya et al., 2018)	A Matchbook is a small simple header (Matchcover) beneath a circular of matches and having a coarse striking surface on the floor.
EXEMPLE 2	
Phrase originale	*Aliteracy (sometimes spelled alliteracy) is the state of being able to read but being uninterested in doing so.*
GRASS (Hijazi et al 2022)	**Aliteracy is spelled alliteracy, sometimes. Aliteracy is the state of being able to read but being un-interested in doing.**
Hybrid (Narayan et Gardent, 2014)	Aliteracy (spelled alliteracy) is the state of being to read but being uninterested in doing.
DSS (Sulem et al., 2018a)	aliteracy sometimes spelled alliteracy is the state of being able to read. being uninterested in doing so. of being able to read.
TSM (Zhu et al., 2010)	Aliteracy (sometimes spelled alliteracy) is the state of being able to read but being uninterested in doing so.
RevILP (Woodsend et Lapata, 2011)	Aliteracy (sometimes spelled alliteracy) is the state of being able to read but being uninterested in doing so.

PBMT-R (Wubben et al., 2012)	Aliteracy (sometimes spelled alliteracy) is the state of being able to read but being want to do so .
DRESS-LS (Zhang et Lapata, 2017)	Aliteracy is the state of being able to read but being uninterested in doing so .
UNTS (Surya et al., 2018)	Aliteracy (sometimes spelled Alliteracy) is the state of being able to read but being concerned in doing so.

Tableau 8 - Sorties des différents systèmes pour deux phrases du corpus de test.

6.2.4. Discussion des résultats

Le Tableau 7 montre que GRASS surpasse les autres systèmes à base de sémantique en termes de BLEU, SARI et SAMSA pour les phrases de référence évaluées.

Plusieurs études ont montré que BLEU n'est pas adapté à la simplification de textes (Alva-Manchego et al., 2021 ; Sulem et al., 2018b ; Van den Bercken et al., 2019), car il n'est pas rare que les phrases sources (de Wikipédia) et les phrases de référence (de Simple Wikipédia) soient identiques ou très similaires, car les éditeurs de Wikipédia les ont juste copiées sans modifications ou avec des modifications mineures. Par conséquent, une simplification automatique qui conserve simplement la phrase source sans changement a souvent des scores BLEU élevés, mais n'est pas plus simple. Alva-Manchego et al. (2021) prouvent que BLEU est une mauvaise mesure pour estimer la simplicité structurelle dans les sorties du système où la division des phrases est privilégiée[50]. C'est pourquoi nous ne discuterons pas les résultats de BLEU.

Quant à SARI, il compare la sortie du système aux références et à la phrase d'entrée. Il se concentre sur la simplification lexicale (Sulem et al., 2018b ; Alva-Manchego et al., 2019 ; Cardon et Grabar, 2020b). Il mesure explicitement la qualité des mots qui sont ajoutés, supprimés et conservés par les systèmes. Nous ne discuterons non plus les résultats de SARI étant donné que notre système ne traite que la simplification syntaxique. Ce qui nous intéresse de ces métriques est le nombre de découpages effectués, ainsi que la mesure SAMSA.

En termes de découpage de phrases (*sentence splits*), DSS surpasse les autres systèmes. En examinant les sorties de DSS, on peut constater que la qualité des sorties est inférieure au nôtre. Le tableau 8 montre deux exemples de phrase de sortie pour chacun des systèmes évalués.

Pour l'exemple 1, DSS effectue le même nombre de découpage que GRASS (trois découpages) mais les phrases sont syntaxiquement incorrectes (phrases agrammaticales, sujet partagé non reconstruit). Pour l'exemple 2, DSS effectue plus de découpage que GRASS (trois découpages pour DSS contre deux découpages pour GRASS), dans ce cas DSS va obtenir un score SAMSA et un pourcentage de découpage plus élevés que GRASS. Mais les phrases obtenues avec DSS sont agrammaticales, moins simples et ne conservent pas

[50] "*BLEU is a bad metric to estimate Structural Simplicity in system outputs where sentence splitting was performed." (Alva-Manchego et al., 2021).*

le sens original de la phrase. C'est pourquoi, les mesures automatiques ne sont pas suffisantes pour évaluer un système de TAL et il est toujours indispensable de mener une évaluation humaine des sorties de n'importe quel système. Nous avons mené une campagne d'évaluation humaine pour les phrases transformées par notre système. La section suivante montre les résultats de l'évaluation humaine de ces phrases.

6.3. Évaluation humaine

Malgré les progrès récents des métriques d'évaluation automatique, il est indispensable de prendre en compte les jugements humains. En effet, certaines notions comme le degré de simplification ou le maintien ou non du sens original de la phrase sont assez *subjectives*. Grabar et Saggion (2022) trouvent que l'évaluation humaine se fait généralement avec des grilles, dans lesquelles 1 correspond à une qualité minimale et 5 à une qualité maximale. Cela dépend des annotateurs et de leur *perception* de la simplicité, la préservation du sens et la grammaticalité. De ce fait, la reproductibilité des annotations est faible, puisque les considérations peuvent varier d'un annotateur à l'autre. Il est donc intéressant d'évaluer l'accord inter-annotateur, c'est-à-dire, le degré d'accord entre des juges humains.

Nous avons mené une campagne d'évaluation humaine pour les 91 phrases transformées par notre système. Les jugements ont été recueillis selon trois dimensions (Todirascu et al., 2022) : l'adaptation (est-ce que la phrase simplifiée a le même sens que la phrase originale ?), la fluence (est-ce que la phrase simplifiée est grammaticale et correcte syntaxiquement ?) et la simplicité (est-ce que la phrase transformée est plus simple que l'originale ?).

La simplicité est la dimension la plus difficile à évaluer, parce qu'elle est très subjective, dans le sens où les annotateurs ont un degré de liberté d'interprétation qui peut varier significativement. En revanche, l'adaptation et la grammaticalité sont définies selon des règles fixes qui déterminent si une phrase est grammaticale et si les éléments sont conservés. Certains travaux ont tenté de rendre cette dimension plus quantifiable en demandant aux annotateurs de compter le nombre de réécritures trouvées dans la phrase plus simple (Xu et al., 2015), mais les articles les plus récents ont laissé le terme « plus simple » à l'interprétation des annotateurs (Zhang et Lapata, 2017 ; Jiang et al., 2020).

6.3.1. Protocole

Pour l'évaluation humaine, nous avons recruté **huit** locuteurs bilingues volontaires. Ces annotateurs ont été divisés en **quatre binômes**. Au total, notre système a simplifié 91 phrases. Chaque simplification est notée deux fois par les deux annotateurs du même groupe. Chaque groupe a annoté en moyenne 23 phrases. Le Tableau 9 est un exemple d'annotation de phrase.

Pour l'évaluation, les annotateurs ont analysé si la phrase est correctement découpée. Il s'agit de voir si le point de découpage est correct et si le sujet partagé est correctement reconstruit. Nous avons fourni un guide d'annotation aux annotateurs (Annexe A). Nous nous sommes inspirés du guide proposé dans le cadre du projet ALECTOR (Gala et al.,

2020a) et SimpleApprenant (Todirascu et al, 2019). L'analyse des erreurs consiste à parcourir chaque phrase et à mettre en évidence les erreurs qui sont ensuite catégorisées. Étant donné une phrase complexe et sa version simplifiée, les jugements sont généralement recueillis selon les trois dimensions mentionnées dans le guide. Ces dimensions sont évaluées sur une échelle de notation – l'échelle de Likert – qui a pour but de mesurer la qualité de la sortie. Nous avons choisi une échelle de 1 à 3, où le score le plus élevé indique la meilleure sortie.

1. **Adaptation**, sens préservé (*adequacy,* i.e. *meaning preservation*) de la sortie par rapport à l'original ; mesurer à quel point les phrases originales et simplifiées sont similaires en termes de sens. Le jugement sur la sémantique doit répondre à la question de savoir si la transformation effectuée préserve la sémantique originale de la phrase.
   1 = pas du tout le même sens (indépendamment de la grammaticalité de la phrase), sens complétement changé ; sujet partagé est mal reconstruit.
   2 = éléments sont différents ou absents / ambigüité ; pas toutes les constructions sont découpées.
   3 = exactement le même sens, parfaitement compréhensible (indépendamment de la grammaticalité de la phrase).

2. **Fluence**, i.e. grammaticalité. Le jugement sur la grammaticalité doit répondre à la question de savoir si la phrase reste grammaticale après les modifications effectuées (indépendamment du maintien du sens).
   1 = phrase agrammaticale.
   2 = nombre limité d'erreurs grammaticales.
   3 = phrase parfaitement grammaticale.

De plus, pour la propriété de grammaticalité, nous évaluons les 3 types d'erreurs :

- Les erreurs *orthographiques* : les erreurs liées à l'orthographe du mot.
- Les erreurs *grammaticales* : les erreurs de conjugaison de verbe,
- Les erreurs *syntaxiques* : erreurs liées à la ponctuation, Ordre des mots, absence d'un élément de la phrase, etc.

3. **Simplicité** linguistique. Il s'agit d'analyser si le découpage a rendu la phrase plus simple à comprendre, sans tenir compte des erreurs de grammaire, au regard des transformations mises en place.
   1 = sortie plus complexe que l'original (les transformations ont rendu la phrase plus difficile à comprendre).
   2 = sortie plus simple (une seule transformation qui a rendu la phrase plus accessible)
   3 = beaucoup plus simple que l'original (deux ou plusieurs transformations qui simplifient la phrase).

Le tableau 9 est un exemple.

Phrase Originale	Phrase simplifiée	Adaptation	Fluence	Simplicité
George Frideric Handel also served as Kapellmeister for George, Elector of Hanover (who eventually became George I of Great Britain).	Also, George Frideric Handel served as Kapellmeister, for George.	3	3	3
	Kapellmeister is Elector of Hanover.	1	3	1
	Kapellmeister became George I of Great Britain, eventually.	1	3	1

Tableau 9 - Exemple d'annotation de phrase.

6.3.2. Résultats

Pour chaque binôme, l'accord inter-annotateurs a été calculé, en utilisant la pondération quadratique κ de Cohen (Cohen, 1968). Puis la moyenne des κ de chaque mesure a été calculée. Les résultats (Tableau 10) montrent un accord satisfaisant.

	Adaptation	Fluence	Simplicité
κ de Cohen	0.658	0.611	0.626

Tableau 10 - Accord inter-annotateurs.

Comme nous n'avons pas les résultats de l'évaluation humaine pour tous les systèmes existants, nous nous sommes comparés aux systèmes à base des représentations sémantiques. Sulem et al., (2018a) ont mené une campagne d'évaluation humaine pour évaluer leur système et les sorties d'Hybrid. Les auteurs utilisent un protocole d'évaluation différent du nôtre. L'adaptation et la fluence sont mesurées à l'aide d'une échelle de 1 à 5. Une échelle de -2 à +2 est utilisée pour mesurer la simplicité, où un score de 0 indique que l'entrée et la sortie sont conservées. Chaque paire d'entrée-sortie est notée par 3 annotateurs par les auteurs. Nous avons sélectionné les scores de deux annotateurs. Pour l'évaluation, nous avons choisi le même sous-ensemble de phrases (les 91 phrases) dans les deux autres systèmes, DSS et Hybrid. Cependant, les auteurs ne fournissent les résultats d'évaluation humaine que pour les 70 premières phrases du corpus Hsplit (Turkcorpus) de 359 phrases. Parmi ces phrases, GRASS a transformé 23 phrases. Les phrases restantes (68 phrases) ont été filtrées des sorties de DSS et Hybrid et ont été annotées par nos annotateurs. Dans ce cas, nos annotateurs ont évalué les sorties de 91 phrases de GRASS et les 68 phrases de DSS et Hybrid (l'évaluation des 23 phrases est déjà fournies).

Comme l'échelle d'évaluation est différentes entre GRASS d'une part et DSS et Hybrid d'autre part et pour pouvoir se comparer, nous avons considéré pour l'adaptation et la fluence que le score 1 de l'échelle 1 à 5 correspond au score 1 de notre échelle de 1 à 3 ; les scores 2 et 3 correspondent au score 2 et les scores 4 et 5 correspondent au score 3

comme le montre le Tableau 11. Le Tableau 12 montre les résultats de l'évaluation en appliquant cette correspondance.

