Université de Montréal

De FrameNet à la Théorie Sens-Texte : Conversion et correspondance

Par

Hubert Corriveau

Département de linguistique et traduction, Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de maîtrise ès arts

en linguistique

Juin 2021

Université de Montréal

Unité académique : Département de linguistique et traduction, Faculté des arts et des sciences

Ce mémoire intitulé

De FrameNet à la Théorie Sens-Texte : Conversion et correspondance

Présenté par

Hubert Corriveau

A été évalué(e) par un jury composé des personnes suivantes

Marie-Claude L'Homme

Présidente-rapportrice

François Lareau

Directeur de recherche

Antoine Venant

Membre du jury

Résumé

Ce projet se décline en deux parties. Dans un premier temps, il s'agit de développer une méthode de conversion automatique des textes annotés selon la sémantique des cadres dans FrameNet en représentations sémantiques de la Théorie Sens-Texte, afin de développer davantage de ressources informatiques pour assurer le développement de différents projets, notamment le réalisateur de textes GenDR. Dans un second temps, cette conversion sera mise à profit pour effectuer une analyse comparative entre les deux formalismes. Nous retiendrons que ces formalismes ne sont pas incompatibles, mais diffèrent par leurs niveaux de granularité et leurs objectifs propres. Nous tracerons quelques parallèles entre les fonctions lexicales et les relations entre cadres, et proposerons une mise en commun des formalismes afin de les enrichir.

Mots-clés : Sémantique, Sémantique des cadres, Théorie Sens-Texte, GAT, Actants

Abstract

This project is divided in two main parts. Firstly, a method allowing for an automatic conversion of FrameNet's Frame Semantics-based text annotations into semantic representations, according to the Meaning-Text Theory framework, will be presented. This method will lead to an increased set of data usable to develop and improve various Meaning-Text Theory-based projects, including GenDR, a text realizer. Secondly, the conversion task will be used to do a comparative analysis of the two frameworks. We will conclude that the two frameworks are not incompatible, but differ in their granularity and goals. We will also draw parallels between the lexical functions and frame-to-frame relationships, and make some suggestions regarding changes to the frameworks in order to enrich them.

Keywords : Semantics, Frame semantics, Meaning-Text theory, NLG, Actants

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

TALN : Traitement automatique des langues naturelles

GenDR : Generic Deep Realizer

TST : Théorie Sens-Texte

RSém : Représentation sémantique

RSyntP : Représentation syntaxique profonde

Tf-Idf : Term frequency-Invert document frequency

SComm : Structure Communicative

ConceptR : Représentation conceptuelle

RSyntS : Représentation syntaxique de surface

Remerciements

J’aimerais remercier Mme Xiaobo-Ren, pour son implication philanthropique, qui a permis de financer ce travail, mon directeur de recherche, François Lareau, pour ses judicieux conseils, et ses recommandations, Marie-Claude L’Homme pour m’avoir permis de travailler dans un projet sur la sémantique des cadres, me faisant, au passage, découvrir FrameNet, mes amis et ma famille pour leur soutien, et finalement, tous les linguistes, philosophes, mathématiciens, psychologues et autres penseurs, cités ou non dans ce travail et qui ont chacun apporté une pierre à l’édifice de la connaissance.

1 Introduction

Lors d'une discussion, un message est formé dans l'esprit de l'un, qui veut véhiculer un message, puis l'articule que ce soit oralement, par signes gestuels ou encore par écrit, avant que l'interlocuteur ne le reçoive pour le décoder et retrouver le sens du message. Bien que le sens soit une notion intuitive, sa modélisation a été, et est encore, source de débats. Le sens peut être approché sous divers paradigmes, notamment à travers du prisme de la linguistique, ou encore de la psychologie. Ainsi, alors que Frege (1892) définit le sens d'un énoncé par les conditions nécessaires à ce que celui-ci soit vrai, (Lenneberg, 1967) associe plutôt le sens d'un énoncé à l'évocation de divers prototypes psychologiques émanant d'une conceptualisation du monde. Si la conception formelle du sens varie légèrement d'une discipline à une autre, la notion reste pourtant la même : celle de la conceptualisation du monde véhiculée par la langue.

Dans ce travail, il sera question d'effectuer automatiquement la conversion entre des phrases annotées sémantiquement et syntaxiquement, issues du projet FrameNet (Baker et coll., 1998) se basant sur la sémantique des cadres (Fillmore, 1976) en représentations sémantiques (RSém) de la Théorie Sens-Texte (Mel'čuk, 1997). Ces deux théories ont été établies parallèlement plutôt qu'en réaction l'une par rapport à l'autre et donc, cette tâche sera l'occasion de faire ressortir les contrastes et similitudes entre les deux formalismes. Finalement, les RSém qui seront produites par cette conversion pourront servir à améliorer GenDR (Lareau et coll., 2018), un réalisateur de texte assurant la réalisation d'énoncés associés à des RSém. L'automatisation de la tâche confère à cette analyse certains avantages. En effet, les méthodes automatiques utilisées dans ce travail permettent d'effectuer la conversion des structures de façon systématique, mettant ainsi en relief les différences pratiques dans le traitement des unités lexicales de chaque théorie. Ces différences seront mises de l'avant par l'analyse du processus de conversion ainsi que du résultat de la conversion. De plus, étant donné que le projet FrameNet est en développement continu, l'automatisation permettra de refaire une analyse comparative entre les théories lorsque de nouvelles données seront disponibles, sans nécessiter la répétition du travail d'analyse complet.

Finalement, cette conversion permet de produire rapidement, et à faible coût, un corpus de RSém, ce qui peut être utile dans le cadre de recherches futures sur la Théorie Sens-Texte. En effet, nous ne disposons pas, à l'heure actuelle, d'un corpus de taille raisonnable de RSém permettant de développer et de tester des systèmes comme GenDR

Ce mémoire est structuré de la façon suivante : tout d'abord, les formalismes seront présentés, séparément puis en contraste ; ensuite, la tâche de conversion d'un formalisme à l'autre sera décortiquée, une évaluation de la conversion sera présentée et finalement, les différences et similitudes entre les deux formalismes seront étudiées de plus près.

1.1 Le Traitement automatique des langues naturelles (TALN)

Puisque ce travail vise notamment à produire des RSém pouvant être utilisées dans un générateur de texte, une mise en contexte s'impose afin de saisir le champ d'études dans lequel ce travail s'inscrit.

Le TALN est un champ d'études à la croisée de la linguistique et de l'informatique, qui a pour but de mettre les techniques informatiques au profit du travail linguistique et vice-versa, que ce soit l'analyse du discours ou de données textuelles, la traduction automatisée, ou encore la génération de textes. D'un côté, il est possible de décoder les données textuelles pour en extraire des informations (l'analyse), d'un autre, il est possible de prendre des informations pour tenter de produire un texte par la suite (la génération ou la synthèse). Ces deux directions ont des visées distinctes et font appel à divers degrés à l'expertise linguistique et informatique. En effet, l'analyse fait appel à diverses techniques de reconnaissance de patrons afin d'extraire des informations. À titre d'exemple, il s'agit notamment des liens qui peuvent être faits entre différents concepts, ou encore de la caractérisation du style d'un auteur en particulier, en analysant ses choix lexicaux, ses thèmes principaux ou encore les différentes collocations qui lui sont privilégiées.

La synthèse fait appel à une variété de concepts linguistiques, afin de remplir sa mission de production du langage, notamment en utilisant les différents concepts linguistiques, afin de produire des énoncés grammaticalement corrects, qu'il s'agisse des règles morphosyntaxiques,

des règles de lexicalisation, ou encore de l'utilisation de la phonologie ou de l'orthographe (selon le médium). Il s'agit de l'angle d'attaque par lequel la TST considère d'abord le langage (Polguère, 2011).

1.1.1 Bref survol historique

Bien que dépendant des outils technologiques et de leur puissance de calcul, le TALN a une histoire qui remonte au début de l'informatique, et certaines méthodes utilisées sont encore plus anciennes.

Le mathématicien russe Andreï Markov a étudié les probabilités conditionnelles pour développer sa théorie à la base des études des processus stochastiques modernes. Pour ce faire, il dénombra les séquences de lettres et leur probabilité d'apparition suivant le contexte (Markov, 1913), ce faisant, il appliqua une méthode statistique à l'analyse de corpus, méthode encore appliquée de nos jours, connue sous le nom de modèles n-gramme, mais raffinée en comparant les mots entiers plutôt que les lettres, que les moyens de l'époque ne permettaient pas.

Le test de Turing (Turing, 1950) a comme objectif d'évaluer si une machine émule suffisamment bien l'intelligence humaine par le biais d'un dialogue. Ainsi, un humain face à un clavier doit dialoguer avec un participant humain ou une machine (en ignorant si c'est un humain ou une machine). Si la machine confond le participant humain dans 70 % des essais, le test est réussi. Ce test a fait grand bruit dans le domaine de l'informatique, en fixant un jalon dans la recherche en intelligence artificielle. Toutefois, une condition nécessaire à la réussite du test de Turing est de maîtriser tant l'analyse que la synthèse en TALN. En effet, la réussite du test est conditionnelle à la démonstration des habiletés langagières (Shieber, 2006).

Une tentative célèbre de développement de robot conversationnel est le système ELIZA (Weizenbaum, 1966), développé au MIT dans les années 60, qui prend la forme d'un ersatz de psychologue rogérien, posant des questions à son interlocuteur, et, à l'aide de patrons, reconnaît certains mots-clefs afin de demander à son interlocuteur de s'exprimer davantage sur un sujet. Bien qu'il ait rencontré un certain succès, ELIZA est loin d'avoir une capacité d'analyse linguistique suffisante pour tromper les participants humains, et encore moins une capacité de synthèse

adéquate, le système étant fait essentiellement pour générer des phrases telles que *pouvez-vous me dire plus au sujet de X* ou encore, *À quoi Y vous fait-il penser ?.*

1.1.2 L'analyse

En analyse de texte, différentes méthodes sont employées pour faire de la reconnaissance d'unités lexicales, de phrasèmes et autres. Parmi les différentes tâches d'analyse en TALN, notons la reconnaissance d'entités nommées [Named Entity Recognition, NER], qui tente de reconnaître les différents noms propres, et l'analyse syntaxique, qui consiste en l'analyse en constituants d'un énoncé. Différentes techniques peuvent être employées pour accomplir ces tâches, notamment l'utilisation de règles fixes (par exemple, avec des critères orthographiques pour la sélection des noms propres en français, ou encore l'utilisation de patrons connus, par exemple les déterminants introduisent un groupe nominal, etc.), ou encore des méthodes d'apprentissage profond, lesquelles se basent sur des données, au moins partiellement annotées dans un formalisme donné.

1.1.3 La génération de texte

La génération de texte est une tâche visant à produire un énoncé en langue naturelle à partir d'une entrée donnée. Parmi les différentes sous-tâches nécessaires à la génération de texte, il y a la planification du document, l'extraction de l'information utile, la structuration des données, des décisions à prendre au sujet du choix des expressions référentielles, et la réalisation (Reiter & Dale, 2000). Cette dernière tâche est la tâche plus linguistique de la génération de texte : elle consiste en la transformation de contenu sémantique donné en énoncés.

Différentes approches existent en réalisation de texte. L'approche retenue dans ce travail est celle adoptée par GenDR (Lareau et coll., 2018) : une approche symbolique stratifiée reproduisant les processus proposés par la TST. Tout d'abord, l'information doit être sélectionnée de façon appropriée et en lien avec le contexte discursif, puis les différents concepts doivent être planifiés dans leur aspect sémantique. Ensuite, à partir des sémantèmes, les éléments lexicaux doivent être sélectionnés pour que la syntaxe et la morphologie puissent faire leur travail de linéarisation de l'entrée, de mise en place des accords et de l'ordonnancement de l'énoncé cible.

La suite de la chaîne de génération de texte peut varier, selon la modalité visée. En effet, les différentes limites de la phonation, de l'audition et de la cognition chez l'humain imposent certaines contraintes si le but est de produire de la parole générée par ordinateur, tandis que les différentes conventions d'écriture viennent régir la suite des choses si la sortie visée était un texte écrit.

1.1.4 Generic Deep Realizer [GenDR]

Développé par Lareau et collègues (2018), GenDR est un réalisateur de texte profond. Dans les faits, il s'agit d'un logiciel effectuant le passage de la RSém vers la RSyntS, par un ensemble de règles et d'informations lexicographiques. Pour effectuer ce passage, GenDR prend en entrée la représentation sémantique dans un format textuel, puis un ensemble d'instructions sont appelées pour effectuer le passage de la RSém vers la RSyntP, puis de la RSyntP vers la RSyntS. Parmi les étapes clefs du passage de la RSém vers la RSyntP, il y a la lexicalisation, tirant notamment profit des fonctions lexicales.

Les différentes instructions font appel à des dictionnaires de type dictionnaire explicatif et combinatoire, pour capter les irrégularités de la langue cible.