Echelle 1 à 5 utilisée pour évaluer l'adaptation et la fluence pour DSS et Hybrid	1	2	3	4	5
Echelle -2 à +2 utilisée pour évaluer la simplicité pour DSS et Hybrid	-2	-1	0	+1	+2
Echelle 1 à 3 correspondante utilisée pour évaluer GRASS	1	2		3	

Tableau 11 - Correspondance entre les deux échelles différentes

Systèmes	Adaptation	Fluence	Simplicité
GRASS (Hijazi et al 2022)	**2.91**	**2.84**	**2.89**
Hybrid (Narayan et Gardent, 2014)	2.69	2.41	2.20
DSS (Sulem et al., 2018a)	2.29	2.07	2.37

Tableau 12 - Résultats de l'évaluation humaine (échelle de Likert de 1 à 3).

6.3.3. Discussion des résultats

Les résultats du Tableau 12 montrent que notre système a été mieux évalué que les autres systèmes à base de représentations. Nous avons remarqué lors de l'évaluation que les systèmes Hybrid et DSS traitent essentiellement les propositions de coordination et relatives ; les formes passives, appositives et les subordinations ne sont pas traitées. GRASS couvre un plus large éventail de structures syntaxiques et cela est dû au choix du formalisme de représentation sémantique. DMRS est adapté aux tâches de compréhension du langage naturel : contrairement à UCCA, DMRS a une étiquette spécifique pour le nom propre ; ainsi, en génération, les noms propres sont reconnus, et la première lettre est en majuscule. DMRS donne des informations sur le mode et le temps du verbe, nos règles sont définies de manière à permettre de conjuguer le verbe au bon temps après séparation.

Enfin, si DSS effectue « plus » de divisions des phrases que les autres systèmes, cela ne signifie pas qu'il les divise « mieux ». DSS a un score élevé pour SAMSA et pour le nombre de division. Cependant, comme l'évaluation humaine le montre, la signification n'est pas toujours conservée et la sortie ne conserve pas l'ordre Sujet-Verbe-Objet (SVO). Le nombre important de découpage ne signifie pas que le système est plus performant, pourtant il est considéré comme tel suivant les métriques automatiques.

6.4. Analyse des erreurs

Pour mieux comprendre les aspects du processus de simplification qui constituent un défi à notre modèle, nous présentons les types d'erreurs les plus récurrents de notre ensemble de tests. Nous divisons les erreurs en deux grands types : 1) les erreurs liées à des problèmes techniques et relatifs aux outils comme le parseur, le générateur ou l'application des règles de transformations et 2) les erreurs linguistiques dans la sortie de notre système.

6.4.1. Erreurs techniques

1. Problèmes de couverture : des phrases qui doivent être simplifiées mais qui n'ont pas été transformées pour deux raisons :
    a) Nos règles ne sont pas appliquées.
        Exemple : Their eyes are quite small, and their visual acuity is poor.

    b) Structures que notre système ne traite pas.
    Exemple : Graham attended Wheaton College from 1939 to 1943, when he graduated with a BA in anthropology).

2. Des parties de la phrase sont supprimées. Ceci revient au générateur (des entités non-reconnues par le générateur).
    a) Complexe : The lawyer, Brandon (Waise Lee), became his idol, and MK Sun grew up to be a lawyer.
    b) Simple : MK Sun grew up to be a lawyer.

3. Application des règles dans les endroits où il ne faut pas. Dans l'exemple suivant, l'information entre parenthèses est reconnue comme une construction appositive par le parseur.
    a) Complexe : Dr. David Lindenmeyer (Australian National University) has argued that the need for nest boxes indicates that logging practices are not ecologically sustainable, for conserving hollow-dependent species like Leadbeater's possum.
    b) Simple : Doctor David Lindenmeyer is national Australian university. Doctor David Lindenmeyer has argued that the need for nest boxes indicates that logging practices are not ecologically sustainable for conserving hollow dependent species like Leadbeater's possum.

6.4.2. Erreurs linguistiques

1. **Erreurs orthographiques** dues au générateur.
    a) Complexe : By early on September 30, wind shear began to dramatically increase and a weakening trend began.
    b) Simple : By early on September thirtieth, wind shear began to increase dramatically. A *weakenning* trend began.

2. Erreurs grammaticales :

- Le sujet partagé est mal reconstruit :

a) Complexe : George Frideric Handel also served as Kapellmeister for George, Elector of Hanover (who eventually became George I of Great Britain).

b) Simple : Also, George Frideric Handel served as Kapellmeister, for George. *Kapellmeister* [Handel] is Elector of Hanover. *Kapellmeister* [Handel] became George I of Great Britain, eventually.

- Mauvaise ponctuation et ordre des mots :

a) Complexe : After the Sena dynasty, Dhaka was successively ruled by the Turkish and Afghan governors descending from the Delhi Sultanate before the arrival of the Mughals in 1608.

b) Simple : The governors Turkish, and Afghan, descending from the Delhi Sultanate, before the arrival of the Mughals after the Sena dynasty, ruled Dhaka, successively, in 1608.

6.5. Discussion

Une caractéristique clé de notre approche est qu'elle est basée sur la sémantique. En règle générale, les opérations de simplification telles que la division de phrases, la réorganisation de phrases, la génération d'expressions de référence et le choix de déterminants sont sémantiquement contraintes. Comme le soutient Siddharthan (2006), saisir correctement les interactions entre ces phénomènes est essentiel pour assurer la cohésion du texte.

Les approches de la simplification syntaxique basées uniquement sur la syntaxe sont confrontées aux problèmes d'identification du point de découpage, de reconstruction de l'élément partagé en phrases découpées et de la suppression des arguments. L'utilisation des informations sémantiques résout ces problèmes. Les représentations sémantiques des phrases donnent une idée claire des événements, facilitant ainsi à la fois l'identification des points de découpage possibles et la reconstruction des éléments partagés dans les phrases résultant du découpage (Narayan et Gardent, 2014 ; Sulem et al., 2018a).

Cependant, l'utilisation des informations sémantiques **uniquement** n'est pas suffisante pour capturer les constructions syntaxiques complexes. D'où la nécessité de combiner les avantages des deux annotations syntaxique et sémantique en utilisant un formalisme qui, comme DMRS, prend en considération et décrit à la fois les informations sémantiques et syntaxiques, au contraire des travaux cités qui utilisent des formalismes sémantiques portant uniquement sur les rôles thématiques. Dans l'exemple (56) ci-dessous, la phrase complexe n'a pas été traitée par aucun système de simplification à base de représentations sémantiques mais traitée par notre système.

(56) Loèche harbours the installations of Onyx, the Swiss interception system for electronic intelligence gathering.

Cela s'explique par le fait que les constructions appositives sont traitées par DMRS qui s'appuie sur des analyses syntaxiques, de sorte qu'il existe un fort chevauchement entre les constituants sémantiques et syntaxiques. Sa sémantique est ancrée dans la forme superficielle des phrases et dans les liens syntaxiques entre les constituants. En addition, elle couvre des informations concernant la coréférence et les informations spatiales, temporelles et discursives.

Conclusion

Dans ce chapitre, après avoir présenté les corpus utilisés pour le développement de nos règles de simplification de textes, ainsi que pour l'évaluation de GRASS, nous avons décrit le processus d'évaluation de notre système de simplification syntaxique basé sur les représentations sémantiques en DMRS.

L'évaluation automatique basée sur des métriques spécifiques (BLEU, SARI et SAMSA) a montré que notre système, et par là même notre méthode, surpasse d'autres systèmes de simplification syntaxique similaires existants, particulièrement sur la qualité des phrases découpées du TurkCorpus. Notre système offre également une meilleure couverture des constructions syntaxiques, tout en permettant une interprétabilité des transformations syntaxiques réalisées contrairement aux systèmes à base d'apprentissage de corpus parallèles.

La campagne d'évaluation humaine a été menée en se basant sur un guide d'annotation pour évaluer trois dimensions : *l'adaptation*, la *fluence* et la *simplicité*. Les résultats de cette évaluation montrent que les valeurs sont légèrement plus hautes pour notre système (échelle de Likert) par rapport aux deux autres systèmes basés également sur les représentations sémantiques. Ceci peut s'expliquer par le choix dans notre système de la notation DMRS qui prend en compte les informations à la fois sémantiques et syntaxiques.

Enfin, l'analyse des erreurs de notre système les associe principalement aux outils de génération et/ou de parsing liés à DMRS que nous avons utilisés (Le problème des erreurs en 'cascade' est bien connu pour les systèmes à base de règles).

L'évaluation des systèmes de simplification de texte est encore un domaine peu étudié. Il est fortement basé sur des approches d'évaluation exploitées dans d'autres tâches de TAL. Pourtant, la tâche de simplification est différente de ces autres tâches principalement parce que sa sortie est subjective. Dans les travaux futurs, il sera nécessaire d'approfondir l'évaluation de la simplification : proposer des principes plus clairs pour l'évaluation manuelle, définir une association plus forte entre les critères d'évaluation et les mesures, et proposer de nouvelles métriques d'évaluation plus flexibles. En outre, si les systèmes de simplification sont conçus pour une population cible spécifique, il semble indispensable que cette population soit impliquée dans le développement de la solution et l'évaluation des résultats (Grabar et Saggion, 2022).

Conclusion Générale

La simplification syntaxique vise à modifier la structure syntaxique d'un texte en supprimant les phénomènes syntaxiques complexes, tels que la coordination, la subordination, les relatives et la voix passive, sans en modifier le sens original. De nombreux travaux existants utilisent les arbres syntaxiques, d'autres, peu nombreux et plus récents, utilisent la sémantique. Enfin, des approches plus récentes exploitent par apprentissage de grands ensembles de données parallèles.

Les travaux basés essentiellement sur les arbres de dépendances syntaxiques, ont depuis plusieurs années montré leurs limites. Les travaux basés sur l'apprentissage de corpus parallèles abordent la simplification comme un cas de traduction monolingue, l'opération de découpage de la phrase n'a pas été abordée par ces systèmes, potentiellement en raison de la rareté de cette opération dans les corpus d'apprentissage. La qualité des simplifications fournies par cette approche est inférieure à celle fournie par les systèmes basés sur des règles, car les transformations syntaxiques complexes ne peuvent pas être capturées par des approches sans informations linguistiques.

En ce qui concerne les rares travaux sur la simplification à base de sémantique, ils présentent des limitations liées au fait qu'ils ne combinent pas sémantique et syntaxe. Ainsi l'utilisation des informations sémantiques *uniquement* n'est pas suffisante. Par exemple, en n'utilisant que de la sémantique, les constructions appositives et les subordinations ne peuvent pas être traitées.

Dans ce travail, nous avons tout d'abord proposé une méthode de simplification syntaxique des phrases en anglais utilisant une représentation sémantique des phrases prenant la forme de graphes de dépendances sémantiques et qui fournit une interface de connexion entre un langage sémantique et une structure de dépendance. Plus précisément notre méthode utilise *Dependency Minimal Recursion Semantic* (DMRS), une représentation sémantique profonde à base de graphes combinant sémantique et syntaxe. Le processus de simplification consiste à représenter la phrase complexe en un graphe DMRS, à le transformer selon un protocole spécifique en un ou plusieurs autre graphes DMRS représentant des phrases plus simples, et enfin la génération de ces graphes.

Pour évaluer cette méthode, nous avons développé un *système automatique de simplification syntaxique* implémentant notre méthode. Ce système, nommé GRASS, est basé sur des *règles de transformation de graphes* DMRS utilisant le système de réécriture de graphes GREW. La simplification syntaxique s'effectue en appliquant un ensemble de règles de transformation de graphes implémentées dans GREW. Le système prend en entrée la représentation sémantique, puis il applique un ensemble de règles de transformation de graphes afin d'obtenir deux (ou plusieurs) graphes DMRS. Les graphes transformés sont enfin générés afin d'obtenir des phrases simplifiées sans modifier leur sens. Nous avons montré que cette approche facilite la complétion (la réécriture de

l'élément partagé dans les phrases découpées) et couvre plus de constructions que les systèmes basés uniquement sur la sémantique.

Notons que l'approche que nous avons retenue dans cette recherche est une approche à base de règles qui n'utilise pas l'apprentissage automatique. Les approches à base d'apprentissage automatique nécessitent de gros corpus parallèles qui ne sont pas toujours disponibles, notamment pour la simplification de textes. De plus ces approches basées sur l'apprentissage profond (*Deep Learning*) conduisent à des systèmes de type « boites noires » incapables d'expliquer les transformations qu'ils réalisent sur les phrases à simplifier. Au contraire, notre méthode à base de règles et son implémentation dans le système GRASS, permet d'interpréter, d'expliquer comment les transformations des phrases à simplifier sont réalisées par le système, permettant ainsi d'analyser les erreurs et leur origine dans le but d'améliorer la méthode et son implémentation dans le système par modification des règles de transformation.

Sur un corpus de référence spécifique à la simplification de textes, le TurkCorpus, le système GRASS développé nous a permis de d'évaluer notre méthode linguistique, tout d'abord automatiquement en utilisant des métriques spécifiques (BLEU, SARI et SAMSA) et ensuite en ayant recours à des experts humains. Les évaluations automatiques et humaines obtenues par notre système surpassent les résultats obtenus par d'autres systèmes de simplification syntaxique existants sur ce même corpus de référence, en termes de grammaticalité et de couverture des constructions syntaxiques traitées. Les rares erreurs de notre système semblent liées à des faiblesses actuelles des outils DMRS utilisés encore en développement.

Les résultats de ces évaluations révèlent quelques limitations de cette méthode basée sur la sémantique.

1) Tout d'abord, cette évaluation a porté sur un corpus spécifique à la simplification syntaxique, le corpus TurkCorpus, et nous n'avons considéré que 359 phrases du ce corpus. Il nous apparaît nécessaire de valider notre méthode et notre système avec d'autres corpus plus conséquents.

2) Dans ce travail de thèse, nous n'avons traité que le découpage de phrase et le passage de la voix passive à la voix active. D'autres opérations de *réorganisation* ou de *substitution* liées à la simplification syntaxique n'ont pas été traitées comme la sémantique ne peut pas les capter. Le générateur utilisé effectue quelques opérations de réorganisation, notamment de déplacement des adverbes et des expressions spatio-temporelles, mais pas des déplacements au niveau des propositions. Pour les opérations de substitutions, nous avons besoin de ressources lexicales spécifiques que les outils mis en œuvre dans notre système ne disposent pas.

3) Notre système utilise divers outils de prétraitement, d'analyse, de réécriture de graphes et de génération principalement liés à la notation DMRS. Il y a ainsi des phrases qui ont été mal analysées/générées, c'est pourquoi elles n'ont pas été simplifiées. L'évolution de ces outils DMRS devrait permettre de réduire ces erreurs du fait que notre système dépend de ces outils.

Ce travail ouvre des perspectives à différents niveaux.

1) Au niveau **linguistique** :

- Notre méthode et le système GRASS la mettant en œuvre traitent de la division de phrase et la transformation de la voix passive en voix active. Il existe d'autres opérations de réorganisation et de substitution morpho-syntaxiques qui n'ont pas été traitées dans la thèse. Ces opérations, qui ne nécessitent pas le recours à de la sémantique exprimées dans DMRS, devraient cependant pouvoir être traitées dans notre méthode et réalisées dans GRASS avec des règles GREW spécifiques, nous pensons notamment au passage du discours indirect au discours direct.

2) Au niveau de l'**ingénierie linguistique :**

- Notre système simplifie des phrases isolées. Le système pourrait être amélioré dans le but de simplifier les phrases complexes au *niveau du texte* tout en préservant suffisamment d'informations, d'intégrité sémantique et de traiter les inférences dans le but de préserver la cohérence du texte pourrait être l'un des travaux futurs.

- Notre thèse se concentre exclusivement sur l'anglais, mais l'initiative DELPH-IN est à l'origine de grammaires dans plusieurs langues, dont le japonais[51] (Siegel et Bender), l'indonésien (Moeljadi et al., 2015) et le chinois (Fan, 2019).

3) Au niveau **expérimental** :

- Comme déjà évoqué, l'évaluation de notre système devrait se faire sur un corpus plus large que le TurkCorpus limité à 359 phrases, dans lequel seules 91 phrases sont concernées par les transformations définies dans GRASS. Nous espérons pouvoir évaluer notre système sur un corpus plus large, comme le corpus Newsela.

- Toujours concernant l'évaluation de notre système, celui-ci ne vise actuellement pas un public cible particulier. Evaluer les résultats de ce système avec un public cible particulier, comme par exemple les apprenants de l'anglais langue seconde (ALS), ou des lecteurs ayant des déficiences spécifiques, pourrait constituer des travaux futurs.

[51] https://github.com/delph-in/docs/wiki/JacyTop

4) Au niveau **intégration** dans d'autres **tâches de TAL plus complexes** :

- La simplification syntaxique réalisée par le système GRASS pourrait être *intégré* en pré ou post traitement, dans d'autres tâches plus complexes du TAL comme la traduction et le résumé automatiques. Notre système réaliserait ainsi une étape de pré-traitement de ces tâches. Il serait intéressant de tester cette intégration et d'évaluer son intérêt pour ces tâches complexes.

- Enfin l'approche et l'architecture logicielle du système GRASS à base de GREW pourraient être utilisées pour réaliser *d'autre tâches de TAL que la simplification*, comme par exemple le *paraphrase* voire le *résumé* des textes automatique. Des méthodes spécifiques à chacune de ces tâches, conduisant toujours à des transformations de représentations en DMRS, pourraient alors être définie, avec ses stratégies et règles de transformation spécifiques.

Références bibliographiques

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a treebank for French. In *Treebanks* (pp. 165-187). Springer, Dordrecht.

Abend, O., & Rappoport, A. (2013). Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 228-238).

Abney, S. P. (1991). Parsing by chunks. In *Principle-based parsing* (pp. 257-278). Springer, Dordrecht.

Agrawal, A., & Mukherjee, A. (2009). Question Answering using FrameNet. *Availablefrom: http://www. cse. iitk. ac. in/users/ashagr/webpage/courses/cs365_project_report. pdf [2009, April 19]*.

Aluísio, S., & Gasperin, C. (2010). Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas* (pp. 46-53).

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., & Fortes, R. P. (2008). Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering* (pp. 240-248).

Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C., & Specia, L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 295-305).

Alva-Manchego, F., Martin, L., Scarton, C., & Specia, L. (2019). EASSE: Easier automatic sentence simplification evaluation. *arXiv preprint arXiv:1908.04567*.

Alva-Manchego, F., Scarton, C., & Specia, L. (2021). The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, *47*(4), 861-889.

Alva-Manchego, F., Scarton, C., & Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, *46*(1), 135-187.

Alvin, L. P. (2014). The passive voice in scientific writing. The current norm in science journals. *Journal of Science Communication*, *13*(1), 1-16.

Babko-Malaya, O., Palmer, M., Xue, N., Joshi, A., & Kulick, S. (2004). Proposition bank ii: Delving deeper. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004* (pp. 17-23).

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Baldwin, T., Beavers, J., Bender, E. M., Flickinger, D., Kim, A., & Oepen, S. (2005). Beauty and the beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar—and the corpus. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 49-70.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... & Schneider, N. (2013, August). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178-186).

Bangalore, S., & Rambow, O. (2000). Exploiting a probabilistic hierarchical model for generation. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Barlacchi, G., & Tonelli, S. (2013). ERNESTA: A sentence simplification tool for children's stories in Italian. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 476-487). Springer, Berlin, Heidelberg.

Barzdins, G. (2014). FrameNet CNL: A knowledge representation and information extraction language. In *International Workshop on Controlled Natural Language* (pp. 90-101). Springer, Cham.

Barzilay, R., & McKeown, K. (2001, July). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics* (pp. 50-57).

Bastianelli, E., Castellucci, G., Croce, D., & Basili, R. (2013, November). Textual inference and meaning representation in human robot interaction. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora* (pp. 65-69).

Battistelli, D., Etienne, A., Rahman, R., Teissèdre, C., & Lecorvé, G. (2022). Une chaîne de traitement pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique (A Processing Chain to Explain the Complexity of Texts for Children From a Linguistic and Psycho-linguistic Point of View). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale* (pp. 236-246).

Beigman Klebanov, B., Knight, K., & Marcu, D. (2004). Text simplification for information-seeking applications. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 735-747). Springer, Berlin, Heidelberg.

Bender, E. M., Flickinger, D., Oepen, S., Packard, W., & Copestake, A. (2015). Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th international conference on Computational Semantics* (pp. 239-249).

Bentin, S., Deutsch, A., & Liberman, I. Y. (1990). Syntactic competence and reading ability in children. *Journal of Experimental Child Psychology*, *49*(1), 147-172.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1533-1544).

Berman, R. A. (1984). Syntactic components of the foreign language reading process. *Reading in a foreign language*, 139-156.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Billami, M. B. (2015). Désambiguïsation lexicale à base de connaissances par sélection distributionnelle et traits sémantiques. In *22ème conférence sur le traitement automatique des langues naturelles et 17ème rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues* (pp. 13-24).

Billami, M., François, T., & Gala, N. (2018). ReSyf: a French lexicon with ranked synonyms. In *27th International Conference on Computational Linguistics (COLING 2018)*.

Biran, O., Brody, S., & Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 496-501).

Birch, A., Abend, O., Bojar, O., & Haddow, B. (2016). HUME: Human UCCA-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030*.

Bitea, I. N. (1977). An attempt at defining apposition in modern English.

Blache, P. (2004). Property grammars: A fully constraint-based theory. In *International Workshop on Constraint Solving and Language Processing* (pp. 1-16). Springer, Berlin, Heidelberg.

Blache, P. (2010). Un modèle de caractérisation de la complexité syntaxique. In *Traitement Automatique des Langues Naturelles* (pp. 1-10).

Blanche-Benveniste, C. (2013a). Propositions pour progression dans la complexité syntaxique. *Revue Tranel (Travaux neuchâtelois de linguistique)*, *58*, 247-269.

Blanche-Benveniste, C. (2013b). Les difficultés syntaxiques et la lecture. *Revue Tranel (Travaux neuchâtelois de linguistique)*, *53*, 39-52.

Bloomfield, L. (1933). Language.. New York: Henry Holt and Company. *Bolinger, L. Dwight.(1957). Locus versus class*, 31-37.

Böhmová, A., Hajič, J., Hajičová, E., & Hladká, B. (2003). The Prague dependency treebank. In *Treebanks* (pp. 103-127). Springer, Dordrecht.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., ... & Turchi, M. (2017, September). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation* (pp. 169-214).

Bonfante, G., Guillaume, B., & Perrier, G. (2018). *Application de la réécriture de graphes au traitement automatique des langues* (Vol. 1). ISTE Group.

Bos, J. (1995). Predicate Logic Unplugged In *Proceedings 10th Amsterdam Colloquium*.

Bott, S., & Saggion, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, *48*(1), 93-120.

Bouamor, H. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues* (Doctoral dissertation, Université Paris Sud-Paris XI).

Boyer, J. Y. (1992). La lisibilité. *Revue française de pédagogie*, 5-14.

Bram, V. A. (1978). Sentence construction in scientific and engineering texts. *IEEE Transactions on Professional Communication*, (4), 162-164.

Bresnan, J. (2001). *Lexical-functional syntax*. Oxford: Blackwell.

Brinton, D. M., & Brinton, L. J. (2010). The linguistic structure of modern English. *The Linguistic Structure of Modern English*, 1-446.

Brouwers, L., Bernhard, D., Ligozat, A. L., & François, T. (2012). Simplification syntaxique de phrases pour le français (Syntactic Simplification for French Sentences)[in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN* (pp. 211-224).

Brouwers, L., Bernhard, D., Ligozat, A. L., & François, T. (2014). Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL 2014* (pp. 47-56).

Brunato, D., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2014). Defining an annotation scheme with a view to automatic text simplification. *Defining an annotation scheme with a view to automatic text simplification*, 87-92.

Brunato, D., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2015). Design and annotation of the first Italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop* (pp. 31-41).