Étant donné que chaque opération est assortie d'une règle de nature linguistique, il devient possible de tester à grande échelle différentes théories linguistiques, en générant diverses phrases à partir de RSém, puis en les soumettant à des tests de grammaticalité, ou autre. Toutefois, pour faire ce genre de test, il est nécessaire d'avoir sous la main suffisamment de données à soumettre en entrée. Or, dans l'état actuel des choses, il existe très peu de structures bien formées sous la forme de RSém. Ce travail contribue à combler ce manque, en effectuant une correspondance entre les annotations de FrameNet et les RSém de la TST.

2 Cadre théorique

2.1 Notations graphiques et conventions d'écriture

Dans ce travail, différents graphiques seront présentés, sous forme de RSém, ou encore des extraits d'annotation tirés du corpus à l'étude. Lorsque le formalisme importe, les nœuds de la sémantique des cadres seront présentés dans un encadré, avec, au-dessus, le lemme, et en dessous, le nom du cadre, comme à la figure ci-après, représentant une annotation de la sémantique des cadres, associant l'unité lexicale POMME au cadre sémantique *Food* :

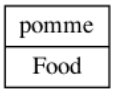


Figure 1. – Exemple d'une association entre une unité lexicale et un cadre

Les liens sémantiques seront représentés par des flèches, depuis une unité lexicale annotée jusqu'à ses dépendants, avec, à côté de chaque lien, l'élément de cadre dont il est question, comme à la figure suivante, représentant *Manger (une) pomme* :

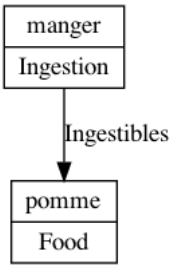


Figure 2. – Exemple de l'annotation d'une unité lexicale et d'un participant

Lorsqu'il s'agit d'un réseau sémantique, les nœuds sont représentés par des ellipses, et les liens présentent les actants sémantiques, avec un numéro :

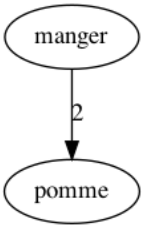


Figure 3. – Exemple de structure pour une RSém

Les exemples d'énoncés, les noms de cadres et de participant d'un cadre seront insérés en italique : *Ceci est un énoncé*

Lorsqu'un sémantème sera sujet de discussion, il sera encadré par des guillemets anglais simples : 'sémantème'

Lorsqu'une unité lexicale sera en cause, elle sera indiquée en PETITES MAJUSCULES

2.2 Sémantique des cadres

Issue des travaux de Fillmore et Baker, entre autres, la **sémantique des cadres** [Frame Semantics] est un paradigme d'étude sémantique et lexicologique associant à chaque unité lexicale un cadre conceptuel.

Selon Fillmore (1982, p.111), un **cadre** est défini comme étant un système de concepts tel que l'évocation d'un élément du système de concepts rend l'ensemble des éléments du système de concepts accessible.

> By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available.

Ainsi, la sémantique des cadres permet d'élaborer des cadres conceptuels, qui permettent de structurer une conception du monde en décrivant les prototypes de différentes catégories au sens de Rosch (1975), auxquelles des unités lexicales sont rattachées. Une fois ces associations établies, il est possible de faire des rapprochements entre différentes unités lexicales. En effet,

parmi les buts de ce paradigme de recherche il y a celui d'établir une ontologie, c'est-à-dire une classification de ce qui est (Guarino et coll., 2009), guidée par les cadres situationnels. Pour ce faire, un ensemble de cadres ont été élaborés par introspection, puis différentes tâches d'annotation ont permis de soulever les différents correctifs à apporter aux cadres, soit parce qu'ils sont trop vagues, trop précis, ambigus ou encore inexistants pour témoigner d'une réalité particulière.

À chaque cadre sont associés différents **participants**. La littérature parle parfois également d'élément de cadre (ou de *Frame Element* [FE]). Toutefois, pour des raisons arbitraires et esthétiques, nous retiendrons le terme « participant ». Le terme « actant » serait certainement approprié pour désigner cette notion, qui est analogue à celle d'actant sémantique en TST. Toutefois, afin d'éliminer de possibles confusions avec la notion d'actant sémantique en TST, nous retiendrons le terme « participant » pour désigner la notion de la sémantique des cadres, et « actant » pour la notion (quasi-) équivalente en TST. Les participants peuvent être conceptualisés comme un ensemble de paramètres présents lorsqu'un cadre est évoqué. Par exemple, lorsqu'un cadre désignant une action accomplie volontairement est évoqué, il y a la présence d'un participant désignant l'agent de l'action. Nous reviendrons plus en détail sur la notion de participant plus loin, à la section 3.1.2

FrameNet (Baker et coll., 1998) est un projet visant à établir les cadres, tout d'abord en les esquissant puis en les validant par le biais d'annotations de corpus de langue anglaise, ce faisant testant l'applicabilité du cadre théorique. Une description plus détaillée des données de FrameNet 1.6 sera faite à la section 4.1.

Différents projets secondaires sont apparus pour tester la validité des cadres et leur généralisation dans une perspective multilingue, et, bien que les résultats ne sont pour l'instant que préliminaires, il semble y avoir un certain potentiel de généralisation des cadres développés pour l'anglais à d'autres langues, principalement indo-européennes (Timponi Torrent et coll., 2018). À quelques reprises dans ce travail, des exemples en français seront associés à des cadres issus du projet original ; ces exemples ne sont qu'à titre indicatif, puisqu'il est possible que les

différents cadres proposés pour l'anglais ne soient pas parfaitement adaptés pour une autre langue.

Puisque les différents cadres sont d'une portée plus large qu'une simple unité lexicale, il est possible d'en tirer profit dans le domaine terminologique, particulièrement utile en contexte multilingue (Boas, 2005). Ainsi, la notion de cadre a été intégrée à différents domaines terminologiques, que ce soit le soccer (Schmidt, 2009) ou encore l'environnement, avec le projet DicoEnviro (L'Homme et coll., 2014), une base de données terminologiques multilingue qui repose aussi sur la Théorie Sens-Texte.

Dans ce programme de développement, les différents cadres sont reliés par des relations sémantiques. Par exemple, un cadre peut hériter des caractéristiques d'un autre, faire appel à une combinaison de cadres ou encore être un changement de perspective. Les relations de cadre à cadre sont définies ci-après.

La relation d'**héritage**, la plus importante (on dénombre 770 cas de cette relation au sein de FrameNet 1.6), est celle où un cadre hérite de toutes les caractéristiques d'un (ou de plusieurs) autre cadre, l'équivalent de l'hyponymie (ou de l'hyperonymie) entre deux cadres. En particulier, les participants du cadre parent doivent se retrouver d'une façon ou d'une autre parmi les participants de l'héritier. Par exemple, le cadre *Food*, désignant la nourriture, hérite du cadre *Physical_entity*, qui désigne une entité physique.

La relation d'**utilisation** [Using], la deuxième plus fréquente, avec 551 occurrences, réfère à la réutilisation d'une partie d'un cadre plus abstrait au sein d'un second cadre, sans qu'ils ne soient liés par une relation d'héritage. Par exemple, *Compliance*, qui désigne le fait de se conformer à une obligation, utilise *Obligation_scenario*, qui désigne l'obligation.

La relation de **perspective** [perspective_on], qui compte 119 occurrences, réfère au lien qui unit deux cadres, où le cadre annoté aborde la même situation qu'un autre cadre, mais avec un point de vue particulier. Par exemple, *Being_at_risk* et *Run_risk* sont deux cadres adoptant une perspective sur le cadre *Risk_scenario*. Ainsi, tandis que dans *Being_at_risk*, une chose positive est en danger d'être endommagée ou perdue, dans *Run_risk*, un protagoniste est dans une situation pouvant avoir un effet négatif sur sa personne. La différence entre les deux cadres étant

la façon d'aborder la situation (*Risk_scenario*), où un risque est encouru. Le protagoniste de Run_risk possède un attribut quelconque qui correspond à la chose référée par *Being_at_risk*.

La relation de **sous-cadre** [Subframe], avec 133 occurrences, dénote la relation entre un cadre et un cadre dénotant une partie de celui-ci. Par exemple, *Visit_host* dénote une situation où un visiteur se rend en un lieu associé à un hôte, avec un but, et a parmi ses sous-cadres *Drop_in_on*, désignant uniquement l'arrivée du visiteur chez l'hôte.

La relation de **précédence** [Precedes], avec 89 occurrences, dénote la relation de précédence habituelle entre deux cadres qui se suivent temporellement. Par exemple, *Sleeping*, qui dénote la situation de sommeil, a une relation de précédence avec *Waking_up*, qui dénote la situation du réveil.

La relation de **causativité** [Is_Causative_of], avec 60 occurrences, désigne une relation où un cadre est la cause d'un autre. Par exemple, *Cure*, qui désigne les procédures menant à la guérison d'une blessure ou d'une maladie quelconque a une relation de causativité avec *Recovery*, qui désigne la rémission ou la guérison d'une maladie ou d'une blessure.

La relation d'**inchoativité** [Is_inchoative_of], avec 19 occurrences, désigne une relation où un cadre désigne la phase initiale d'un autre. Par exemple, *Rotting,* qui désigne le fait de pourrir, est en relation d'inchoativité avec *Being_rotten*, qui désigne le fait d'être pourri, puisque lorsque quelque chose pourrit, la chose en question est pourrie *de facto.* Le tout peut être résumé en un tableau explicatif :

Tableau 1. – Résumé des relations entre cadres :

Relation (fr)	Relation (an)	Nombre d'occurrences	Exemple	
			Cadre 1	Cadre 2
Héritage	Inheritance	770	*Food*	*Physical_entity*
Utilisation	Using	551	*Compliance*	*Obligation_scenario*
Perspective	Perspective_on	119	*Being_at_risk*	*Run_risk*
Sous-cadre	Subframe	133	*Visit_host*	*Drop_in_on*
Précédence	Precedes	89	*Sleeping*	*Waking_up*
Causativité	Is_Causative_of	60	*Cure*	*Recovery*
Inchoativité	Is_Inchoative_of	19	*Rotting*	*Being_Rotten*

Il importe de mentionner que chaque unité lexicale se doit d'être désambiguïsée le mieux possible avant d'être associée à un cadre, ce qui permet de conserver certains contrastes, qui existent également en TST. Par exemple, le lemme FONDRE correspond à (au moins) deux unités lexicales : FONDRE$_{II.1}$,[1] désignant le fait, pour un solide, de devenir liquide, et FONDRE$_{I.1}$, causer que x fonde$_{II.1}$.

(1) Le beurre fond$_{II.1}$.

(2) Le soleil de midi a fondu$_{I.1}$ le beurre.

En sémantique des cadres, ces deux unités lexicales pourraient être associées à *Change_of_phase*, et à *Cause_change_of_phase*, respectivement. (leurs équivalents anglais, *melt* et *melt*, le sont)

> *Change_of_phase* : In this frame a Patient undergoes a change of phase. Note that this frame contrasts with Change_of_consistency in that this frame describes a change of a Patient between different phases (i.e. solid to liquid or frozen to "unfrozen").

[1] La numérotation choisie est celle du petit Robert

> *Cause_change_of_phase* : A Cause or Agent causes a Patient to undergo a change of phase. The Result of the change may be given, along with the Initial_state and the Circumstances under which the change can occur. Note that this frame contrasts with Cause_change_of_consistency in that this frame describes causation of a change of a Patient between different phases (i.e. solid to liquid or frozen to "unfrozen"

Dans les deux cas, un patient — le beurre — change de phase — de solide à liquide. La seule différence est la présence d'une cause (le soleil de midi) dans l'exemple (2), contrairement à l'exemple (1).

Finalement, la méthode de développement de la sémantique des cadres est basée sur l'annotation de corpus, dans le but de tester et de raffiner les différents cadres. Cette annotation massive de corpus nous sera utile, étant donné que certains textes ont été annotés sémantiquement et syntaxiquement en suivant les principes de la sémantique des cadres, et nous ont donc fourni un corpus particulièrement intéressant pour la linguistique informatique.

2.3 La Théorie Sens-Texte

Le but secondaire de notre travail étant de fournir des représentations sémantiques dans le formalisme de la Théorie Sens-Texte, il est approprié d'en faire un survol, et d'entrer davantage dans les détails pour certains aspects : la structure attendue des représentations sémantiques, ainsi que le nœud communicativement dominant, qui nous serviront pour effectuer la tâche de conversion.