Brunato, D., Cimino, A., Dell'Orletta, F., & Venturi, G. (2016). Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 351-361).

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, *104*(2), 215-226.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., & Pinkal, M. (2009). FrameNet for the semantic analysis of German: Annotation, representation and automation. *Multilingual FrameNets in Computational Lexicography: methods and applications*, *200*, 209-244.

Burton-Roberts, N. (1975). Nominal apposition. *Foundations of language*, *13*(3), 391-419.

Callison-Burch, C. (2007). *Paraphrasing and translation* (Doctoral dissertation, University of Edinburgh).

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *11th conference of the european chapter of the association for computational linguistics* (pp. 249-256).

Caplan, D. and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioural and Brain Sciences*, 22, 77–126.

Candito, M., Amsili, P., Barque, L., Benamara, F., De Chalendar, G., Djemaa, M., ... & Vieu, L. (2014). Developing a french framenet: Methodology and first results. In *LREC-The 9th edition of the Language Resources and Evaluation Conference*.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and language*, *3*(4), 572-582.

Cardon, R., Bibal, A., Wilkens, R., Alfter, D., Norré, M., Müller, A., Watrin, P., François T., (2022). Linguistic Corpus Annotation for Automatic Text Simplification Evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Cardon, R. (2018). Approche lexicale de la simplification automatique de textes médicaux. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN* (p. 159).

Cardon, R., & Grabar, N. (2019). Détection automatique de phrases parallèles dans un corpus biomédical comparable technique/simplifié (Automatic detection of parallel sentences in comparable biomedical corpora). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II: Articles courts* (pp. 255-264).

Cardon, R., & Grabar, N. (2020a). Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français. *Traitement Automatique des Langues*.

Cardon, R., & Grabar, N. (2020b). French biomedical text simplification: When small and precise helps. In *The 28th International Conference on Computational Linguistics*.

Cardon, R., & Grabar, N. (2021). Recherche de phrases parallèles à partir de corpus comparables pour la simplification de textes médicaux en français. In *Atelier SimpleText–GDR IA (2021)-Simplification et Vulgarisation des Textes Scientifiques*.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology* (pp. 7-10).

Carroll, J. A., Minnen, G., Pearce, D., Canning, Y., Devlin, S., & Tait, J. (1999). Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 269-270).

Carroll, J., & Oepen, S. (2005). High efficiency realization for a wide-coverage unification grammar. In *International Conference on Natural Language Processing* (pp. 165-176). Springer, Berlin, Heidelberg.

Coster, W., & Kauchak, D. (2011). Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation* (pp. 1-9).

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Chandrasekar, R., Doran, C., & Bangalore, S. (1996). Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Chandrasekar, R., & Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge-Based Systems*, *10*(3), 183-190

Chapman, K., Abraham, C., Jenkins, V., & Fallowfield, L. (2003). Lay understanding of terms used in cancer consultations. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, *12*(6), 557-566.

Chaumartin, F. R. (2008). ANTELOPE-Une plateforme industrielle de traitement linguistique. *Revue TAL*, *49*(2), pp-43.

Chen, Y. N., Wang, W. Y., & Rudnicky, A. I. (2013). Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 120-125). IEEE.

Chomsky, N. (1957). Syntactic structure. Mouton.

Chomsky, N. (1965). Aspects of the Theory of Syntax.

Choomthong, D. (2011). A case study of learning English passive of Thai EFL learners: Difficulties and learning strategies. In *The Asian conference on language learning official proceedings* (pp. 74-87).

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 93-98).

Choshen, L., & Abend, O. (2018). Automatic metric validation for grammatical error correction. *arXiv preprint arXiv:1804.11225*.

Chuquet, J. (2000). *Complexité syntaxique et sémantique: études de corpus*. Université de Poitiers UFR Langues et littératures; Maison des sciences de l'homme et de la société.

Clark, H. H., & Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Memory & Cognition*, *7*(1), 35-41.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Cohn, T., & Lapata, M. (2013). An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(3), 1-35.

Copestake, A. (2006). *Robust minimal recursion semantics*. Technical report, Cambridge Computer Lab.

Copestake, A. (2007). Semantic composition with (robust) minimal recursion semantics. In *ACL 2007 Workshop on Deep Linguistic Processing* (pp. 73-80).

Copestake, A. (2009, March). Invited Talk: Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 1-9).

Copestake, A., Emerson, G., Goodman, M. W., Horvat, M., Kuhnle, A., & Muszyńska, E. (2016, May). Resources for building applications with dependency minimal recursion semantics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1240-1247).

Copestake, A. A., & Flickinger, D. (2000, May). An Open Source Grammar Development Environment and Broad-coverage English Grammar Using HPSG. In *LREC* (pp. 591-600).

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on language and computation*, *3*(2-3), 281-332.

Coster, W., & Kauchak, D. (2011a). Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 665-669).

Coster, W., & Kauchak, D. (2011b). Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation* (pp. 1-9).

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Crain, S. (1985). On not being led up the garden path. *Natural language parsing*, 320-358.

Crystal, D. (2011). *A dictionary of linguistics and phonetics*. John Wiley & Sons.

Curran, J. R., Clark, S., & Bos, J. (2007). Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion volume proceedings of the demo and poster sessions* (pp. 33-36).

Davidson, D. (1967). Truth and meaning. In *Philosophy, language, and artificial intelligence* (pp. 93-111). Springer, Dordrecht.

Debusmann, R. (2006). Extensible Dependency Grammar: a modular grammar formalism based on multigraph description.

Desmets, M., Hamon, S., & Lavieu, B. (2003). Les grammaires HPSG. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (48), 57-76.

Devlin, S. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Devlin, S., & Canning, Y. (1999). Automatic text simplification for readers with aphasia. In *The British Aphasiology Society Biennial International Conference*. London: City University.

Devlin, S., & Unthank, G. (2006). Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 225-226).

Dixon, R. M. (2009). *Basic linguistic theory volume 1: Methodology* (Vol. 1). OUP Oxford.

Dixon, R. M. (2010). *Basic linguistic theory volume 2: Grammatical topics* (Vol. 2). Oxford University Press on Demand.

Dixon, R. M. (2012). Basic linguistic theory, vol. 3: Further grammatical topics.

Djemaa, M., Candito, M., Muller, P., & Vieu, L. (2016). Corpus annotation within the French FrameNet: a domain-by-domain methodology. In *Tenth international conference on language resources and evaluation (LREC 2016)*.

Dmitrieva, A., & Tiedemann, J. (2021). Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. ACL Anthology.

Dong, Y., Li, Z., Rezagholizadeh, M., & Cheung, J. C. K. (2019). EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. *arXiv preprint arXiv:1906.08104*.

Dras, M. (1999). A meta-level grammar: redefining synchronous TAG for translation and paraphrase. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 80-87).

Drndarević, B., Štajner, S., Bott, S., Bautista, S., & Saggion, H. (2013). Automatic text simplification in spanish: A comparative evaluation of complementing modules. In *International Conference*

on Intelligent Text Processing and Computational Linguistics (pp. 488-500). Springer, Berlin, Heidelberg.

Dumortier, J. L. (2001). Lisibilité du discours didactique: réflexions sur la compréhension en lecture des différents écrits disciplinaires. *Liège: université de Liège: Service de didactique des langues et littératures romanes*.

Edwards, N. (1980). *Difficulty in Text as a Function of Syntactic Complexity: A Study of Syntactic Complexity Within and Between Sentences*. Open University (United Kingdom).

Elhadad, M., & Robin, J. (1996). An overview of SURGE: A reusable comprehensive syntactic realization component.

Ellsworth, M., & Janin, A. (2007). Mutaphrase: Paraphrasing with framenet. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing* (pp. 143-150).

Engelhardt, P. E., & Ferreira, F. (2010). Processing coordination ambiguity. *Language and speech*, *53*(4), 494-509.

Erteschik-Shir, N., & Nomi, E. S. (1997). *The dynamics of focus structure*. Cambridge University Press.

Evans, R. J. (2011). Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and linguistic computing*, *26*(4), 371-388.

Fajardo, I., Ávila, V., Ferrer, A., Tavares, G., Gómez, M., & Hernández, A. (2014). Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension. *Journal of applied research in intellectual disabilities*, *27*(3), 212-225.

Fan, Z. (2019). *Building an HPSG Chinese grammar (Zhong)* (Doctoral dissertation, Nanyang Technological University).

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 229-237).

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech* (Vol. 280, No. 1, pp. 20-32).

Fillmore, C. (1982). Frame Semantics. || In Linguistics in the Morning Calm. 111-137. *Seoul: Hanshin Publishing Co*.

Flickinger, D. (2000). On building a more effcient grammar by exploiting types. *Natural Language Engineering*, *6*(1), 15-28.

Flickinger, D., Bender, E. M., & Oepen, S. (2014). Towards an encyclopedia of compositional semantics: Documenting the interface of the English Resource Grammar. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 875-881).

François, T., Müller, A., Degryse, B., & Fairon, C. (2018). AMesure, une plateforme web pour soutenir la rédaction simple de textes administratifs. *Repères-Dorif*, *16*.

François, T., Müller, A., Rolin, E., & Norré, M. (2020). Amesure: A web platform to assist the clear writing of administrative texts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 1-7).

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Natural Language & Linguistic Theory*, *5*(4), 519-559.

Frazier, L., & Clifton, C. (1996). *Construal*. Mit Press.

Freyhoff, G., Hess, G., Kerr, L., Menzel, E., Tronbacke, B., & Van Der Veken, K. (1998). Make It Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability for authors, editors, information providers, translators and other interested persons. *International League of Societies for Persons with Mental Handicap European Association, Brussels*.

Fujita, A. (2005). *Automatic generation of syntactically well-formed and semantically appropriate paraphrases* (Doctoral dissertation, Ph. D. thesis, Nara Institute of Science and Technology).

Gala Pavia, N. (2003). *Un modèle d'analyseur synthaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires* (Doctoral dissertation, Paris 11).

Gala, N., François, T., Bernhard, D., & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN'2014* (pp. 91-102).

Gala, N., François, T., Javourey-Drevet, L., & Ziegler, J. C. (2018). La simplification de textes, une aide à l'apprentissage de la lecture. *Langue française*, (3), 123-131.

Gala, N., & Javourey-Drevet, L. (2020). Mots «faciles» et mots «difficiles» dans ReSyf: un outil pour la didactique du lexique mobilisant polysémie, synonymie et complexité. *Lidil. Revue de linguistique et de didactique des langues*, (62).

Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J. C. (2020a). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Language Resources and Evaluation for Language Technologies (LREC)*.

Gala, N., Todirascu, A., Bernhard, D., Wilkens, R., & Meyer, J. P. (2020b). Transformations syntaxiques pour une aide à l'apprentissage de la lecture: typologie, adéquation et corpus adaptés. In *SHS Web of Conferences* (Vol. 78, p. 14006).

Gala, N., Todirascu, A., Javourey-Drevet, L., Bernhard, D., Wilkens, R., & Meyer, J. P. (2020). *Recommandations pour des transformations de textes français afin d'améliorer leur lisibilité et leur compréhension*.

Gala, N., & Ziegler, J. (2016). Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (pp. 59-66).

Gardent, C., & Parmentier, Y. (2007). Semtag, une architecture pour le développement et l'utilisation de grammaires d'arbres adjoints à portée sémantique. In *14e Conférence sur le Traitement Automatique des Langues Naturelles-TALN 2007* (pp. 175-184).

Garg, S., Galstyan, A., Hermjakob, U., & Marcu, D. (2016). Extracting biomolecular interactions using semantic parsing of biomedical text. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Gasperin, C., Maziero, E., & Aluisio, S. M. (2010). Challenging choices for text simplification. In *International Conference on Computational Processing of the Portuguese Language* (pp. 40-50). Springer, Berlin, Heidelberg.

Genest, P. E., & Lapalme, G. (2011). Framework for abstractive summarization using text-to-text generation. In *Proceedings of the workshop on monolingual text-to-text generation* (pp. 64-73).

Gerber, L., & Hovy, E. (1998). Improving translation quality by manipulating sentence length. In *Conference of the Association for Machine Translation in the Americas* (pp. 448-460). Springer, Berlin, Heidelberg.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal dependencies workshop 2018*.