2.3.1.1 Survol de la Théorie Sens-Texte

Découlant des travaux de Mel'čuk et de Zholkovsky dans les années 1960 (Milićević, 2006), à qui de nombreux linguistes se sont associés par la suite, la **Théorie Sens-Texte** (TST) a pour mission d'être un cadre théorique permettant l'élaboration de **Modèles Sens-Texte** (MST), qui sont des modèles des langues du monde (Mel'čuk, 1997), par une approche empirique. Il s'agit d'une modélisation de la production linguistique, en prenant comme point de départ le sens devant être véhiculé et comme point d'arrivée, un texte, qui est défini comme une suite de stimuli destinés à véhiculer le sens. Le texte est typiquement un texte écrit, ou encore de la langue orale. Cet aspect directionnel distingue d'emblée la TST et la linguistique générative, qui attribue un rôle

central à la syntaxe. Ainsi, contrairement au modèle en Y ou en 'T inversé' de la grammaire générative (Chomsky, 1993), qui suppose l'existence d'une syntaxe profonde pourvue d'une interface avec une syntaxe de surface occupant une place centrale, liée aux formes phonologiques et sémantiques, la TST propose un modèle directionnel, linéaire, où la syntaxe agit comme une représentation intermédiaire permettant de passer d'une représentation sémantique à une sortie linguistique intelligible, qu'elle soit sonore ou écrite. Pour ce faire, la TST implique un certain nombre de représentations intermédiaires depuis la représentation sémantique (RSém) : la syntaxe profonde (RSyntP), la syntaxe de surface (RSyntS), la représentation morphologique profonde (RMorphP) et la représentation morphologique de surface (RMorphS), pour finalement en arriver à la représentation phonologique profonde (RPhonP). Après la représentation phonologique profonde, il y a une forte dépendance à la modalité, qu'elle soit sonore ou textuelle, pour la production langagière.

À première vue, le modèle est complet ; toutefois, il ne faut pas omettre les facteurs communicatifs et pragmatiques permettant d'avoir une représentation sémantique. En effet, Mel'čuk postule l'existence d'une structure communicative en parallèle aux structures de la RSém et de la RSyntP, permettant alors de contraindre les sorties à ce qui est approprié dans le contexte discursif. De la structure communicative, nous utiliserons le concept de **nœud communicativement dominant**, qui sera introduit à la section 2.3.1.4 . De plus, pour générer une représentation sémantique, il est nécessaire d'avoir une certaine modélisation du monde, une partie de laquelle devant être communiquée. Mel'čuk la désigne comme étant la **représentation conceptuelle** (ConceptR). Cette représentation conceptuelle désigne l'ensemble des faits du monde connus du locuteur. Toutefois, cette représentation conceptuelle n'est pas formalisée (Mel'čuk, 2020). Cette modélisation du monde sera abordée dans la conclusion où un parallèle sera tracé entre la représentation conceptuelle suggérée par Mel'čuk et la notion de cadre sémantique.

Finalement, dans (Mel'čuk, 2001), il est fait mention qu'une RSém nécessite également au moins trois autres modélisations : une du locuteur, une du destinataire, et une de la situation pour être produite. Ainsi, le modèle, tel que postulé par Mel'čuk est le suivant :

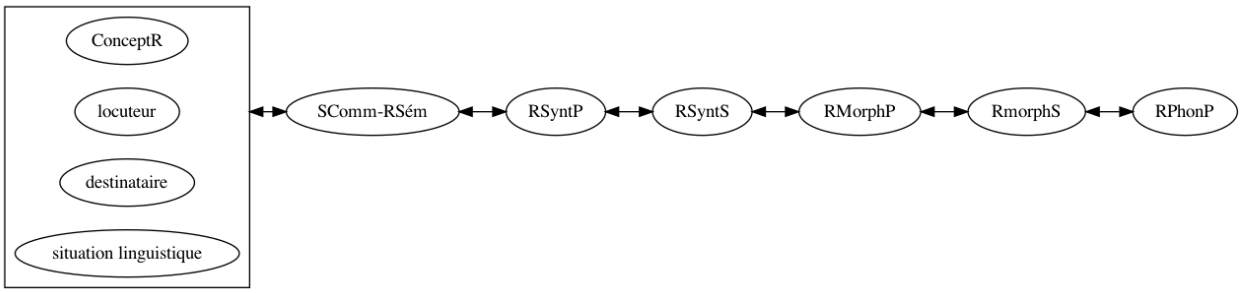


Figure 4. – Modèle Sens-Texte

2.3.1.2 La notion d'actant en TST

Dans les travaux en TST, différentes définitions équivalentes sont proposées pour définir la notion d'**actant**. Dans les sections qui suivent, nous résumerons celle de (Polguère, 2008), qui a comme avantage pour le présent exposé d'être énoncée en langue naturelle, contrairement à celle présentée dans (Mel'čuk, 2004), qui est plus technique et formelle, nous invitons nos lecteurs à consulter les travaux sur les actants de Mel'čuk pour une discussion plus approfondie sur la notion d'actant et sa formalisation.

2.3.1.2.1 Le sens lexical

Polguère (2008) définit le **sens** comme étant la propriété que partage une expression linguistique avec l'ensemble de ses paraphrases. Cette notion de sens dépend donc de la notion de paraphrase, qui est supposée transparente pour un locuteur, lequel peut, par le simple fait de connaître une langue, savoir si deux expressions sont équivalentes sémantiquement. Ainsi, les deux énoncés suivants :

(3) Paul a mangé une pomme

(4) Paul a mastiqué, avalé et digéré le fruit d'un pommier

sont jugés comme étant des énoncés véhiculant le même sens par les locuteurs du français, sans avoir à définir explicitement la notion de paraphrase.

2.3.1.2.2 Les sens liants et non-liants

Les unités lexicales sont dotées d'un sens, qui peut être de deux natures : liant, ou non liant. Un sens liant en est un dont l'information exprimée n'est pas complète sans l'apport d'autre sens. La plupart des verbes ont des sens liants, par exemple, le verbe *manger* appelle naturellement un mangeur et une chose mangée, sans quoi le sens n'est pas complet. Des noms peuvent également avoir un sens liant, par exemple les noms de relation, comme *mère*, qui n'est pas complet sans que la mère n'ait eu un enfant. Finalement, les modificateurs sont aussi habituellement des sens liants, comme *grand*, qui nécessite une chose pouvant être qualifiée par sa taille. En TST, un sens liant dénotant un fait est appelé **prédicat**, la plupart des verbes sont des prédicats, et un sens liant dénotant une entité, est appelé **quasi-prédicat,** comme dans le cas de *mère*.

Un sens non-liant est un sens qui est complet en soi. Par exemple, les noms propres comme *Mozart* ou *Jean-Paul Sartre* ont un sens non liant, étant donné qu'ils ne sont que des références à un élément du monde. Également, certains noms communs ont un sens non liant : *roche, air, étoile*, dénotant chacun des éléments du monde, puisque ces éléments n'ont pas, en soi, de propriétés ayant un lien de dépendance avec un autre élément. Les sens non-liants sont désignés en TST par le terme **nom sémantique**.

2.3.1.2.3 Les actants

Finalement, les actants sémantiques d'un (quasi-)prédicat sont les différents sens devant lui être liés afin qu'il soit complet. Par exemple, les actants de *manger* sont le mangeur et la chose mangée, et l'actant de *mère* est son enfant. Cette notion sera élaborée davantage à la section 3.1.1. Chaque actant occupe une position actancielle, dont on désigne le nombre par le terme **valence**.

2.3.1.2.4 Les fonctions lexicales

Certains sens liants dénotant un sens abstrait et général sont considérés de façon particulière par la TST, et sont décrits à l'aide de **fonctions lexicales** (Mel'čuk et coll., 1995), qui sont des fonctions ayant en entrée une unité lexicale et en sortie, un ensemble d'unités lexicales.

Plus concrètement, lors de la production d'un énoncé, ces sens liants sont réalisés linguistiquement par une unité lexicale ayant une relation privilégiée avec l'argument de la

fonction. Par exemple, la fonction lexicale d'intensification, notée MAGN, peut se combiner à diverses unités lexicales pour afin de dénoter la notion d'intensité, et avoir comme résultat des éléments aussi variés que les éléments d'entrée :

(5) MAGN(PEUR)= *bleue*

(6) MAGN(FORT)= comme un bœuf

Ces valeurs de sorties différentes dépendent de l'argument de la fonction. Ainsi, si le sens de BLEUE dans *peur bleue* est une paraphrase d'*intense*, il en est autre chose dans un contexte comme *peine bleue*, qui ne saurait être analysable par un locuteur. De la même façon, *peur comme un bœuf* ne serait analysable autrement qu'en tant qu'analogie entre la peur ressentie par celui éprouvant la peur et celle d'un bœuf, sans mention de son intensité.

De plus, d'autres fonctions lexicales n'expriment aucun sens, et ne sont utiles qu'à la réalisation syntaxique d'une unité lexicale. Il s'agit plutôt de fonctions lexicales renvoyant une unité lexicale sémantiquement vide. Par exemple, la fonction OPER_i renvoie un prédicat permettant la réalisation syntaxique de son actant sémantique *i* en tant que sujet, comme dans l'exemple ci-dessous,

(7) OPER_1(CRIME)= *commettre*

La fonction lexicale OPER_1 permet la réalisation en syntaxe de l'actant sémantique 1 de son argument. Ici, l'actant sémantique 1 de CRIME est le criminel, devenant le sujet sémantique de *commettre* lors de la réalisation syntaxique. Ainsi, *commettre* dans un énoncé du type *X commet un crime* n'a pas de réel sens.

2.3.1.3 Les représentations sémantiques (RSém)

Une **représentation sémantique** (RSém) est un graphe dirigé, reliant chaque unité de sens, appelée **sémantème**, à ses actants, numérotés en fonction de la valence des prédicats, selon le rôle qu'ils occupent dans le réseau sémantique.

2.3.1.4 Le nœud communicativement dominant

Parallèlement à la structure sémantique, il y a également une **structure communicative** (SComm), qui sert à définir l'organisation d'énoncés en tant que message. Cette structure est relativement complexe ; toutefois, seule la notion de **nœud communicativement dominant** au sein d'une RSém nous sera utile pour ce travail, et nous invitons le lecteur à consulter (Mel'čuk, 2001) pour une discussion approfondie des différents éléments de la SComm.

Selon (Mel'čuk, 2001), pour une paire de nœuds A et B, le nœud A domine directement le nœud B si le résultat de l'union linguistique[2] entre A et B est, sémantiquement, une hyponymie de A.

Par exemple, pour un syntagme comme *table rouge*, étant donné que l'union linguistique de TABLE et de ROUGE dénote d'abord une table, et non le fait que quelque chose soit rouge, on dira que TABLE domine communicativement ROUGE, même si, du point de vue sémantique, ROUGE est un prédicat ayant TABLE comme actant, alors que pour un syntagme comme *Paul dort*, DORMIR domine communicativement PAUL, puisque l'union linguistique de DORMIR et de PAUL est une instance de sommeil.

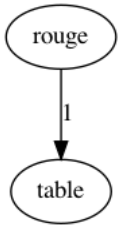


Figure 5. – *Table rouge*

Il importe de mentionner que le fait de dominer communicativement est indépendant du sens des relations sémantiques au sein d'une représentation sémantique, étant donné que pour un réseau sémantique comme illustré ci-dessous, autant 'augmenter' que 'population' peuvent être un nœud communicativement dominant, selon que le message véhiculé soit au sujet de la population (qui est croissante) ou encore de la croissance (de la population).

[2] En Théorie Sens-Texte, l'union linguistique est l'opération de base consistant à associer deux éléments ensemble

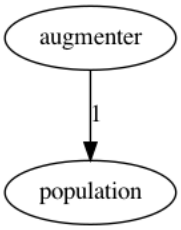


Figure 6. – *Croissance de la population ou population croissante ?*

De façon similaire, (Mel'čuk, 2001) définit un nœud comme dominant indirectement un autre s'il existe un chemin entre des nœuds tel que pour chaque nœud du chemin, un nœud domine l'autre, passant d'un nœud dominant à un nœud dominé à chaque fois. Finalement, un nœud en domine un autre s'il le domine directement ou indirectement.

Lors du passage de la RSém à la RSyntP, le nœud communicativement dominant est réalisé par le nœud racine de l'arbre ou du sous-arbre syntaxique généré (Polguère, 1990). Ainsi, si un nœud domine l'ensemble des autres nœuds d'une RSém, celui-ci sera la racine de l'arbre syntaxique généré.

Nous utiliserons cette notion afin de réduire des ensembles de sémantèmes reliés à l'élément communicativement dominant, dans les cas où l'information dont nous disposons est parcellaire.

3 Différences entre TST et sémantique des cadres

Outre le fait que ces deux formalismes n'opèrent pas au même niveau d'abstraction, la sémantique des cadres étant plus abstraite que la TST, certaines différences entre les formalismes sont bel et bien présentes, et seront approfondies ci-après.

3.1 Les actants et participants

3.1.1 Les actants et circonstants en TST

Tel qu'il a été mentionné précédemment, les actants sémantiques d'une unité lexicale sont les différents sens qui sont nécessaires pour avoir une signification complète d'un énoncé. Par exemple, pour *manger*, il y a nécessairement un mangeur et une chose mangée.

Mel'čuk (2004) fait une distinction entre deux types de participants pour la situation linguistique d'une unité de sens : obligatoire et optionnel.

Pour qu'un **participant d'une situation linguistique** soit **obligatoire**, il faut que son absence empêche l'utilisation de l'unité lexicale souhaitée. Par exemple, sans mangeur, il n'y a pas l'action désignée par *manger*, et sans nourriture, non plus. De la même façon, une mère qui n'a jamais eu d'enfant n'est pas une mère.