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

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of experimental psychology: General*, *113*(2), 256.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: a component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 245.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length?. *Cognitive Science*, *34*(2), 286-310.

von Glasersfeld, E. (1970). The problem of syntactic complexity in reading and readability. *Journal of Reading Behavior*, *3*(2), 1-14.

Golub, L. S., & Kidder, C. (1974). Syntactic density and the computer. *Elementary English*, *51*(8), 1128-1131.

Goodman, M. W. (2018). *Semantic operations for transfer-based machine translation* (Doctoral dissertation).

Grabar, N., & Cardon, R. (2018). Clear-simple corpus for medical french. In *ATA*.

Grabar, N., Koptient, A., & Cardon, R. (2021). Traitement Automatique de Langues pour la simplification de documents de santé. *Bulletin de l'Association Française pour l'Intelligence Artificielle*.

Grabar, N., & Saggion, H. (2022). Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale* (pp. 453-463).

Gray, W. S., & Leary, B. E. (1935). What makes a book readable.

Guillaume, B., Bonfante, G., Masson, P., Morey, M., & Perrier, G. (2012, June). Grew: un outil de réécriture de graphes pour le TAL (Grew: a Graph Rewriting Tool for NLP)[in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 5: Software Demonstrations* (pp. 1-2).

Guillaume, B., de Marneffe, M. C., & Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL*, *60*(2), 71-95.

Guo, H., Pasunuru, R., & Bansal, M. (2018). Dynamic multi-level multi-task learning for sentence simplification. *arXiv preprint arXiv:1806.07304*.

Hajič, J. (1998). Building a syntactically annotated corpus: The prague dependency treebank. *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, 106-132.

Hajič, J., Hajicová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fucíková, E., Mikulová, M., Pajas, P., Popelka, J. and Semecký, J., (2012, May). Announcing Prague Czech-English Dependency Treebank 2.0. In *LREC* (pp. 3153-3160).

Halliday, M., & Matthiessen, C. (2014). *An introduction to functional grammar*. Routledge.

Hamilton, H. W., & Deese, J. (1971). Comprehensibility and subject-verb relations in complex sentences. *Journal of Verbal Learning and Verbal Behavior*, *10*(2), 163-170.

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of verbal learning and verbal behavior*, *13*(5), 512-521.

Hershcovich, D., Abend, O., & Rappoport, A. (2017). A transition-based directed acyclic graph parser for UCCA. *arXiv preprint arXiv:1704.00552*.

Hijazi, R. (2020). Transformations syntaxiques entre niveaux de simplification dans le corpus Newsela. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL* (pp. 137-150). ATALA; AFCP.

Hijazi, R., Espinasse, B, Gala, N. (2022). GRASS: A syntactic text simplification system based on semantic representations. In *11th International Conference on Natural Language Processing (NLP 2022)*. Copenhagen (Denmark).

Hinkel, E. (2001). Why English passive is difficult to teach (and learn). In *New perspectives on grammar teaching in second language classrooms* (pp. 245-270). Routledge.

Hinkel, E. (2004). Tense, aspect and the passive voice in L1 and L2 academic texts. *Language teaching research*, *8*(1), 5-29.

Hoeks, J. C. J. (1999). *The processing of coordination: semantic and pragmatic constraints on ambiguity resolution*. [Sl: sn].

Hoeks, J. C., Hendriks, P., Vonk, W., Brown, C. M., & Hagoort, P. (2006). Processing the noun phrase versus sentence coordination ambiguity: Thematic information does not completely eliminate processing difficulty. *Quarterly Journal of Experimental Psychology*, *59*(9), 1581-1599.

Hoeks, J. C., Vonk, W., & Schriefers, H. (2002). Processing coordinated structures in context: The effect of topic-structure on ambiguity resolution. *Journal of Memory and Language*, *46*(1), 99-119.

Hollenbach, B. E. (1983). Apposition and X-bar rules.

Hosenfeld, C. (1984). Case studies of ninth grade readers. *Reading in a foreign language*, *4*, 231-249.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Huddleston, R. (1984). *Introduction to the Grammar of English*. Cambridge University Press.

Huddleston, R., & Pullum, G. (2002). *The Cambridge grammar of the English language*. Cambridge University Press.

Huggins, A. W. F., & Adams, M. J. (2017). Syntactic aspects of reading comprehension. In *Theoretical issues in reading comprehension* (pp. 87-112). Routledge.

Hwang, W., Hajishirzi, H., Ostendorf, M., & Wu, W. (2015). Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 211-217).

Ibrahim, A. (2013). Une mesure unifiée de la complexité linguistique: l'analyse matricielle définitoire. *Nouvelles perspectives en sciences sociales: revue internationale de systémique complexe et d'études relationnelles*, *9*(1), 17-80.

Inui, K., Fujita, A., Takahashi, T., Iida, R., & Iwakura, T. (2003). Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing* (pp. 9-16).

Ivanova, A., Oepen, S., Øvrelid, L., & Flickinger, D. (2012). Who did what to whom?: A contrastive study of syntacto-semantic dependencies. In *Proceedings of the sixth linguistic annotation workshop* (pp. 2-11). Association for Computational Linguistics.

Javourey-Drevet, L. (2021). *La simplification de textes comme outil pour améliorer la fluidité et la compréhension de lecture chez les enfants à l'école primaire: une étude en longitudinal avec des textes littéraires et scientifiques chez des enfants entre 7 et 9 ans* (Doctoral dissertation, Aix-Marseille).

Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. (2020). Neural CRF model for sentence alignment in text simplification. *arXiv preprint arXiv:2005.02324*.

Johnson M. (1988), Attribute-Value Logic and the Theory of Grammar, CSLI Lecture Notes, Chicago University Press.

Johnson, C., Fillmore, C., Petruck, M., Baker, C., Ellsworth, M., Ruppenhofer, J., & Wood, E. (2002). *FrameNet: Theory and Practice. International Computer Science Institute*. Technical Report-02009. Berkeley, CA.

Shieber, S. M., & Schabes, Y. (1991). Synchronous tree-adjoining grammars.

Jakubowicz, C. (2007). Grammaire universelle et trouble spécifique du langage. *Noam Chomsky. Cahiers de l'Herne*, 164-175.

Jakubowicz, C., & Tuller, L. (2008). Specific language impairment in French. *Studies in French applied linguistics*, 97-134.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, *274*(5284), 114-116.

Kahane, S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. In *Actes de la 8Ëme confÈrence sur le Traitement Automatique des Langues Naturelles. Tutoriels* (pp. 17-76).

Kahane, S., & Gerdes, K. (2022). *Syntaxe théorique et formelle, Volume 1: Modélisation, unités, structures*. BoD–Books on Demand.

Kaplan, R. M., & Bresnan, J. (1995). Formal system for grammatical representation. *Formal issues in lexical-functional grammar*, (47), 29.

Kate, R. J., & Wong, Y. W. (2010). Semantic parsing. The task, the state of the art and the future. In *Tutorial abstracts of the 20th Meeting of the Association for Computational Linguistics* (p. 6).

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1537-1546).

Kingsbury, P., & Palmer, M. (2003). Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories* (Vol. 3).

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch.

Kintsch, W., & Miller, J. R. (1981). Readability: A view from cognitive psychology. *Teaching: Research Reviews*, 220-232.

Klare, G. R. (1985). *How to write readable English*. Hutchinson.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177-180).

Koller, A., & Striegnitz, K. (2002). Generation as dependency parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 17-24).

Konstas, I., & Lapata, M. (2012). Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 369-378).

Koptient, A., & Grabar, N. (2020). Fine-grained text simplification in French: steps towards a better grammaticality. In *ISHIMR Proceedings of the 18th International Symposium on Health Information Management Research*.

Kouylekov, M., & Oepen, S. (2014). Semantic technologies for querying linguistic annotations: An experiment focusing on graph-structured data.

Kriz, R., Miltsakaki, E., Apidianaki, M., & Callison-Burch, C. (2018). Simplification using paraphrases and context-based lexical substitution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 207-217).

Kriz, R., Sedoc, J., Apidianaki, M., Zheng, C., Kumar, G., Miltsakaki, E., & Callison-Burch, C. (2019). Complexity-weighted loss and diverse reranking for sentence simplification. *arXiv preprint arXiv:1904.02767*.

Kuhlmann, M., & Oepen, S. (2016). Towards a catalogue of linguistic graph banks. *Computational Linguistics*, *42*(4), 819-827.

Kusters, W., & Muysken, P. C. (2001). The complexities of arguing about complexity.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th international symposium on natural language processing* (p. 7).

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied linguistics*, *35*(5), 607-614.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge university press.

Lamercerie, A. (2021). *Principe de transduction sémantique pour l'application de théories d'interfaces sur des documents de spécification* (Doctoral dissertation, Rennes 1).

Langkilde, I. (2002). An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the international natural language generation conference* (pp. 17-24).

Langkilde, I., & Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Lavoie, B., Rambow, O., & Reiter, E. (1997, March). Customizable descriptions of object-oriented models. In *Fifth Conference on Applied Natural Language Processing* (pp. 253-256).

Lecocq, P., Casalis, S., Leuwers, C., & Watteau, N. (1996). *Apprentissage de la lecture et compréhension d'énoncés*. Presses Univ. Septentrion.

Lerner, E. B., Jehle, D. V., Janicke, D. M., & Moscati, R. M. (2000). Medical communication: do our patients understand?. *The American journal of emergency medicine*, *18*(7), 764-766.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156-166.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2018). Toward abstractive summarization using semantic representations. *arXiv preprint arXiv:1805.10399*.

Liu, D., & Gildea, D. (2010). Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 716-724).

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and cognitive processes*, *9*(2), 157-201.

Mallinson, J., & Lapata, M. (2019). Controllable sentence simplification: Employing syntactic and lexical constraints. *arXiv preprint arXiv:1910.04387*.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English : The Penn Treebank. Computational Linguistics, 19(2) :313–330.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., ... & Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Martin, L. (2021). *Automatic sentence simplification using controllable and unsupervised methods* (Doctoral dissertation, Sorbonne Université).

Martins, D., & Le Bouédec, B. (1998). La production d'inférences lors de la compréhension de textes chez des adultes: une analyse de la littérature. *L'année psychologique*, *98*(3), 511-543.

Martins, A. F., Junczys-Dowmunt, M., Kepler, F. N., Astudillo, R., Hokamp, C., & Grundkiewicz, R. (2017). Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, *5*, 205-218.

Martos, J., Freire, S., González, A., Gil, D., & Sebastian, M. (2012). D2. 1: Functional requirements specifications and user preference survey. *Project FIRST*.

Marzinotto, G. (2019). *Semantic frame based analysis using machine learning techniques: improving the cross-domain generalization of semantic parsers* (Doctoral dissertation, Aix-Marseille).

McCawley, J. D. (1995). An overview of\appositive" constructions. In *Proceedings of ESCOL* (Vol. 95).

McDonald, R., & Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics*, *37*(1), 197-230.

McKoon, G., Ward, G., Ratcliff, R., & Sproat, R. (1993). Morphosyntactic and pragmatic factors affecting the accessibility of discourse entities. *Journal of Memory and Language*, *32*(1), 56-75.

Medero, J., & Ostendorf, M. (2011). Identifying targets for syntactic simplification. In *Speech and Language Technology in Education*.

Mehler, J. (1963). Some effects of grammatical transformations on the recall of English sentences. *Journal of verbal Learning and verbal Behavior*, *2*(4), 346-351.

Mel'čuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press.

Màrquez, L., Carreras, X., Litkowski, K. C., & Stevenson, S. (2008). Semantic role labeling: an introduction to the special issue. *Computational linguistics*, *34*(2), 145-159.

Meltzer, J. A., McArdle, J. J., Schafer, R. J., & Braun, A. R. (2010). Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. *Cerebral cortex*, *20*(8), 1853-1864.

Meyer, C. F. (1992). *Apposition in contemporary English*. Cambridge Univ Pr.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. (2004). The NomBank project: An interim report. In *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004* (pp. 24-31).

Michalon, O. (2017). *Modèles statistiques pour la prédiction de cadres sémantiques* (Doctoral dissertation, Aix-Marseille).

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.

Miller, G. A. (1962). Some psychological studies of grammar. *American psychologist*, *17*(11), 748.

Mishra, K., Soni, A., Sharma, R., & Sharma, D. M. (2014). Exploring the effects of sentence simplification on Hindi to English machine translation system. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)* (pp. 21-29).

Miyao, Y., Oepen, S., & Zeman, D. (2014). In-house: An ensemble of pre-existing off-the-shelf parsers. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 335-340).

Moeljadi, D., Bond, F., & Song, S. (2015). Building an HPSG-based indonesian resource grammar (INDRA). In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop* (pp. 9-16).

Monville-Burston, M. (2013). Complexité et transfert dans l'acquisition du français langue étrangère: le cas des apprenants chypriotes du FLE. *Travaux de linguistique*, (1), 97-134.

Moreb, B. (2016). The frequency of the passive voice in freshman academic books.

Morey, M. (2011). *Étiquetage grammatical symbolique et interface syntaxe-sémantique des formalismes grammaticaux lexicalisés polarisés* (Doctoral dissertation, Université de Lorraine).

Morgan, P. J. (1978). *An investigation of the structure and mode of classification strategies of retarded and achieving readers: a modified replication*. Ohio University.

Morgan, M. F., & Moni, K. B. (2008). Literacy: Meeting the challenge of limited literacy resources for adolescents and adults with intellectual disabilities. *British Journal of Special Education*, *35*(2), 92-101.

Murcia, M. C., & Freeman, D. L. (1999). The grammar book.

Muszyńska, E. (2016). Graph-and surface-level sentence chunking. In *Proceedings of the ACL 2016 Student Research Workshop* (pp. 93-99).

Nakanishi, H., Miyao, Y., & Tsujii, J. I. (2005). Probabilistic models for disambiguation of an HPSG-based chart generator. In *Proceedings of the Ninth International Workshop on Parsing Technology* (pp. 93-102).

Narayan, S., & Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics* (pp. 435-445).

Narayan, S., & Gardent, C. (2015). Unsupervised sentence simplification using deep semantics. *arXiv preprint arXiv:1507.08452*.

Narayan, S., Gardent, C., Cohen, S. B., & Shimorina, A. (2017). Split and rephrase. *arXiv preprint arXiv:1707.06971*.

Neilson, K. J. (2016). A text analysis of how passive voice in a biology textbook impacts English language learners.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE™//A lexical database for contemporary french: LEXIQUE™. *L'année psychologique*, *101*(3), 447-462.

Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2019). Transforming complex sentences into a semantic hierarchy. *arXiv preprint arXiv:1906.01038*.

Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 85-91).

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Zeman, D. (2016, May). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666).

Norbury, C. F. (2005). Barking up the wrong tree? Lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of experimental child psychology*, *90*(2), 142-171.

Norman, S., Kemper, S., Kynette, D., Cheung, H., & Anagnopoulos, C. (1991). Syntactic complexity and adults' running memory span. *Journal of Gerontology*, *46*(6), P346-P351.

Rogers, J. (1996). A model-theoretic framework for theories of syntax. *arXiv preprint cmp-lg/9604023*.

O'Connor, I. M., & Klein, P. D. (2004). Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders. *Journal of autism and developmental disorders*, *34*(2), 115-127.

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). Lingo redwoods. *Research on Language and Computation*, *2*(4), 575-596.

Oepen, S., & Lønning, J. T. (2006). Discriminant-Based MRS Banking. In *LREC* (pp. 1250-1255).

Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Cinková, S., Flickinger, D., ... & Uresova, Z. (2015, June). Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 915-926).

Oepen, S., Abend, O., Hajic, J., Hershcovich, D., Kuhlmann, M., O'Gorman, T., ... & Urešová, Z. (2019). MRP 2019: Cross-framework Meaning Representation Parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning* (pp. 1-27).

Ortega, L. (2012). Interlanguage complexity. *Linguistic complexity: Second language acquisition, indigenization, contact*, *13*, 127.

Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, *60*, 549-593.

Paetzold, G., & Specia, L. (2013). Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117-134.

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

Parsons, T. (1990). Events in the semantics of English: A study in subatomic semantics.

Pan, X., Cassidy, T., Hermjakob, U., Ji, H., & Knight, K. (2015). Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1130-1139).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 425-430).

Pavlick, E., & Callison-Burch, C. (2016). Simple PPDB: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 143-148).

Pearson, P. D. (1974). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations. *Reading Research Quarterly*, 155-192.

Perrier, G. (2003). *Les grammaires d'interaction* (Doctoral dissertation, Université Nancy 2).

Perrier, G. (2021). Étude des dépendances syntaxiques non projectives en français. *Revue TAL*, *62*(1).

Perrier, G., Candito, M., Guillaume, B., Ribeyre, C., Fort, K., & Seddah, D. (2014, July). Un schéma d'annotation en dépendances syntaxiques profondes pour le français. In *TALN-Traitement Automatique des Langues Naturelles* (pp. 574-579).

Petersen, S. E., & Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*.

PlainLanguage (2011). Federal plain language guidelines.

Pollard, C., & Sag, I. A. (1988). *Information-based syntax and semantics: Vol. 1: fundamentals*. Center for the Study of Language and Information.

Pollard, C. J. (1999). Strong generative capacity in HPSG. *Lexical and constructional aspects of linguistic explanation*, *1*, 281-297.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Price, J. (1984). *How to write a computer manual: A handbook of software documentation*. Benjamin-Cummings Publishing Co., Inc..

Pullum, G. K. (2014). Fear and loathing of the English passive. *Language & Communication*, *37*, 60-74.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A Comprehensive Grammar of the English Language (2. udg.).

Ramadier, L., & Lafourcade, M. (2018). Utilisation d'une base de connaissances de spécialité et de sens commun pour la simplification de comptes-rendus radiologiques (Radiological text simplification using a general knowledge base). In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN* (pp. 447-454).

Reddy, S., Täckström, O., Collins, M., Kwiatkowski, T., Das, D., Steedman, M., & Lapata, M. (2016). Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, *4*, 127-140.

Reddy, S., Täckström, O., Petrov, S., Steedman, M., & Lapata, M. (2017). Universal semantic parsing. *arXiv preprint arXiv:1702.03196*.

Rello, L. (2014). *DysWebxia: a text accessibility model for people with dyslexia* (Doctoral dissertation, Universitat Pompeu Fabra).

Rello, L., Baeza-Yates, R., Bott, S., & Saggion, H. (2013a). Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (pp. 1-10).

Rello, L., Baeza-Yates, R., & Saggion, H. (2013b). The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 501-512). Springer, Berlin, Heidelberg.

Rescher, N. (1998). *Complexity: A philosophical overview*. Transaction Publishers.

Ribeyre, C. (2016). *Méthodes d'analyse supervisée pour l'interface syntaxe-sémantique* (Doctoral dissertation, Université Paris Diderot).

Riegel, M., Pellai, J. C., & Rioul, R. (1998). Grammaire méthodique du français. Collection Linguistique Nouvelle. Presses Universitaires de France, Paris, 3e3.

Rioul, R. (1983). Les appositions dans la grammaire française. *L'information grammaticale*, *18*(1), 21-29.

Romero, C. (2013). Comment le sens peut-il être complexe? L'exemple des comparaisons d'intensité. *Nouvelles perspectives en sciences sociales: revue internationale de systémique complexe et d'études relationnelles*, *9*(1), 171-198.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437-442.

Sag, I., Wasow, T., & Bender, E. (1999). Syntactic Theory: A formal approach. *Standford: CSLI Publications*.

Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1-137.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47), 341-342.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., & Drndarevic, B. (2015). Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, *6*(4), 1-36.

Savin, H. B., & Perchonock, E. (1965). Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, *4*(5), 348-353.

Scarton, C., Paetzold, G., & Specia, L. (2018). Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Scarton, C and Specia L. (2018). "Learning Simplifications for Specific Target Audiences." In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2, pp. 712–718.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies : An Improved Representation for Natural Language Understanding Tasks. In Proceedings of LREC 2016), pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).

Scott, S. E. (2009). *Knowledge for teaching reading comprehension: Mapping the terrain*. University of Michigan.

Seretan, V. (2012). Acquisition of syntactic simplification rules for french. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Sgall, P. (1992). *Underlying structure of sentences and its relations to semantics*. na.

Sgall, P., Hajicová, E., Hajicová, E., Panevová, J., & Panevova, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, *4*(1), 58-70.

Shardlow, M., & Nawaz, R. (2019). Neural text simplification of clinical letters with a domain specific phrase table.

Shen, D., & Lapata, M. (2007). Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 12-21).

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, *4*(1), 77-109.

Siddharthan, A. (2011). Text simplification using typed dependencies: a comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 2-11). Association for Computational Linguistics.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, *165*(2), 259-298.

Siegel, M., & Bender, E. M. (2002). Efficient deep processing of Japanese. *arXiv preprint cs/0207005*.

Skehan, P. (1991). Individual differences in second language learning. *Studies in second language acquisition*, *13*(2), 275-298.

Smith, D. A., & Eisner, J. (2006). Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings on the Workshop on Statistical Machine Translation* (pp. 23-30).

Sopher, H. (1971). Apposition.

Soler, A. G., Apidianaki, M., & Allauzen, A. (2018). A comparative study of word embeddings and other features for lexical complexity detection in French. In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN* (pp. 499-508).

Sorin, N. (1996). De la lisibilité linguistique à une lisibilité sémiotique. *Revue québécoise de linguistique*, *25*(1), 61-97.

Specia, L. (2010). Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language* (pp. 30-39). Springer, Berlin, Heidelberg.

Štajner, S. (2016). New data-driven approaches to text simplification.

Štajner, S., Drndarevic, B., & Saggion, H. (2013). Corpus-based sentence deletion and split decisions for Spanish text simplification.

Štajner, S., & Popović, M. (2016). Can text simplification help machine translation?. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation* (pp. 230-242).

Štajner, S., & Saggion, H. (2018). Data-driven text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts* (pp. 19-23).

Staub, A., & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: evidence from either... or. *Journal of experimental psychology: Learning, memory, and cognition*, *32*(2), 425.

Steedman, M., & Baldridge, J. (2009). Combinatory categorial grammar. nontransformational syntax: A guide to current models. *Blackwell, Oxford*, *9*, 13-67.

Subirats, C. (2009). 5. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In *Multilingual FrameNets in computational lexicography* (pp. 135-162). De Gruyter Mouton.

Sulem, E., Abend, O., & Rappoport, A. (2018a). Simple and effective text simplification using semantic and neural methods. *arXiv preprint arXiv:1810.05104*.

Sulem, E., Abend, O., & Rappoport, A. (2018b). Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

Surya, S., Mishra, A., Laha, A., Jain, P., & Sankaranarayanan, K. (2018). Unsupervised neural text simplification. *arXiv preprint arXiv:1810.07931*.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension.

Tesnière, L. (1959). Eléments de syntaxe structurale.

Thompson, C. K., & Shapiro, L. P. (2007). Complexity in treatment of syntactic deficits. *American Journal of Speech-Language Pathology*.

Todirascu, A., Cargill, M., & Buhnila, I. (2019). Erreurs d'apprenants: typologie et annotations. In *10e Journées Internationales de la Linguistique de Corpus*.

Todirascu, A., François, T., Bernhard, D., Gala, N., Ligozat, A. L., & Khobzi, R. (2017). Chaînes de référence et lisibilité des textes: Le projet ALLuSIF. *Langue française*, (3), 35-52.

Todirascu, A., Wilkens, R., Rolin, E., François, T., Bernhard, D., & Gala, N. (2022). HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French. In *12th International Conference on Language Resources and Evaluation (LREC)*.

Tomita, M. (1985). *An efficient context-free parsing algorithm for natural languages and its applications*. Carnegie Mellon University.

Torrent, T. T., Ellsworth, M., Baker, C., & Matos, E. E. (2018). The multilingual framenet shared annotation task: a preliminary report. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 62-68).

Toutanova, K., Manning, C. D., Flickinger, D., & Oepen, S. (2005). Stochastic HPSG parse disambiguation using the Redwoods corpus. *Research on Language and Computation*, *3*(1), 83-105.