Un **participant d'une situation linguistique** peut aussi être **facultatif**, par exemple, dans le cas de 'manger', il est habituel d'employer un instrument quelconque, qu'il s'agisse des mains du mangeur, d'une fourchette, d'une paire de baguettes ou d'un autre instrument. Toutefois, l'instrument peut être omis entièrement sans que l'action de manger ne soit remise en cause, pensons notamment à un animal qui broute de l'herbe, ou un humain décidant de manger directement une pomme d'un arbre, sans la cueillir. Dans ces deux cas, l'emploi de l'unité lexicale MANGER est conforme à ce qui est attendu.

Aux participants obligatoires et facultatifs d'une situation linguistique sont associés, à condition de pouvoir être exprimés linguistiquement, des positions actancielles sémantiques. Bien que les actants soient toujours obligatoires, nous utiliserons les termes **actant obligatoire** et **actant**

facultatif pour désigner respectivement les actants correspondant aux participants obligatoires et aux participants facultatifs d'une situation linguistique, pour ne pas entretenir de confusion avec les participants en sémantique des cadres.

Bien qu'ils participent au sens d'un énoncé, les **circonstants**, qui agissent à titre de modificateurs, s'opposent aux actants, étant donné qu'ils ne peuvent prétendre à faire partie de la définition lexicographique d'une unité lexicale donnée. Ainsi, en sémantique, ils ont une structure distincte de celle des actants. Toutefois, la distinction en pratique entre actant et circonstant n'est pas toujours triviale (c.f. Mel'čuk, 2004)

3.1.2 Les participants en Sémantique des cadres

La sémantique des cadres admet la notion de **participant**, qui est réalisée linguistiquement par une unité lexicale jouant un rôle sémantique au sein d'un cadre, de façon analogue à un actant qui joue un rôle sémantique dans une situation linguistique au sens de Mel'čuk. Les participants sont divisés entre participant obligatoire [*core*]*,* périphérique [*Circumstancial*] et extrathématique [*extra-thematic*]. Les participants *core* ne peuvent pas être évacués d'un énoncé sans qu'ils soient présents de façon implicite. Par exemple, dans l'exemple *Je bois*, il y a nécessairement un liquide qui est bu ; même s'il n'est pas exprimé de façon explicite, il y a lieu de considérer que la chose bue existe.

Les **participants obligatoires** (*Core frame element*) : Les participants obligatoires ne sont jamais absents de la situation linguistique. Toutefois, ils peuvent être omis de l'énoncé, auquel cas ils sont présents, mais non mentionnés. On peut donc concevoir ces participants comme étant l'intersection des actants des prédicats appartenant à un même cadre sémantique, au sens de la TST (Mel'čuk & Milićević, 2020). En effet, selon Fillmore (1976, p.29) ,

> Comprehension can be thought of as an active process during which the comprehender – to the degree that it interests him – seeks to fill in the details of the frames that have been introduced, either by looking for the needed information in the rest of the text, by filling it in from his awareness of the current situation, or from his own system of beliefs, or by asking his interlocutor to say more

Ainsi, suivant les principes de la sémantique des cadres, lorsqu'une unité lexicale est mentionnée, un cadre est évoqué et les différents participants obligatoires doivent être présents dans le

message communiqué, que ce soit par une mention explicite, ou alors l'interlocuteur tentera d'associer à chacun des participants manquant un sémantème quelconque, en se fiant au contexte et à sa représentation conceptuelle du monde, la sémantique des cadres prévoyant trois cas de figure pour des omissions :

> L'élément omis peut être indéfini [*indefinite null instanciation*], par exemple, dans un énoncé comme *Myriam cuisine,* il est entendu que MYRIAM prépare des aliments afin qu'ils soient mangés ; toutefois, les aliments en soi ne sont pas définis.
>
> L'élément omis peut être défini par le contexte discursif [definite null instanciation]. Par exemple, dans un énoncé comme *Myriam a été embauchée,* ayant lieu dans le cadre d'une réunion des ressources humaines, le nom de l'employeur n'est pas explicitement mentionné, bien qu'il eût été possible de produire l'énoncé *Myriam a été embauchée chez Bio+ inc.*
>
> L'élément omis peut également l'être en raison de l'utilisation d'une construction [*constructional null instanciation*]. Par exemple, dans un énoncé comme *La salade a été mangée*, la personne ayant mangé la salade n'est pas explicitement mentionnée, étant donné l'utilisation de la voie passive.

Les **participants périphériques** (*Peripheral frame elements*) ne définissent pas un cadre et peuvent ainsi être omis. Par exemple, le lieu ou encore le temps sont des participants périphériques typiques (mais ne sont pas périphériques dans le cas d'un cadre référant directement au temps ou au lieu).

Les **participants extrathématiques** (*Extra-thematic frame elements*) quant à eux permettent de fournir un arrière-plan à un cadre et peuvent également être omis. Conceptuellement, ces participants relèvent d'un autre cadre qui permet de fournir un arrière-plan au cadre d'intérêt. Par exemple, un cadre comme *Using*, dénotant l'utilisation d'un instrument par un agent, a comme participant extrathématique *Containing_event*, qui désigne le contexte dans lequel l'instrument est utilisé par l'agent. En soi, ce participant n'a pas besoin d'être véhiculé par la situation linguistique, ni même d'être conceptualisé par les locuteurs, mais peut y être ajouté.

De la division des participants en trois catégories, on peut faire correspondre les participants obligatoires aux actants obligatoires, les participants périphériques aux actants facultatifs, et les participants extrathématiques aux circonstants. Cette correspondance sera discutée à la section 5.2.3.

3.1.3 Différences structurelles entre participant et actant

Selon Mel'čuk et Milićević, (2020) la relation de dépendance sémantique est une relation antiréflexive[3] et antisymétrique[4]. Toutefois, en sémantique des cadres, l'unité lexicale associée à un cadre peut elle-même être l'expression d'un élément du cadre évoqué, violant la condition d'antiréflexivité, ce qui démontre que la notion de participant est différente de celle d'actant sémantique. Une telle relation apparaît quasi systématiquement lorsqu'il s'agit de noms sémantiques ou de quasi-prédicats, étant donné qu'au moins un participant est évoqué par un cadre. Par exemple, le cadre *Weapon* désigne les armes, et a comme seul participant obligatoire *Weapon*, qui désigne l'arme en soi. Ainsi, une unité lexicale dénotant une arme aura généralement comme participant 'Weapon', l'arme en soi.

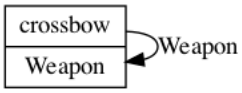


Figure 7. – Annotation réflexive en sémantique des cadres

En ce qui a trait à l'antisymétrie, il arrive parfois en sémantique des cadres que deux éléments soient reliés de façon symétrique, par exemple avec *enriched uranium*, où URANIUM est associé au cadre sémantique *Substance*, dénotant les substances, et qui a notamment comme participant facultatif *Descriptor*, une description de la substance. En retour, *Enriched* [enrichi] est associé au cadre *Degree_of_Processing*, désignant qu'un matériau a été transformé pour répondre à une fonction, et a comme participant obligatoire *Material* [matériau].

[3] Les deux éléments de la relation doivent être distincts
[4] Une relation entre deux éléments prévoit que la relation de sens inverse n'existe pas

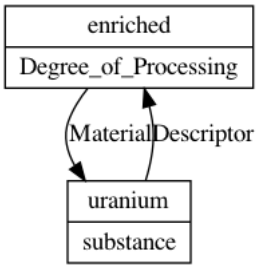


Figure 8. – Annotation symétrique en sémantique des cadres

3.1.4 Différences de statut (non) obligatoire

En comparant l'exemple du prédicat MANGER, présenté à la section 3.1.1, à ce qu'il en serait si l'on tentait une association avec les données de FrameNet, il est possible d'associer MANGER au cadre sémantique *Ingestion*, qui désigne le fait d'ingérer quelque chose, et qui a comme participants obligatoires *Ingestor* et *Ingestibles,* comme participants périphériques *Duration, Degree, Manner, Time, Source, Instrument, Means, Place* et *Purpose* et aucun participant extrathématique. En contraste, l'instrument est considéré comme un actant (facultatif) dans l'analyse de manger en TST alors que les autres participants périphériques seraient associés à des circonstants. Il y a donc une distinction entre ce qui est considéré comme étant obligatoire et facultatif entre les théories, puisque l'instrument, qui a un statut d'actant en TST, a le même statut, en sémantique des cadres, qu'un circonstant comme le temps.

Une autre distinction importante entre les deux théories réside dans la méthode employée pour établir la liste d'actants ou de participants dans les théories. Dans le cas de FrameNet, les cadres ont d'abord été esquissés par introspection, puis différents corpus ont été annotés selon les cadres établis par introspection pour valider et revoir la théorie, tandis qu'en TST, les différentes unités lexicales étudiées l'ont d'abord été de fond en comble, par introspection et analyse de corpus, une à une, afin d'avoir la description la plus juste et la plus précise de chaque unité.

Il importe de rappeler que les données de FrameNet sont issues principalement de tâches d'annotation partagées, et qu'une révision des cadres est faite de façon périodique. Par conséquent, il est possible qu'il y ait des erreurs dans la définition du cadre *Ingestion*, que ce soit

par la façon même dont le cadre a été esquissé, ou encore par suite d'erreurs humaines, comme il est possible d'en retrouver dans tous les formalismes.

3.1.5 Dénomination des relations sémantiques

Bien que ces deux formalismes partagent certaines ressemblances, la TST refuse de nommer explicitement la relation sémantique entre deux actants. En effet, selon (Mel'čuk, 2004), le fait de nommer les relations sémantiques est inadéquat, et témoigne d'un syntactocentrisme trop important. Toujours selon Mel'čuk, les rôles sémantiques doivent être numérotés selon la définition lexicographique, qui est elle-même construite en fonction de la saillance des actants. Ainsi, le sémantème MANGER est défini comme «X mange Y : X ingère Y dans le but de nourrir X» , et donc l'actant 1 est X, et l'actant 2, Y, tandis qu'en sémantique des cadres, une définition similaire, mais cette fois-ci pour un cadre, serait *l'agent X ingère l'aliment Y dans le but de se nourrir* avec comme rôles sémantiques *Agent* et *Patient*. Cette façon de faire est critiquée par Mel'čuk, étant donné qu'elle complique la définition, en introduisant de nouveaux (quasi-)prédicats.

Puisque les sémantèmes participant à une définition peuvent être remplacés par des sémantèmes sémantiquement plus simples (Aristote, c. 350 av. J-C/2015), voire jusqu'à des primitifs sémantiques (cf. Wierzbicka, 1996), il importe de limiter l'utilisation *ad lib* de noms d'actants, étant donné que la définition même d'agent, de patient, d'objet (et autre) sont des *definiens* complexes.

Toutefois, le fait d'associer un nom à un actant permet de tenir compte des similarités entre les différentes unités lexicales. En effet, en sémantique des cadres, les unités lexicales sont regroupées par cadre sémantique, et chaque cadre a un ensemble de participants obligatoires devant se retrouver comme participant pour chaque unité lexicale, en vertu d'une certaine similarité. De plus, ces dénominations se retrouvent non seulement dans un cadre précis, mais également dans tous les cadres qui en dépendent. Ainsi, la notion d'un actant X d'une unité lexicale Y au sein d'un cadre Z est généralisable aux autres unités lexicales participant du même cadre, ainsi qu'aux cadres qui héritent du cadre Z.

3.2 Modificateurs

Une distinction importante entre les deux formalismes est celle du traitement des modificateurs. En effet, en sémantique des cadres, certains modificateurs sont présents comme participants, alors qu'ils n'ont pas une valeur d'actant en TST. En effet, il peut être questionnable de considérer que, lorsqu'un cadre est évoqué, celui-ci comporte un participant *Descriptor*, associé à une description générale. Ainsi, tandis qu'en TST, *pomme rouge* serait représenté comme à la figure 9, en sémantique des cadres, on obtient une représentation bien différente, telle qu'illustré à la figure 10.

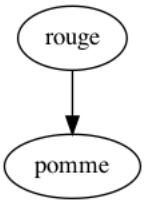


Figure 9. – *Pomme rouge*, en TST

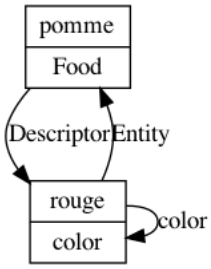


Figure 10. – *Pomme rouge,* en sémantique des cadres

En effet, la version de la sémantique des cadres inclut davantage de liens, notamment entre 'pomme' et 'rouge', violant alors la contrainte d'antisymétrie décrite par (Mel'čuk & Milićević, 2020). Cette différence peut toutefois être expliquée par la façon dont les cadres sont établis et révisés. En effet, étant donné que chaque cadre est évoqué par un grand nombre d'unités lexicales, il est possible qu'un élément *Descriptor* relativement vague soit pertinent pour un sous-groupe d'unités lexicales, qui partageraient une caractéristique particulière et ainsi formeraient un cadre en elles-mêmes, en relation d'héritage avec le cadre dont il était question.