Tran, T. M., Chekroud, H., Thiery, P., & Julienne, A. (2009). Internet et soins: un tiers invisible dans la relation médecine/patient. *Ethica Clinica*, *53*, 34-43.

Tur, G., Hakkani-Tür, D., & Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, *45*(2), 171-186

Turner, A., & Greene, E. (1977). *The construction and use of a propositional text base* (pp. 77-63). Boulder: Institute for the Study of Intellectual Behavior, University of Colorado.

Van den Bercken, L., Sips, R. J., & Lofi, C. (2019). Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference* (pp. 3286-3292).

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, *43*(6), 1606-1618.

Vecchiato, S., & Gerolimich, S. (2013). La langue médicale est-elle «trop complexe»?. *Nouvelles perspectives en sciences sociales: revue internationale de systémique complexe et d'études relationnelles*, *9*(1), 81-122.

Vickrey, D., & Koller, D. (2008). Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT* (pp. 344-352).

Vu, T., Hu, B., Munkhdalai, T., & Yu, H. (2018). Sentence simplification with memory-augmented neural networks. *arXiv preprint arXiv:1804.07445*.

White, M. (2004). Reining in CCG chart realization. In *International Conference on Natural Language Generation* (pp. 182-191). Springer, Berlin, Heidelberg.

White, M., & Rajkumar, R. (2009). Perceptron reranking for CCG realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 410-419).

Wieting, J., & Gimpel, K. (2017). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Wilkens, R., Oberle, B., & Todirascu, A. (2020). Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)* (pp. 93-100).

Wilkens, R., & Todirascu, A. (2020). Un corpus d'évaluation pour un système de simplification discursive. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles* (pp. 361-369). ATALA; AFCP.

Wong, Y. W., & Mooney, R. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 960-967).

Woodsend, K., & Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 409-420).

Wubben, S., Van Den Bosch, A., & Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1015-1024).

Xenouleas, S., Malakasiotis, P., Apidianaki, M., & Androutsopoulos, I. (2019). SUM-QE: a BERT-based Summary Quality Estimation Model Supplementary Material. In *Proceedings of EMNLP*.

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, *3*, 283-297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401-415.

Yan, C. (2021). *Complexité syntaxique et flux de dépendance: études quantitatives dans les treebanks universal dependencies* (Doctoral dissertation, Université de Nanterre-Paris X).

Zhang, X., & Lapata, M. (2017). Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

Zhong, H., & Stent, A. (2005). Building surface realizers automatically from corpora. *Proceedings of UCNLG*, *5*, 49-54.

Zhu, Z., Bernhard, D., & Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 1353-1361). Association for Computational Linguistics.

Table des Figures

Table des Tableaux

Annexes

A. Guide d'annotation humaine

Introduction

Nous avons développé un système de simplification de textes que nous souhaitons évaluer. Lors de l'évaluation humaine, les sorties d'un système de simplification sont soumises au jugement d'utilisateurs humains.

Notre système de simplification porte essentiellement sur le découpage de phrase (*Sentence splitting*). Cette opération a pour but de diviser les phrases longues et contenant des constructions syntaxiques complexes (coordinations, subordinations, appositives, etc.) en phrases plus courtes et plus compréhensibles. D'autres opérations de transformations sont prises en compte comme la transformation de la voix passive en voix active.

Pour l'évaluation, les annotateurs analysent si la phrase est correctement découpée. Il s'agit de voir si le point de découpage est correct et si le sujet partagé est correctement reconstruit.

Nous sommes inspirés du guide d'annotation des erreurs proposé dans le cadre des projets Alector (Gala et al., 2020) et SimpleApprenant (Todirascu et al, 2019). L'analyse des erreurs consiste à parcourir chaque phrase et à mettre en évidence les erreurs qui sont ensuite catégorisées.

Mesures

Les performances des systèmes de SAT sont généralement évaluées par des jugements humains : la **préservation du sens** (adaptation, sens conservé par rapport à l'original), la **grammaticalité** (fluence) et la **simplicité** (relative à l'original) sur une échelle Likert (1-3), où le score le plus élevé indique la meilleure sortie. La simplicité est une mesure qu'on considère comme subjective, dans le sens où les annotateurs ont un degré de liberté d'interprétation des critères, selon ce qu'il voit comme simple. Alors que l'adaptation et la grammaticalité sont définies selon des règles fixes qui déterminent si une phrase est grammaticale et si les éléments sont conservés.

Les valeurs proposées pour chaque variable sont :

1. **Adaptation**, sens préservé (*adequacy*, i.e. *meaning preservation*) de la sortie par rapport à l'original ; mesurer à quel point les phrases originales et simplifiées sont similaires en termes de sens. Le jugement sur la sémantique doit répondre à la question de savoir si la transformation effectuée préserve la sémantique originale de la phrase.
   1 = pas du tout le même sens (indépendamment de la grammaticalité de la phrase), sens complétement changé ; sujet partagé est mal reconstruit.
   2 = éléments sont différents ou absents / ambigüité ; pas toutes les constructions sont découpées.
   3 = exactement le même sens, parfaitement compréhensible (indépendamment de la grammaticalité de la phrase).

2. **Fluence**, i.e. grammaticalité. Le jugement sur la grammaticalité doit répondre à la question de savoir si la phrase reste grammaticale après les modifications effectuées (indépendamment du maintien du sens).
   1 = phrase agrammaticale.
   2 = nombre limité d'erreurs grammaticales.
   3 = phrase parfaitement grammaticale.

De plus, pour la propriété de grammaticalité, nous évaluons les 3 types d'erreurs (présentés plus bas).

3. **Simplicité** linguistique. Il s'agit d'analyser si le découpage a rendu la phrase plus simple à comprendre avec une structure SVO, sans tenir compte des erreurs de grammaire, au regard des transformations mises en place.
   1 = sortie plus complexe que l'original (les transformations ont rendu la phrase plus difficile à comprendre)
   2 = sortie plus simple (une seule transformation qui a rendu la phrase plus accessible)
   3 = beaucoup plus simple que l'original (deux ou plusieurs transformations qui simplifient la phrase).

Exemple 1

Phrase Originale	Phrase simplifiée	Adaptation	Fluence	Simplicité
George Frideric Handel also served as Kapellmeister for George, Elector of Hanover (who eventually became George I of Great Britain).	Also, George Frideric Handel served as Kapellmeister, for George.	3	3	3
	Kapellmeister is Elector of Hanover.	1	3	1
	Kapellmeister became George I of Great Britain, eventually.	1	3	1

Exemple 2

Phrase Originale	Phrase simplifiée	Adaptation	Fluence	Simplicité
The lawyer, Brandon (Waise Lee), became his idol, and MK Sun grew up to be a lawyer.	-	1	1	1
	MK Sun grew up to be a lawyer.	3	3	3

Types d'erreurs

Nous proposons d'annoter les catégories d'erreurs identifiées dans la sortie du système.

Nous avons 3 grandes catégories d'erreurs : orthographiques, grammaticales et syntaxiques.

1. **Les erreurs orthographiques (étiquette Ortho) :**

Cette sous-catégorie concerne les erreurs liées à l'orthographe du mot.
Ex : A weakenning [weakening] trend began.

2. **Les erreurs grammaticales (étiquette Gram) :**

Gram_ConjugaisonVb : Cette sous-catégorie concerne les erreurs de conjugaison de verbe (le verbe peut être mal conjugué, être au mauvais temps ou pas conjugué du tout).
Ex : Most of the songs were written by Richard M. Sherman and Robert B. Sherman. → Richard M. Sherman and Robert B. Sherman ~~written~~ [write] most of the songs.

3. **Les erreurs syntaxiques (étiquette Synt) :**

- *Absence de ponctuation :* Salem is a city in Essex County[,] Massachusetts[,] United States.
- *Rajout de ponctuation dans un endroit où il ne faut pas :* The mucous helps to catch ants and termites that [,] adhere to it.
- *Ordre des mots :*
  - Mauvaise position de l'adjectif : Alessandro Mazzola is a former Italian football player →Alessandro Mazzola is an Italian former football player
- *Mauvaise ponctuation* (résultant d'une mauvaise segmentation des phrases) :

  Therefore, these PDFs can not be distributed without further manipulation. If they contain images. →Therefore, these PDFs can not be distributed without further manipulation if they contain images.

- *Absence de verbe* : Calvin Baker [is] an American novelist.
- *Absence d'objet :* But he rejected [them] all.
- *Mauvaise réécriture du syntagme nominal sujet* : The tarantula, the trickster character, spun a black cord. → The tarantula is the trickster character. The ~~character~~ [tarantula] spun a black cord.
- *Absence de préposition :* Mazzola is born [on] 8 November 1942.
- *Absence de déterminant :* Lindenmeyer is [an] Australian National University.

B. Système de règles

Paquets de règles de transformations des appositives

```
rule select_sujet
{
	pattern
	{
		R [gpred="appos"];
		m: R -[ARG1:NEQ]-> N;
		p: R -[ARG2:NEQ]-> M; M[pos="n"]
	}
	commands {
		N.select=yes; del_edge m; N.subject=yes }
	}

package propagate {
	rule down {
	pattern { M[select]; Z1[!select]; M -> Z1 }
	commands { Z1.select = yes }}

	rule up {
	pattern { M[select]; Z2[!select]; Z2 -> M; Z2<M }
	commands { Z2.select = yes }} }

rule right_limit {
	pattern {LR_selected[select=yes];LR[!select];LR > LR_selected;}
	without {X[select=yes]; LR_selected < X;}
	commands {LR.last=yes;}}

rule copy_node{
	pattern {S[select=yes];LR[last=yes];}
	without {X[select=yes];S > X;}
	commands{del_feat S.select;add_node S_copy:<LR;append_feats S ==> S_copy;del_feat
S.subject;add_edge S -[copy:yes]-> S_copy}}

rule copy_edge{
	pattern {A1 -[copy:yes]-> A2;B1 -[copy:yes]-> B2;e: A1 -> B1}
	without {f: A2 -> B2; e.label = f.label}
	commands{add_edge f: A2 -> B2; f.label = e.label;}}

rule del_copy_edge{
	pattern { c: A1 -[copy:yes]-> A2 }
	commands { del_edge c }}

rule add_verb{
```

```
        pattern { S [subject=yes];R [gpred=appos];m: R-[ARG2:NEQ]-> M; F[last=yes]}
        without {S2[subject=yes];S > S2}
        commands{add_node V :>
S;V.lemma=be;V.pos="v";V.sense=id;V.SF=prop;V.TENSE=pres;V.MOOD=indicative;V.cvarsort
=e;add_edge V -[ARG1:NEQ]-> S;del_feat S.subject; del_feat F.last; add_edge V -[ARG2:NEQ]->
M;}}

rule del_appos{
        pattern { R [gpred=appos];}
        commands{del_node R; }}

strat appos
{Seq (select_sujet,Onf(propagate),
right_limit,
Onf(copy_node),
Onf(copy_edge),
Onf(del_copy_edge),
add_verb,
del_appos,
)}
```

Paquets de règles de transformation des coordinations

```
% ------------------------------------------------------------------------------------------------------------
% RULE I COORDINATIONS : un Sujet non partagé avec "AND" à supprimer
% ------------------------------------------------------------------------------------------------------------
rule select_coord {
  pattern {
    A[lemma = "and"];
    V1[pos = "v"];
    V2[pos = "v"];
    m: A - > V1;
    a: A - > V2;
  }
  commands {
    A.select = yes
  }
}
rule deletion {
  pattern {
    A[lemma = "and"];
    e: K - [MOD: EQ] - > L;
    A << e;
  }
  commands {
```

```
      del_edge e;
      del_node A;
   }
}
strat coord_nonpart_and {
   Seq(select_coord, deletion)
}


% ----------------------------------------------------------------------------------------------------------
% RULE II COORDINATIONS : un Sujet non partagé avec "BUT, OR, NOR, SO" à garder
% ----------------------------------------------------------------------------------------------------------

rule select_coordi {
   pattern {
      A[lemma = "but" | "or" | "nor" | "so"];
      V1[pos = "v"];
      V2[pos = "v"];
      m: A - > V1;
      V2 >> A;
      a: A - > V2;
   }
   commands {
      A.select = yes;
      del_edge m
   }
}
rule deletion {
   pattern {
      A[lemma = "but" | "or" | "nor" | "so"];
      e: K - [MOD: EQ] - > L;
      A << e;
   }
   commands {
      del_edge e;
   }
}
strat coord_nonpart_but {
   Seq(select_coordi, deletion)
}


% ----------------------------------------------------------------------------------------------------------
% RULE III COORDINATIONS : un Sujet partagé avec "AND" à supprimer
% ----------------------------------------------------------------------------------------------------------

rule coord_suj_part_and {
   pattern {
      A[lemma = "and"];
```

```
      V1[pos = "v"];
      V2[pos = "v"];
      m: A - > V1;
      a: A - > V2;
      s1: V1 - > S;
      s2: V2 - > S;
      V1 << V2
   }
   without {
      V1 < A;
      A < V2
   }
   commands {
      del_edge s1;
      A.coord = yes;
      add_edge k: V2 - > A;
      k.label = s2.label;
      S.select = yes;
      S.sub = yes;
      S.sujet = yes;
      V2.vb = yes;
      del_edge s2;
   }
}
rule deletion {
   pattern {
      A[coord = yes];
      e: K - [MOD: EQ] - > L;
      A << e;
   }
   commands {
      del_edge e;
   }
}
package propagate {
   rule down {
      pattern {
         N[select];
         Z1[!select];
         N - > Z1;
      }
      commands {
         Z1.select = yes
      }
   }
   rule up {
      pattern {
         N[select];
         Z2[!select];
```

```
        Z2 - > N;
        Z2 < N
      }
      commands {
        Z2.select = yes
      }
    }
}
rule copy_node {
    pattern {
      S[select = yes];
      V2[vb = yes];
    }
    without {
      X[select = yes];
      S > X;
    }
    commands {
      del_feat S.select;
      add_node S_copy: < V2;
      append_feats S == > S_copy;
      add_edge S - [copy: yes] - > S_copy;
      del_feat S.sujet
    }
}
rule copy_edge {
    pattern {
      A1 - [copy: yes] - > A2;
      B1 - [copy: yes] - > B2;
      e: A1 - > B1
    }
    without {
      f: A2 - > B2;e.label = f.label
    }
    commands {
      add_edge f: A2 - > B2;
      f.label = e.label;
    }
}
rule del_copy_edge {
    pattern {
      c: A1 - [copy: yes] - > A2;
    }
    commands {
      del_edge c;
    }
}
rule return_node {
    pattern {
```

```
      A[lemma = "and"];
      V1[pos = "v"];
      A - > V1;
      c: V2 - > A;
      V2[vb = yes];
      S2[sujet = yes];
   }
   commands {
      add_edge d: V2 - > S2;
      d.label = c.label;
      del_edge c;
      del_feat S2.sujet;
      del_feat S2.sub;
      del_feat V2.vb;
   }
}
rule del_nodes {
   pattern {
      A[lemma = "and"];
      V1[pos = "v"];
      S1[sub = yes];
      A - > V1;
      A >> V1;
   }
   commands {
      add_edge V1 - [ARG1: NEQ] - > S1;
      del_feat S1.sub;
      del_node A;
   }
}
strat coord_suj_part_and {
   Seq(coord_suj_part_and, deletion, Onf(propagate), Onf(copy_node), Onf(copy_edge),
Onf(del_copy_edge), return_node, del_nodes, )
}



% ----------------------------------------------------------------------------------------------------------
% RULE II COORDINATIONS : un Sujet partagé avec "BUT, OR, NOR, SO" à garder
% ----------------------------------------------------------------------------------------------------------

rule coord_suj_part_but {
   pattern {
      A[lemma = "but" | "or" | "nor" | "so"];
      V1[pos = "v"];
      V2[pos = "v"];
      m: A - [ARG1: EQ] - > V1;
      a: A - [ARG2: EQ] - > V2;
      s1: V1 - > S;
```

```
      s2: V2 - > S;
      f: V2 - [MOD: EQ] - > V1
   }
   commands {
      del_edge s1;
      del_edge s2;
      del_edge f;
      S.select = yes;
      S.sub = yes;
      S.sujet = yes;
      V2.vb = yes
   }
}
package propagate {
   rule down {
      pattern {
         N[select];
         Z1[!select];
         N - > Z1;
      }
      commands {
         Z1.select = yes
      }
   }
   rule up {
      pattern {
         N[select];
         Z2[!select];
         Z2 - > N;
         Z2 - [RSTR: H] - > N
      }
      commands {
         Z2.select = yes
      }
   }
}
rule copy_node {
   pattern {
      S[select = yes];
      V2[vb = yes];
   }
   without {
      X[select = yes];
      S > X;
   }
   commands {
      del_feat S.select;
      add_node S_copy: < V2;
      append_feats S == > S_copy;
```

```
      add_edge S - [copy: yes] - > S_copy;
      del_feat S.sujet
   }
}
rule copy_edge {
   pattern {
      A1 - [copy: yes] - > A2;
      B1 - [copy: yes] - > B2;
      e: A1 - > B1
   }
   without {
      f: A2 - > B2;e.label = f.label
   }
   commands {
      add_edge f: A2 - > B2;
      f.label = e.label;
   }
}
rule del_copy_edge {
   pattern {
      c: A1 - [copy: yes] - > A2
   }
   commands {
      del_edge c;
   }
}
rule return_node {
   pattern {
      A[lemma = "but" | "or" | "nor" | "so"];
      V1[pos = "v"];
      c: A - > V1;
      S1[sub = yes];
      V2[vb = yes];
      S2[sujet = yes];
   }
   commands {
      del_edge c;
      add_edge V2 - [ARG1: NEQ] - > S2;
      add_edge V1 - [ARG1: NEQ] - > S1;
      del_feat S1.sub;
      del_feat S2.sujet;
      del_feat S2.sub;
      del_feat V2.vb;
   }
}
strat coord_sujepart_but {
   Seq(coord_suj_part_but, Onf(propagate), Onf(copy_node), Onf(copy_edge),
Onf(del_copy_edge), return_node, )
}
```

Paquets de règles de transformation des subordinations

```
% ----------------------------------------------------------------------------------------------------------
% RULE SUBORDINATION
% ----------------------------------------------------------------------------------------------------------

rule subord {
   pattern {
      A[gpred = "subord"];
      V1[pos = v];
      V2[pos = v];
      n1: A - [ARG1: H] - > V1;
      n2: A - [ARG2: H] - > V2;
      s1: V1 - > S;
   }
   commands {
      V2.PROG = V1.PROG;
      V2.TENSE = V1.TENSE;
      S.select = yes;
      S.sujet = yes;
      del_node A;
      S.subj = yes;
      del_edge s1;
      V1.vb1 = yes;
      V2.vb2 = yes
   }
}
package propagate {
   rule down {
      pattern {
         N[select];
         Z1[!select];
         N - > Z1;
      }
      commands {
         Z1.select = yes
      }
   }
   rule up {
      pattern {
         N[select];
         Z2[!select];
         Z2 - > N;
      }
      commands {
         Z2.select = yes
      }
```

```
  }
}
rule copy_node {
  pattern {
    S[select = yes];
    V2[vb2 = yes];
  }
  without {
    X[select = yes];
    S > X;
  }
  commands {
    del_feat S.select;
    add_node S_copy: < V2;
    append_feats S == > S_copy;
    add_edge S - [copy: yes] - > S_copy;
    del_feat S.subj
  }
}
rule copy_edge {
  pattern {
    A1 - [copy: yes] - > A2;
    B1 - [copy: yes] - > B2;
    e: A1 - > B1
  }
  without {
    f: A2 - > B2;e.label = f.label
  }
  commands {
    add_edge f: A2 - > B2;
    f.label = e.label;
  }
}
rule del_copy_edge {
  pattern {
    c: A1 - [copy: yes] - > A2
  }
  commands {
    del_edge c;
  }
}
rule add_nodes {
  pattern {
    S2[subj = yes];
    V2[vb2 = yes];
    S1[sujet = yes, !subj];
    V1[vb1 = yes];
    V2 >> S2;
    S1 << V1;
```

```
  }
  commands {
    add_edge V2 - [ARG1: NEQ] - > S2;
    add_edge V1 - [ARG1: NEQ] - > S1;
    del_feat S2.subj;
    del_feat S1.sujet;
    del_feat S2.sujet;
    del_feat V1.vb1;
    del_feat V2.vb2
  }
}
strat subordination {
  Seq (subord, Onf(propagate), Onf(copy_node), Onf(copy_edge), Onf(del_copy_edge), add_nodes, )}
```

Paquets de règles de transformation des propositions relatives

```
% ------------------------------------------------------------------------------------------------------------
% RULE RELATIVES
% ------------------------------------------------------------------------------------------------------------

rule relative {
  pattern {
    V1[];
    V2[pos = v];
    V1 - [ARG2: NEQ] - > S;
    a: V2 - [ARG1: EQ] - > S;
    D - [RSTR: H] - > S;
  }
  without {
    D >> S  }
  without {
    D > S   }
  commands {
    add_node S2: < V2;
    append_feats S == > S2;
    add_node D2: < S2;
    append_feats D == > D2;
    add_edge D2 - [RSTR: H] - > S2;
    add_edge V2 - [ARG1: NEQ] - > S2;
    del_edge a  }}
```

Paquets de règles de transformation de la voix passive en voix active

```
rule pass {
   pattern {
      V[pos = "v"];
      n1: V - [ARG2: NEQ] - > N1;
      n2: V - [ARG1: NEQ] - > N2;
   }
   commands {
      N1.sujet = yes;
      N1.ob = yes;
      N2.objet = yes;
      N2.sub = yes;
      V.verb = yes;
      V.vb = yes;
      del_edge n1;
      del_edge n2;
   }
}
package propagate_sujet {
   rule down {
      pattern {
         N1[sujet];
         Z1[!sujet];
         N1 - > Z1
      }
      commands {
         Z1.sujet = yes
      }
   }
   rule up {
      pattern {
         N1[sujet];
         Z2[!sujet];
         Z2 - > N1
      }
      commands {
         Z2.sujet = yes
      }
   }
}
package propagate_objet {
   rule down {
      pattern {
         N2[objet];
         Z1[!objet];
```

```
        N2 - > Z1
      }
      commands {
        Z1.objet = yes
      }
    }
    rule up {
      pattern {
        N2[objet];
        Z2[!objet];
        Z2 - > N2
      }
      commands {
        Z2.objet = yes
      }
    }
}
package propagate_verb {
    rule down {
      pattern {
        V[verb];
        Z1[!verb];
        V - > Z1
      }
      commands {
        Z1.verb = yes
      }
    }
    rule up {
      pattern {
        V[verb];
        Z2[!verb];
        Z2 - > V
      }
      commands {
        Z2.verb = yes
      }
    }
}
rule right_limit {
    pattern {
      LR_selected[objet = yes];
    }
    without {
      X[objet = yes];
      LR_selected < X;
    }
    commands {
      LR_selected.last = yes;
```

```
      del_feat LR_selected.objet;
   }
}
rule permutation_verb {
   pattern {
      LR[last];
      V[verb];
   }
   commands {
      unorder V;
      insert V: > LR;
      del_feat V.verb;
   }
}
rule permutation_sujet {
   pattern {
      N[sujet];
      V[vb];
   }
   commands {
      unorder N;
      insert N: > V;
      del_feat N.sujet;
   }
}
rule add_edges {
   pattern {
      V[vb];
      N[ob];
      M[sub];
   }
   commands {
      add_edge V - [ARG1: NEQ] - > M;
      add_edge V - [ARG2: NEQ] - > N;
      del_feat V.vb;
      del_feat N.ob;
      del_feat M.sub;
   }
}
rule del_feats_last {
   pattern {
      Z[last];
   }
   commands {
      del_feat Z.last;
   }
}
rule del_feats_objet {
   pattern {
```

```
      Z[objet];
   }
   commands {
      del_feat Z.objet;
   }
}
strat passive {
   Seq(pass, Onf(propagate_sujet), Onf(propagate_objet), Onf(propagate_verb), right_limit,
Onf(permutation_verb), Onf(permutation_sujet), add_edges, Onf(del_feats_last), Onf(del_feats_objet))
}
```