Par exemple, si le cadre *Food* n'avait pas été prévu lors de l'étape d'introspection, les aliments se retrouveraient associés à *Physical_entity* (dont *Food* hérite). Or, en analysant l'ensemble des éléments annotés au sein de *Physical_entity*, il serait alors possible de réaliser qu'une unité lexicale comme NUTRITIOUS apparaît fréquemment comme participant, ce qui permettrait de supposer l'existence d'un cadre *Food*, et y placer tout ce qui semble approprié lors d'une révision des cadres. Ainsi, le traitement des modificateurs peut être vu comme un artefact du modèle de développement des cadres. Il n'est pas impossible que le traitement actuel des modificateurs laisse graduellement sa place à des représentations davantage compatibles avec le point de vue de la TST.

4 Implémentation

Maintenant que les formalismes ont été introduits et que leurs ressemblances et différences flagrantes ont été présentées, nous pouvons nous attaquer au cœur du travail : la conversion d'annotations en sémantique des cadres vers des représentations sémantiques en TST. Ce chapitre présente les détails de notre implémentation.

4.1 Données

Les données utilisées proviennent du corpus de langue anglaise du projet FrameNet, version 1.6 (Baker, 2015), plus spécifiquement les annotations complètes d'énoncés. En effet, parmi les annotations de FrameNet, Le projet FrameNet comprend 13 320 unités lexicales associées à un cadre, 1206 cadres, comptant chacun de 1 à 53 participants (dont 1 à 32 participants obligatoires), pour un total de 18 940 participants, 1263 participants ayant le même nom, dont 667 n'apparaissent que dans un seul cadre et 596 dans plus d'un cadre (jusqu'à 806 cadres différents pour *Time*). Les données sont présentées au format XML, et sont séparées en deux sources d'information : les définitions de cadre, ainsi qu'un ensemble d'annotations d'unités lexicales. Par exemple, nous retrouvons ci-dessous la définition du cadre *Part_Edge*, telle qu'elle est présente au sein de notre corpus (après avoir retiré certaines informations qui ne sont pas pertinentes pour ce travail, telles que les informations sur les annotateurs, les identifiants uniques associés au cadre et aux participants et les dates de modification). La liste des participants a été tronquée :

```
<frame>
<framename="Part_edge">
     <definition>
          <def-root>This describes a part_whole relationship where
          the <fen>Part</fen> provides the boundary between the
          <fen>Whole</fen> and what is not that object. The
          <fen>Part</fen> can exist along a continuum of width. It
          can be modified by an <fen>Orientation</fen> that
          specifies the particular portion of the edge of a
          <fen>Part_prop</fen> which describes a property of the
          <fen>Part</fen>.
               <ex>We came to the <fex
               name="Orientation">southern</fex> <t><fex
               name="Part">edge</fex></t> <fex name="Whole"> of
               the town </fex> .
               </ex>
          </def-root>
     </definition>
     <semType name="Transparent Noun"/>
     <FE coreType="Peripheral » name="Orientation">
          <definition>
               <def-root>This is a property of the <fen>Part</fen>
               that determines the portion of the <fen>Part</fen>.
            </def-root>
          </definition>
     </FE>
     <FE coreType="Core" name="Whole">
          (...)
     </FE>
     <frameRelation type="Inherits from">
          <relatedFrame>Part_whole</relatedFrame>
     </frameRelation>
     <lexUnit name="edge.n">
          <definition>COD:.the outside limit of an object, area,
          or surface.</definition>
          <sentenceCount annotated="104" total="390"/>
          <lexeme POS="N" name="edge"/>
     </lexUnit>
</frame>
```

Deux corpus sont disponibles : un grand corpus très faiblement annoté, où un petit nombre d'unités lexicales ont été annotées pour un grand nombre d'énoncés, et un petit corpus d'annotations complètes, pour lesquelles la quasi-totalité des unités lexicales sémantiquement pleines a été annotée, mais avec une quantité plus restreinte d'énoncés. Ce second corpus est

plus intéressant pour nous, puisque le but de ce travail est de produire des réseaux sémantiques de phrases, et pour ce faire, il faut des annotations les plus complètes possibles.

Le corpus d'annotations d'énoncés contient 9529 énoncés, provenant de différentes sources, notamment de l'American National Corpus (Macleod et coll., 2000), du projet PropBank (Kingsbury & Palmer, 2002), de Advanced Question & Answering for Intelligence project (National Institute of Standards and Technology, 2010), et quelques documents divers. Une description plus exhaustive du corpus peut être retrouvée en ligne[5]. Toutefois, ce corpus est peu représentatif de la langue orale, étant donné qu'il s'agit de sources écrites, et, par sa répartition des unités lexicales, est peu représentatif de la langue courante, en accordant notamment une place très importante à la géopolitique et à l'armement.

Pour chaque phrase, l'ensemble — ou presque — des unités lexicales sémantiquement pleines (cibles d'annotation) ont été associées à un cadre, et pour chaque participant, les informations suivantes sont fournies : les indices de début et de fin de la chaîne de caractères où est situé l'élément (que ce soit un mot-forme ou un syntagme), le nom du participant, la relation syntaxique entretenue avec la cible d'annotation, et s'il s'agit d'un pronom relatif ou d'un antécédent. Un exemple d'information retirée de FrameNet est disponible ci-dessous, pour l'annotation de l'unité lexicale DROVE, au sein de la phrase *His piety or his superstition drove him to try once more*, en ayant comme information que DROVE est associé au cadre Subjective_influence

	His	piety	or	his	superstition	drove	him		to	try	once	more
Élément de cadre	Situation					(cible)	Cognizer		Action			
Syntaxe	External						Object		Dependant			
Pronom relatif/antécédent												

[5] https://framenet.icsi.berkeley.edu/fndrupal/fulltextIndex

Ainsi, pour un énoncé tel que *He waited three days longer, and then his piety or his superstition drove him to try once more*, on obtient, en réunissant l'ensemble des unités lexicales annotées, une structure comme celle-ci-dessous, représentée graphiquement :

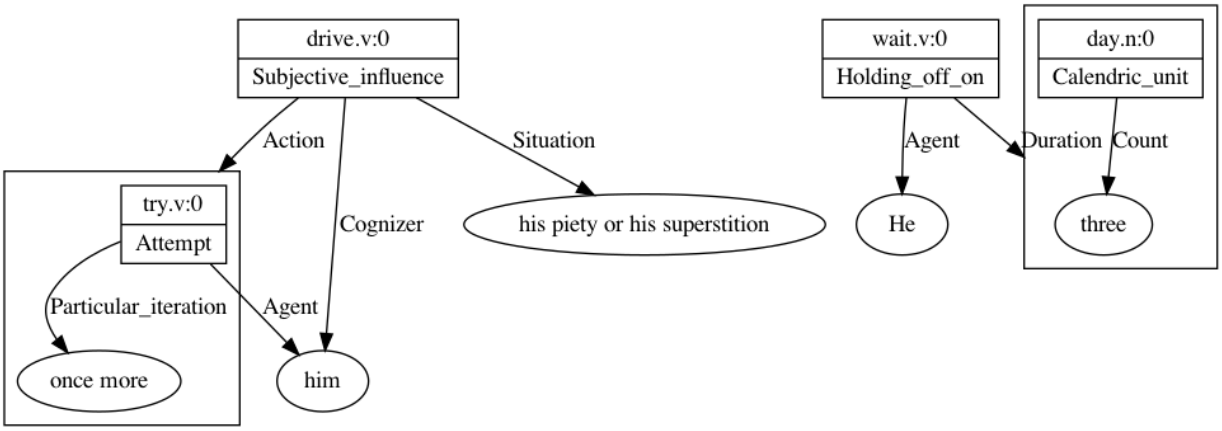


Figure 11. – Exemple d'annotation complète

De plus, les verbes supports sont indiqués dans les annotations des substantifs. Toutefois, les données n'indiquent que si un verbe est un verbe support, sans préciser davantage de quel type de verbe support il s'agit ; nous les avons traités comme des verbes sémantiquement vides et les avons ignorés dans le cadre de ce travail puisqu'ils ne correspondent à rien dans la RSém.

4.2 Choix méthodologiques

Étant donné que certains concepts sont analysés de façon variable d'un formalisme à un autre, ou encore d'une discipline à une autre, certains choix méthodologiques ont dû être faits et certaines définitions ont dû être élaborées afin de lever des ambiguïtés.

Les structures syntaxiques en TST sont décrites comme étant à la fois des arbres, et également des graphes dirigés (Milićević, 2006). Nous utiliserons le terme **degré entrant** (d'un nœud) pour désigner le nombre de liens qui pointent vers un nœud.

De plus, ces structures sont définies comme ayant un seul nœud ayant un degré entrant de 0, et les différents arcs du graphe représentent une relation de dépendance syntaxique, qui est définie comme étant antitransitive, antiréflexive et antisymmétrique. Ainsi, il s'agit d'un arbre enraciné,

puisque la condition de l'unique nœud avec un degré entrant de 0 empêche d'avoir deux nœuds parents n'ayant eux-mêmes aucun parent pour un nœud (figure 12) et l'antitransitivité de la relation de dépendance syntaxique ne permet pas d'avoir de nœud qui dépende de plus d'un nœud (figure 13)

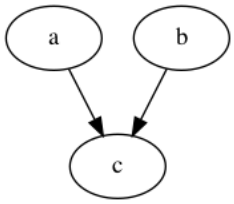


Figure 12. – Deux nœuds avec un degré entrant de 0

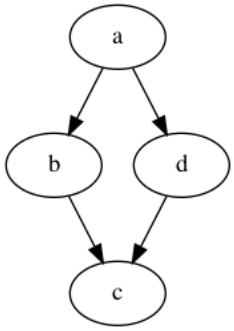


Figure 13. – Deux nœuds parents pour un même nœud.

Étant donné que les conjonctions ne sont normalement pas annotées en sémantique des cadres, elles ont dû être ajoutées par un algorithme spécifique au sein du script. Pour ce faire, nous avons identifié à l'aide d'expressions régulières et des informations sur les annotations si une conjonction était présente, ainsi que la liste des éléments joints. Une fois les conjonctions extraites, celles-ci ont été associées aux différents éléments réunis par celle-ci. Les seules conjonctions ajoutées manuellement sont *and* ['et'] et *or* ['ou'], qui sont classiquement considérées comme étant les équivalents en langue naturelle des fonctions booléennes de conjonction et de disjonction, toutes deux commutatives. Ainsi, l'ordre des actants importe peu, d'un point de vue strictement sémantique, bien qu'il y ait une différence du point de vue communicatif. D'autres conjonctions auraient pu être ajoutées manuellement, or, elles ne sont pas toutes commutatives, et donner un ordre risque d'introduire des erreurs au sein des

structures. En ce qui a trait aux conjonctions, nous utiliserons une structure telle que présente à la figure ci-dessous :

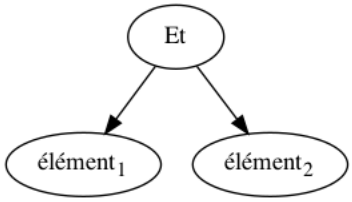


Figure 14. – Structure utilisée pour les conjonctions

Ces conjonctions peuvent être considérées comme des prédicats à deux arguments, en tant qu'équivalents linguistiques de la conjonction et de la disjonction au sens mathématique. (Mel'čuk, 1988) Toutefois, elles ont été traitées comme des prédicats à un nombre d'arguments indéfini. Pour ce faire, nous considérerons la présence d'une structure à plusieurs arguments comme étant une composition de prédicats binaires.

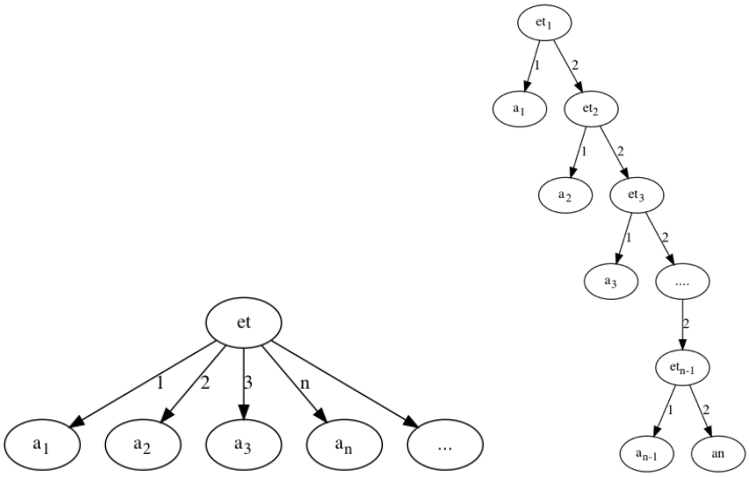


Figure 15. – Différentes formalisations du prédicat ET

De cette façon, les représentations graphiques des RSém produites sont plus simples à analyser, et la conversion d'une représentation illustrée à la figure 15 à l'autre est triviale.

Finalement, l'ajout des conjonctions permet de s'assurer d'avoir un nœud communicativement dominant selon notre algorithme. En effet, il arrive que l'annotation brute de la structure soit

analogue à celle en 16, avec un réseau sémantique aux composantes disjointes, plutôt qu'un réseau sémantique formé d'une seule composante. Ainsi, il est parfois nécessaire de reconnecter certains sous-graphes en corrigeant des structures. Comme on peut le voir dans la figure suivante, il arrive parfois que deux éléments soient coordonnés, et que ces éléments coordonnés agissent à titre de participant pour un autre cadre. Toutefois, si ces deux éléments ne sont pas déjà connectés entre eux, il est nécessaire d'ajouter la conjonction pour créer une structure adéquate.

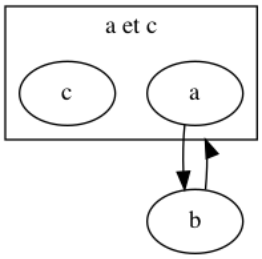


Figure 16. – Cas d'ajout de conjonction

Ainsi, cette figure illustre un cas où *a et c* forme un élément du cadre évoqué par *b* , mais pour laquelle aucune relation sémantique n'existe entre *a* et *c*. Après l'ajout de la conjonction, et la sélection du nœud communicativement dominant (présenté plus loin, à la section 4.3.1.3, p.52), on obtient une RSém ayant la forme d'un graphe connecté, comme illustré à la figure 17.

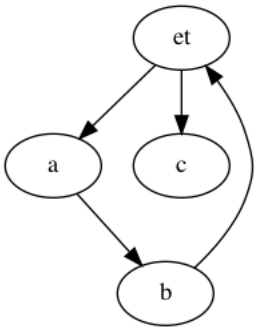


Figure 17. – Structure après ajout de la conjonction

4.3 Algorithmes

Pour effectuer le travail, un script a été fait en langage Python 3. Pour chaque phrase, les éléments suivants ont été extraits : le texte brut de la phrase, les différentes cibles qui ont été annotées : leur emplacement, leur lemme et leur cadre ; l'emplacement des participants au sein

du texte, avec leur fonction, ainsi que leur relation syntaxique avec l'élément lexical auquel ils sont reliés. Finalement une information supplémentaire présente a été utilisée : le lien entre pronom et antécédent.

Ensuite, pour chaque phrase, les annotations ont été extraites, ainsi que les informations relatives aux différents cadres qui sont utilisés. Une fois les annotations extraites, les différentes entrées ont été lemmatisées, puis mises en relation suivant les principes de la TST. Ces ajustements sont discutés dans la prochaine section. Une fois la structure finale produite, elle est transcrite dans un fichier en format lisible par GenDR. L'ensemble du code se retrouve en ligne sur GitHub[6].

Pour effectuer la conversion, pour chaque chaîne de caractères présente dans les annotations (que ce soit comme cible ou comme élément de cadre), une comparaison a eu lieu, afin de déterminer si une ou plusieurs autres chaînes présentes dans l'annotation de la phrase complète étaient imbriquées. Si plusieurs chaînes étaient imbriquées et que la chaîne comportait textuellement 'and' ou 'or', les différentes chaînes imbriquées ont été associées entre elles par une conjonction. Si plusieurs éléments étaient imbriqués, et que ceux-ci sont eux-mêmes reliés entre eux, nous avons présumé qu'une relation sémantique vers cette chaîne était équivalente à une relation sémantique vers le nœud communicativement dominant du sous-réseau formé par les éléments imbriqués. Si un seul élément était imbriqué, nous avons présumé qu'une relation vers la chaîne comportant l'unique élément était équivalente à une relation vers l'unique élément imbriqué de la chaîne. Par la suite, pour chaque cible d'annotation, un nœud a été créé dans la RSém, auquel ont été reliés tous les nœuds représentant un élément de cadre du cadre évoqué.

En somme, les données présentes en entrée, qui peuvent être représentées de façon simplifiée comme à la figure 18, sont transformées à la suite du travail de conversion en représentations similaires à ce qui est attendu en TST, tel qu'illustré à la figure 19

[6] https://github.com/HubertCorr/Memoire

Paul aime manger avec Marie.			
Unité lexicale annotée :	Cadre sémantique :	Éléments de cadre :	Relation syntaxique :
Aimer.v	Desiring	Experiencer : *Paul*	External
		Event : *Manger avec marie*	Object
Manger	Ingestion	Ingestor : *Paul*	External
		Co_part : *avec Marie*	Dep.

Figure 18. – Exemple de format des données

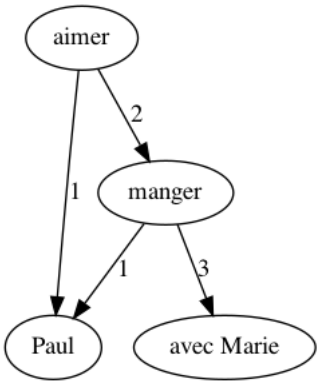


Figure 19. – Exemple de structure souhaitable

La transformation n'est pas parfaite, étant donné la présence du syntagme prépositionnel 'avec Marie' comme s'il s'agissait d'un sémantème. Toutefois, ceci est dû à la méthode d'annotation employée par FrameNet, qui prescrit l'annotation des unités lexicales ou des syntagmes dépendant de la projection maximale de la cible d'annotation, sauf pour quelques rares exceptions (par exemple, l'utilisation de verbe support ou de construction à montée) (Ruppenhofer et coll., 2016).

La présence du syntagme prépositionnel *avec Marie* s'explique également par le fait que l'unité lexicale MARIE n'a pas été annotée indépendamment dans le corpus, en sa qualité de nom propre, peu utile pour l'étude du langage.

En théorie, il semble possible de remédier au moins partiellement à de telles situations, puisque l'annotation précise le type de syntagme et que l'inventaire des prépositions au sein d'une langue

cible est limité. Ainsi, en établissant une liste des prépositions d'une langue donnée, il serait possible de les retirer de nos annotations. Toutefois, une telle technique est fortement dépendante de la langue de l'annotation, ce qui limite la possibilité de généraliser, et également, risque d'introduire des erreurs, étant donné qu'il faut faire une distinction entre les prépositions libres et les prépositions régies. En effet, certaines unités lexicales régissent l'usage de prépositions précises, qui ne font pas partie des RSém, puisqu'elles n'ont pas de réelle valeur sémantique. Par exemple, *faire part* régit la préposition DE pour exprimer ce dont il est question. Ainsi, inclure l'ensemble de ces prépositions régies, qui varient d'une unité lexicale à l'autre revient à établir un dictionnaire de régimes syntaxiques des lexies, ou en utiliser les données, ce qui est hors de la portée de ce travail, bien qu'il existe différentes ressources existantes, notamment VerbNet, dont les données pour l'anglais ont été intégrées à GenDR (Galarreta-Piquette, 2018).

4.3.1 Choix linguistiques et problèmes

4.3.1.1 Problèmes encourus

Pour effectuer le transfert depuis des annotations en sémantique des cadres vers une RSém, certains problèmes liés à l'annotation des données ont été rencontrés.

Tout d'abord, les annotations de FrameNet précisent, pour chaque cadre évoqué, quel est le lemme de la cible d'annotation et son emplacement dans la chaîne de caractère, ainsi que le nom de chacun des participants et leur emplacement au sein de la chaîne de caractère, sans que les participants ne soient lemmatisés. De plus, pour chaque participant, sa relation syntaxique avec la cible d'annotation est nommée. Ainsi, il a fallu extraire les lemmes associés à une chaîne de caractères, et à défaut, la chaîne de caractères elle-même, telle que présente dans le texte. De plus, lorsqu'un cadre est évoqué dans l'annotation de l'énoncé et que la cible d'annotation se retrouve au sein d'un participant d'un autre cadre, sans qu'il n'y ait d'autre annotation liée au participant, le participant du second cadre est remplacé par la cible d'annotation du premier, comme à la figure 20.

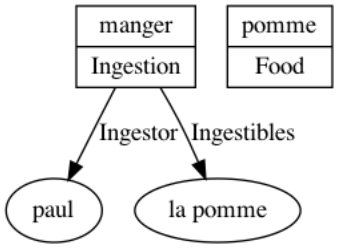


Figure 20. – Erreur d'encodage

Ici, POMME est la cible d'annotation du cadre *Food*, et *la pomme* est un participant du cadre *Ingestion*, contenant la chaîne de caractères associée à POMME, et il n'y a pas d'autre élément annoté dans la chaîne de caractères représentant le participant, le déterminant n'étant pas annoté séparément. Ainsi, le participant *Ingestibles* est remplacé par POMME, afin d'obtenir la structure illustrée à la figure 21.

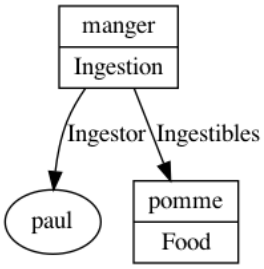


Figure 21. – Réparation de l'erreur d'encodage

Certains participants sont des syntagmes, et des éléments du syntagme peuvent également prendre part à des annotations, que ce soit en tant que cible ou que participant. Or, utiliser simplement la chaîne de caractères peut mener à des réseaux sémantiques erronés, comportant de nombreux doublons, comme l'illustre la figure 22, où *jouer* et *aux cartes* se retrouvent à la fois séparément et dans le nœud *jouer aux cartes*.

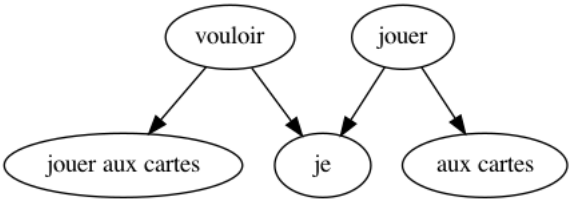


Figure 22. – Doublons causés par la chaîne de caractères

Afin d'éviter de tels cas de figure, chaque participant a été analysé afin de déterminer si une autre annotation était présente au sein du participant. Si aucune cible d'annotation ne se situe dans la chaîne de caractères, celle-ci est considérée comme correcte. Cependant, si une ou plusieurs cibles d'annotation se retrouvent au sein du participant, le participant a été remplacé par le nœud communicativement dominant de la structure sémantique définie par la cible d'annotation et ses participants présents dans la chaîne de caractères qui agit à titre de participant.

Ainsi, nous avons dû effectuer la sélection du nœud communicativement dominant à partir du peu d'information présente dans la structure. Plusieurs approches ont été considérées. Nous aurions pu, à l'aide d'une procédure itérative, attribuer un poids à chaque nœud, poids qui se verrait ajouté au poids du nœud duquel il dépend. Toutefois, une telle approche peut dépendre du nœud de départ, et nécessite que la structure sémantique soit acyclique pour s'assurer que le processus se termine. Or, les réseaux sémantiques ne sont pas nécessairement des arbres. De plus, dans le cas d'un arbre enraciné, le nœud avec le poids le plus important suivant cette procédure coïncide avec le nœud racine. Une preuve formelle est fournie en annexe.

Nous aurions également pu sélectionner le nœud ayant le plus grand nombre de liens. Toutefois, il n'est pas garanti que ce nœud existe, dès qu'il existe plus d'un nœud dans la structure.

Finalement, nous avons décidé qu'il valait mieux sélectionner la tête syntaxique du syntagme, puisque nous avons également accès à des informations de nature syntaxique, à défaut de pouvoir mieux analyser le tout. Cette façon de faire correspond aux propositions de Polguère (1990) et de Mel'čuk (2001) en ce qui a trait à l'arborisation d'une RSém. En effet, selon la TST, le nœud communicativement dominant se traduit généralement, en syntaxe profonde, par le nœud racine de l'arbre syntaxique associé. Ainsi, il est possible, dans la recherche du nœud communicativement dominant, de postuler l'hypothèse inverse : que le nœud racine de l'arbre syntaxique corresponde au nœud communicativement dominant de la structure sémantique. Ainsi, le nœud racine de la structure syntaxique associée a été choisi comme nœud communicativement dominant. Cette procédure a comme second avantage d'être économe au niveau des ressources informatiques, étant donné qu'il n'est pas nécessaire de parcourir la structure maintes fois pour retrouver le nœud communicativement dominant.

4.3.1.2 Résolution des propositions relatives

Au sein du corpus d'annotation de FrameNet, les pronoms relatifs sont annotés comme tels, et leurs antécédents sont également identifiés. Le guide d'annotation de FrameNet prescrit que tant le pronom relatif que l'antécédent doivent être annotés comme participant d'un cadre cible. Ainsi, un énoncé comme *a fire that had STARTED in the basement* sera annoté comme au tableau suivant (avec comme cible STARTED, associée au cadre *Catching_fire*).

	A	*fire*	*that*	*had*	*started*	*in*	*the*	*basement*
Participant		Fire	Fire		Cible	Place		
Antécédent/pronom relatif		Rel	Ant					

Tableau 2. – Annotation d'une relative

Pour éviter d'avoir des éléments sémantiquement vides au sein de la structure, l'annotation de *start.v* est transformée pour omettre le pronom relatif, et obtenir la structure sémantique suivante :

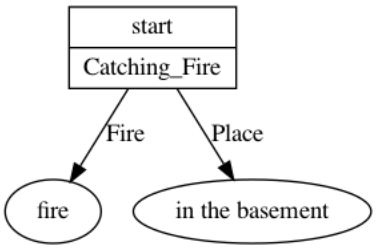


Figure 23. – Annotation après omission du pronom relatif

Or, les énoncés ne sont que rarement seuls et hors contexte. Dans ce cas-ci, la phrase complète annotée est la suivante *:*

> At about 8:30 (eastern time) firefighters were called to the home on 48 Granger Place to put out a fire that had started in the basement. [Vers 8 h 30 (heure de l'Est), les pompiers ont été appelés à la maison du 48 place Granger pour éteindre un feu qui a commencé dans le sous-sol]

L'annotation de *put out* [éteindre], associé au cadre : *putting_out*, qui dénote qu'un agent cause l'extinction d'un feu est présente ainsi en corpus :

	At about 8:30 (eastern time)	*firefighters*	*Were called to the home on 48 Granger Place to*	*Put out*	*A fire That had started in the basement*
Participant :		Agent		Cible	Fire
Syntaxe :		External			Object

Figure 24. – Annotation de l'énoncé complet

Une telle structure pose problème, étant donné la présence de l'antécédent et du pronom relatif au sein d'un groupe annoté, ce qui crée une structure imbriquée. Ces structures imbriquées doivent faire l'objet de la sélection du nœud communicativement dominant afin d'avoir une RSém.

4.3.1.3 Nœud communicativement dominant

Pour effectuer la sélection du nœud communicativement dominant, l'algorithme sélectionne, pour chaque sous-graphe sémantique, le nœud qui est la racine de l'arbre syntaxique correspondant. Cet algorithme fonctionne de manière récursive pour traiter les structures ayant comme participant un ensemble de nœuds, par exemple lorsqu'un participant est une structure complexe. Pour ce faire, les annotations syntaxiques sont extraites de nos données, puis seuls les nœuds faisant partie du sous-graphe sont sélectionnés, afin de limiter l'analyse syntaxique à un sous-graphe.

Par la suite, on établit la liste des relations syntaxiques et sémantiques, et en s'appuyant sur le fait qu'on a déterminé le nœud communicativement dominant de chaque structure complexe. Finalement, le nœud communicativement dominant sélectionné par l'algorithme est celui qui est le nœud racine de l'arbre syntaxique, déterminé en sélectionnant le nœud ne dépendant d'aucun autre nœud au sein de la structure syntaxique, à condition qu'il soit unique. La condition d'unicité n'est pas nécessairement remplie, les annotations étant fréquemment incomplètes, lorsqu'il n'est pas possible de faire la sélection par le seul critère de la sélection du nœud non dominé syntaxiquement, l'algorithme se rabat sur les informations sémantiques. Ainsi, pour départager

entre plusieurs candidats au titre de nœud communicativement dominant, celui qui est choisi est le nœud qui ne dépend d'aucun autre nœud en sémantique, et, à défaut, le nœud ayant le plus grand nombre de dépendants sémantiques. Si l'ensemble de ces critères ne permet pas de déterminer un nœud unique, parmi les nœuds répondant au plus grand nombre de critères, un nœud est choisi au hasard.

Une fois la sélection des nœuds communicativement dominants de chaque sous-graphe, faite, une redirection a été établie, afin de rediriger toute relation pointant vers le sous-graphe entier en une relation pointant vers le nœud communicativement dominant. Par exemple, pour une structure d'entrée comme ci-dessous,

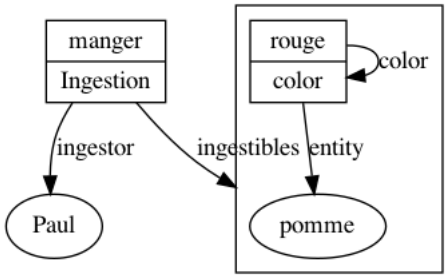


Figure 25. – *Pomme mange la pomme rouge* : structure d'entrée

Après la sélection du nœud communicativement dominant ('pomme') du sous-graphe *pomme rouge*, les relations sémantiques de 'manger' sont mises à jour :

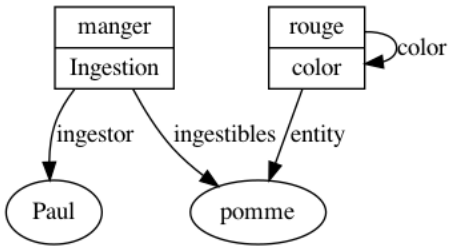


Figure 26. – *Pomme mange la pomme rouge* : représentation intermédiaire

4.3.1.4 Numérotation des actants

Un élément clef de la sémantique en TST est la numérotation des actants d'une unité lexicale, ce qui est *a priori* incompatible avec la sémantique des cadres, qui ne numérote pas ses participants, mais les désigne plutôt par le rôle qu'ils jouent au sein du cadre d'intérêt.

Pour passer de cette désignation à la représentation en TST, quatre avenues ont été considérées. Une première approche est d'évaluer manuellement chaque lexème associé à un cadre, et mettre en place une heuristique associant aux différentes unités lexicales un lien entre les rôles sémantiques et leur numéro d'actant. Une telle approche ne peut vraisemblablement être réalisée que manuellement, pour les 13 320 unités lexicales présentement associées à un cadre, et ne permet pas de mettre à jour aisément la transposition entre sémantique des cadres et TST dans le cas de révision ou d'ajout de cadres ou d'unités lexicales. Compte tenu de la lourdeur de la tâche, cette avenue a été jugée irréaliste.

Ainsi, une seconde approche faisant appel à leur nom de participant a été envisagée. Cette approche présupposait que l'on puisse savoir, en toute généralité, quel numéro devrait être le plus souvent associé à un participant. Toutefois, une telle avenue n'est pas praticable, étant donné que les participants ont des noms ambigus, par exemple, *degree* est défini d'au moins trois façons différentes, selon le cadre auquel il est rattaché (Accuracy, Aesthetics, ou encore Quantified_mass) :

> Degree (Accuracy) : How closely the expected or actual location that the Instrument interacts with corresponds to the Target_value.
>
> Degree (Aesthetics) : A modifier expressing the deviation of the implicit value either from the norm or from the value for another Entity.
>
> Degree (Quantified_mass) : This frame element selects some gradable attribute and modifies the expected value for it.

De plus, ce participant a une importance différente d'un cadre à l'autre : s'il est très important pour un cadre comme *Accuracy*, qui s'intéresse à la précision (au sens large), *Degree* a bien moins d'importance dans un cadre comme *Quantified_mass*, où cet élément agit comme modificateur d'une expression.

La désambiguïsation est certes possible, mais même dans la situation — hypothétique — où les participants ne seraient pas ambigus, il ne serait toujours pas possible d'attribuer un ordre relatif aux 987 participants existants sans faire appel à l'arbitraire, ou sans avoir déjà un ordre relatif entre les participants.

Une troisième approche a été d'établir un ordre de base pour l'ensemble des cadres : pour chaque cadre et pour chaque participant, lorsque le cadre est évoqué dans le corpus, un poids de précédence est attribué aux participants selon leur ordre d'apparition eu sein de l'énoncé. Ensuite, les poids de précédence pour chaque paire de participants sont additionnés au travers des cadres, dans le but d'obtenir une sorte d'ordre de précédence moyen.

Toutefois, cette approche est inadéquate pour décrire les participants des cadres ayant une relation de changement de perspective (voir section 2.2), étant donné que deux cadres ayant une relation de perspective auront des participants dont la saillance est différente (par exemple, dans *Selling* et *Buying*, les rôles du vendeur et de l'acheteur sont intervertis et par conséquent, ne devraient pas être numérotés de la même façon). Ainsi, un tel ordonnancement a des défauts majeurs auxquels il est difficile, voire impossible de remédier. En effet, cette procédure suppose que la relation d'ordre entre les participants est quasi-universelle, tout en supposant que cet ordre diffère d'un cadre à l'autre. De plus, la supposition que les ordres diffèrent d'un cadre à l'autre, tout en cherchant l'ordre le plus près d'un universel, pose la question de la valeur de cette procédure.

Également, cette analyse se base sur les différents cadres, qui varient grandement en termes d'usage : on peut supposer raisonnablement que le cadre *Cooking_creation*, qui désigne essentiellement toute chose cuisinée soit plus utilisé qu'un cadre comme *Weapon*, désignant quelque chose qui est une arme. Toutefois, une telle supposition est contraire aux données de notre corpus, qui recense 3 occurrences de *Cooking_creation* contre 968 occurrences de *Weapon*. Étant donné la taille relativement petite du corpus, il n'est pas étonnant qu'il ne soit pas représentatif du langage général, et puisque peu de corpus peuvent prétendre être quelque peu représentatifs de la langue parlée, nous ne pouvons donc nous fier aux fréquences relatives des

cadres pour définir un ordre, mais seulement sur un *a priori*, soit que les différents cadres peuvent être évoqués dans la langue courante.

Une quatrième approche fut donc retenue. Elle considère une numérotation des participants qui diffère pour chaque cadre, en prenant comme structure l'ordre d'apparition des rôles sémantiques en fonction de leur ordre d'apparition dans la définition textuelle du cadre. Ainsi, les participants du cadre *Event*, défini ci-dessous, auraient comme correspondance TST-Sémantique des cadres : 1-Event, 2-Place, 3- Time.

Event : An **Event** takes place at a **Place** and **Time**.

Une telle façon de procéder a comme avantage non négligeable qu'il est alors possible de produire l'ordre d'une façon semblable à la pratique habituelle en TST pour la numérotation des actants. En effet, en TST, la description lexicographique place habituellement les différents actants en ordre de saillance cognitive, les plus importants d'abord, ce qui correspond à l'ordre de la numérotation des actants, tandis que la méthode proposée pour la numérotation des actants en se basant sur la définition du cadre suppose qu'il s'agit de ce qu'il y a de plus proche d'une description lexicographique, ce qui semble être le cas. Notons toutefois que cette méthode est fondée sur l'hypothèse que l'ordre linéaire syntaxique en anglais correspond à la saillance des actants, permettant alors de les numéroter en ordre d'importance.

Par contre, cette méthode comporte certains inconvénients. En effet, dans la description des cadres, il n'est pas toujours possible de déterminer l'ordre de l'ensemble des participants, certains actants n'étant pas toujours mentionnés dans la description du cadre. Ainsi, une procédure a été élaborée pour obtenir une liste plus complète des éléments, à l'aide d'une stratégie de propagation. Deux méthodes sont employées pour extraire la liste des éléments dans la définition. Tout d'abord, dans certains cadres sémantiques, les noms des participants sont encadrés par des balises XML `<fen>` et `</fen>`, permettant de délimiter clairement les noms des participants, tel qu'illustré dans la présentation des données de cadre au format XML (p.40). Toutefois, bien que ces balises délimitent sans équivoque les participants, elles ne sont utilisées que dans environ 1 % du corpus. Ainsi, une seconde stratégie a été mise en place pour pallier ce manque de données : il s'agit d'effectuer une recherche dans le texte de la définition formelle

des différents cadres, afin de détecter la première mention de chaque participant, et de déclarer cet ordre comme étant le meilleur.

Finalement, pour combler les dernières lacunes existantes, l'ordonnancement des actants d'un cadre ainsi obtenu sera utilisé pour ordonner les actants des cadres en relation d'héritage. En effet, dans chaque paire de cadres parent-héritier, les participants obligatoires du parent doivent se retrouver parmi les participants de l'héritier. Toutefois, les participants n'apparaissent pas toujours sous le même nom, ceux-ci apparaissant parfois sous un nom plus précis, par exemple le participant nommé *Agent* dans un cadre apparaît sous les noms *Helper, Cognizer, Executive_Authority,* ou encore *Perpetrator*, selon qu'il s'agisse du cadre *Intentionnaly_act, Assistance, Choosing, Clemency* ou encore *Piracy*. Une procédure, décrite à la section 4.3.1.5 a été établie pour effectuer la concordance entre le nom des participants dans un cadre parent et de ses homologues dans le cadre héritier. Une fois cette concordance faite, la procédure décrite ci-après permet d'insérer les participants du cadre parent dans le cadre héritier.

Pour une paire de cadres, avec un cadre parent et un cadre héritier, les participants du cadre parent sont associés à leurs homologues du cadre héritier, selon l'algorithme d'association des participants homologues. Au sein des cadres parent et héritier, tout ou partie des participants sont ordonnés, de la façon décrite plus haut. Chaque participant ordonné du cadre parent qui n'est pas ordonné dans le cadre héritier est placé à la dernière position dans le cadre héritier qui suit l'ensemble des participants ordonnés du cadre parent au sein du cadre héritier, mais avant les éléments qui le suivent au sein du cadre héritier, lorsque possible.

Par exemple, si, après l'extraction de l'ordre pour une paire parent-héritier, on a l'ordre au tableau suivant :

	Ordre présent				Éléments non ordonnés
Cadre parent	A	B	C	D	
Cadre héritier	A	C	K		B,D

Tableau 3. – Exemple d'ordre après extraction de l'information

Lors de l'étape de propagation, l'élément *B* du cadre héritier sera inséré après *A*, qui le précède dans le cadre parent, mais avant *C*, qui le suit. L'élément *D* sera quant à lui inséré à la toute fin, étant donné qu'il suit *A*, *B* et *C* dans le cadre parent, et qu'aucun élément ne le suit dans le cadre héritier, nous laissant avec l'ordre suivant :

	Ordre après propagation				
Cadre parent	A	B	C	D	
Cadre héritier	A	B	C	K	D

Tableau 4. – Exemple d'ordre après propagation du cadre parent dans le cadre héritier

Cette procédure est appliquée pour chaque paire de cadres parent-héritier, depuis les cadres n'ayant pas de parents, et en descendant l'arborescence. Étant donné qu'il y a des cas d'héritage multiple (un cadre pouvant hériter de deux parents), la procédure est répétée jusqu'à ce que la procédure n'ait plus d'effet, l'ensemble des participants pouvant être ordonnés l'étant.

Une fois cette étape d'établissement de l'ordre accomplie, la procédure est refaite, mais à l'envers, en utilisant les participants ordonnés dans un cadre héritier qui ont une correspondance dans le cadre parent. En effet, il peut arriver que des éléments d'un cadre héritier qui ont un correspondant dans le cadre parent soient ordonnés au sein d'un cadre héritier, mais pas dans le cadre parent. En répétant la procédure dans l'ordre inverse, on peut alors espérer ordonner le maximum de participants.

Finalement, les participants restants au sein de chaque cadre sont ajoutés à la suite, dans l'ordre d'apparition des participants dans la liste des participants au sein des données XML de FrameNet.

En résumé, pour chaque cadre, les éléments de cadre ont été placés en ordre selon leur ordre d'apparition dans la définition du cadre, lorsque possible, sinon l'ordre obtenu avec le cadre parent a été utilisé, puis l'ordre obtenu avec les différents cadres héritiers. Finalement, si après toutes ces opérations, certains participants n'étaient toujours pas dans un ordre précis, ceux-ci ont été simplement placés à la suite des éléments ordonnés.

4.3.1.5 Association d'éléments de cadre homologues

Étant donné que, pour un participant d'un cadre parent, celui-ci peut se retrouver sous divers noms parmi les cadres qui héritent de celui-ci (voir section précédente, 4.3.1.4), il est nécessaire d'établir une certaine correspondance entre un participant d'un parent, et un autre d'un cadre héritier.

Il aurait été envisageable *a priori* d'utiliser l'algorithme de Needleman-Wunsch (Needleman & Wunsch, 1970) pour faire un travail d'alignement. Toutefois, cet algorithme implique d'attribuer manuellement des pondérations pour les différentes pénalités d'insertion, de suppression et de substitution, pour chaque substitution possible. Étant donné les 987 participants possibles, il s'agit alors d'attribuer une pondération pour les 486 591 possibilités de substitution, les 987 possibilités de non-substitution et une pénalité pour suppression, ce qui s'avérait irréalisable.

Pour effectuer la concordance entre les participants du cadre parent à ceux de l'héritier, trois modèles ont été testés sur un sous-corpus de l'ensemble des paires parent-héritier, et évalués en ce qui a trait à la précision de l'association, en se comparant à une association manuelle pour 75 paires de cadres parent-héritier. La précision est calculée comme étant le rapport entre le nombre de participants bien associés pour chaque cadre et le nombre de participants. À titre comparatif, l'espérance de la précision du hasard pur et la précision d'un modèle trivial ont été évaluées.

Le modèle de hasard est défini comme étant un ordre attribué par le biais d'un tirage sans remise pour élément de la liste des participants, de façon uniforme. Les détails du calcul de l'espérance mathématique du nombre de bonnes associations du modèle en fonction du nombre de participants dans chaque cadre se retrouvent en annexe, et ce calcul nous donne une précision

évaluée à 12 %. Ce modèle n'est toutefois pas réaliste, étant donné qu'un modèle trivial peut être élaboré avec une précision bien meilleure.

Le modèle trivial suppose que deux participants ayant le même nom sont associés et se fie au hasard pour le reste. Celui-ci nous donne une précision sur le corpus de test de 59 %. Cette précision est évaluée en dénombrant les participants ayant le même nom, additionnée de l'espérance du modèle de hasard pur, calculée en prenant comme nombre de participants, pour chaque cadre, le nombre de participants moins le nombre de participants ayant le même nom.

Le premier modèle évalué est un modèle vectoriel par plongement de mots ayant été entraîné à l'aide de la bibliothèque Gensim (Řehůřek & Sojka, 2010), sur un corpus très limité constitué des descriptions textuelles des participants dans les descriptions de cadres de FrameNet, après le prétraitement suivant : suppression des entrées en double, uniformisation de la casse, suppression de tout caractère différent d'une lettre, d'un chiffre, du trait d'union ou du trait de soulignement, et suppression des espaces en double. Par la suite, pour chaque participant au sein du cadre parent, la liste des participants homologues potentiels du cadre qui en hérite est produite, puis parmi cette liste, la similarité cosinus entre les différents termes est utilisée pour sélectionner le terme le plus semblable.

Ce modèle a une précision de 55,5 %, tandis que le hasard pur donne une précision de 12 % et le modèle trivial, 59 %. Les deux autres modèles testés font usage de la méthode d'analyse Tf-Idf, qui est discutée ci-après.

4.3.1.5.1 Tf-Idf

Term frequency-invert document frequency [ci après Tf-Idf] est une méthode d'analyse permettant d'obtenir une représentation vectorielle d'un document en multipliant, pour chaque jeton présent dans un document, sa fréquence au sein du document par l'inverse de la fréquence à laquelle un document du corpus contient le jeton au moins une fois. Nous avons utilisé l'implémentation de Tf-Idf fournie par le module Python Scikit-learn (Pedregosa et coll., 2011) pour ce travail. La formule utilisée par Scikit-learn pour le calcul du IDF est la suivante : $\log \frac{n}{1+df(j)} + 1$, avec *n*, le nombre de documents, et *df(j)*, le nombre de documents dans lequel le jeton *j* est présent. Ainsi, des jetons très fréquents dans un document se verront attribuer un

poids important, étant donné leur grande contribution au sens du document, et un jeton généralement très fréquent dans tout le corpus aura un poids faible, étant donné que cela indique que le jeton n'est que très peu représentatif d'un document. Finalement, le vecteur défini par TF-IDF est normalisé, afin d'éviter que la longueur du document ait un effet.

Par exemple, pour deux documents très courts (en ignorant la ponctuation),

(8) La petite, petite pomme jaune

(9) La grosse, la poire, pas la pomme !

Document	Unité lexicale	TF	IDF	TF-IDF	Normalisé
(8)	La	1	0,8239	0,8239	0,3268
	petite	2	1	2	0,7932
	pomme	1	0,8239	0,8239	0,3268
	jaune	1	1	1	0,3966
(9)	La	3	0,8239	2,4717	0,7900
	grosse	1	1	1	0,3196
	poire	1	1	1	0,3196
	pas	1	1	1	0,3196
	pomme	1	0,8239	0,8239	0,2633

Tableau 5. – Exemple de calcul de TF-IDF

Finalement, une normalisation est effectuée, pour contrebalancer le facteur de longueur du document, étant donné qu'un document plus court ou plus long aurait alors des facteurs plus ou moins importants. Cette méthode permet d'obtenir une pondération représentative des différents jetons qui composent un document, en le distinguant des autres. Dans notre cas, la vectorisation Tf-Idf permet de décomposer les définitions de chaque participant en une sommation des vecteurs de mots qui le composent. Par la suite, un vecteur représentant le document est créé en effectuant une combinaison linéaire des différents vecteurs représentant

un jeton, en multipliant chaque vecteur par le poids obtenu par la méthode TF-IDF et en les additionnant. Par exemple, en supposant que nous ayons les vecteurs suivants :

Unité lexicale	vecteur						
La	0	0	0	0	0	0	1
Petite	0	0	0	0	0	1	0
pomme	0	0	0	0	1	0	0
Jaune	0	0	0	1	0	0	0
grosse	0	0	1	0	0	0	0
Poire	0	1	0	0	0	0	0
Pas	1	0	0	0	0	0	0

Tableau 6. – Exemples de vecteurs de jetons

Une fois la combinaison linéaire effectuée, nous obtenons les vecteurs suivants pour les documents 5 et 6 :

Document	Vecteur						
5	0	0	0	0,3966	0,3268	0,7932	0,3268
6	0,3196	0,3196	0,3196	0	0,2633	0	0,7900

Tableau 7. – Exemples de vecteurs de documents

Dans les deux modèles d'association des éléments de cadre homologue suivants, TF-Idf a été utilisé pour transformer un ensemble de documents en représentations vectorielles pouvant être utilisées pour établir une mesure de distance entre deux définitions d'éléments de cadre.

4.3.1.5.2 Autres modèles testés

Un second modèle a été testé. Un ensemble de vecteurs de mots a été entraîné sur un corpus constitué de l'anthologie des conférences de l'ACL, ainsi que d'articles en accès libre traitant de la sémantique des cadres, le corpus ayant été constitué entre le 15 et le 30 août 2020, et

contenant des textes publiés entre 1965 et 2020[7]. Nous avons choisi ce corpus parce qu'il est représentatif du sujet à l'étude, en particulier les articles traitant de la sémantique des cadres. Étant donné que ce corpus contient des articles scientifiques sur la linguistique informatique, il est attendu que le corpus ait un style et des choix lexicaux relativement uniformes, lorsqu'il s'agit de décrire une même réalité. Les différents documents ont été récupérés au format PDF, puis convertis en texte brut à l'aide du module python PDFplumber (Singer-Vine, 2020). Par la suite, un prétraitement a été effectué sur le corpus, soit la suppression de la casse, et la suppression des caractères non alphabétiques, à l'exception des espaces, des traits d'union et de soulignement ainsi que des chiffres collés à un trait d'union ou de soulignement, afin de ne pas supprimer les noms d'éléments de cadres. Les suites de plus de 16 caractères alphabétiques ont été retirées, pour éviter les problèmes d'espace supprimés lors de la conversion des documents. Ensuite, nous avons entraîné des vecteurs de mots sur ce corpus à l'aide de la bibliothèque Gensim, en ne conservant que les jetons apparaissant plus de 30 fois dans le corpus d'entraînement pour éliminer le bruit. Une fois ces vecteurs construits, les différents documents ont été vectorisés par la méthode Tf-Idf, puis chaque vecteur de document fut comparé aux autres vecteurs de documents de la même façon que pour le premier modèle, qui utilisait des vecteurs représentant les définitions, le tout nous donnant une précision de 39,3 %, et en forçant les éléments ayant le même nom à être liés entre eux, on obtient une précision de 68 %, soit mieux que le modèle trivial.

Finalement, l'utilisation de vecteurs entraînés sur un corpus plus large a été tentée. Pour ce faire, les vecteurs GloVe entraînés à la fois sur Wikipédia en anglais (2014) et Gigaword, un corpus de fils de presse, ont été téléchargés puis, de ces vecteurs, seuls ceux correspondant à des jetons présents dans les définitions des participants ont été conservés, par souci d'espace disque et de mémoire vive. La même procédure que plus haut a été utilisée comme algorithme, ce qui donne une précision de 68,69 % lorsque les noms identiques sont forcés, et de 38,25 % lorsque l'on ne force pas les participants aux noms identiques à être associés.

[7] Une version à jour de l'anthologie peut être retrouvée en format bibtex en suivant le lien ci-après https://www.aclweb.org/anthology/anthology.bib.gz

Soulignons que la précision maximale est de 97,37 %, étant donné que, lors de l'association manuelle, certains éléments n'ont pas pu être associés, vu la difficulté d'association de certains participants, qui ne trouvaient pas d'homologue, tandis que les algorithmes d'association des actants homologues sont élaborés en supposant que tous les participants peuvent être associés à un homologue, et associent donc un maximum de participants entre eux.

Ainsi, le meilleur modèle d'alignement évalué est celui reposant sur GloVe. Bien que la précision évaluée ne soit que de 68,69 % sur le corpus de test, puisque l'alignement n'étant censé être utilisé que pour pallier les informations manquantes, il semble que cette précision soit raisonnable.

Modèle	Gensim + TF-IDF		GloVe	Hasard	Trivial	Max
Corpus	Description des cadres	Corpus de l'ACL				
Précision	55,5 %	68 %	68,69 %	12 %	59 %	97,37 %

Tableau 8. – Précision des algorithmes d'alignement des participants

4.3.1.6 Modificateurs

Au niveau algorithmique, il n'est pas possible de convertir les annotations de modificateurs parfaitement, étant donné leur nature variable, entre un actant d'un sémantème et un simple modificateur.

En effet, dans une expression comme *premier ministre du Québec*, il semble raisonnable de considérer une structure comme ci-dessous, étant donné la nature prédicative de l'unité lexicale PREMIER MINISTRE.

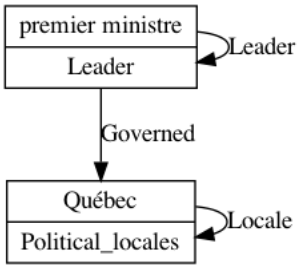


Figure 27. – *Premier ministre du Québec* en sémantique des cadres

Ainsi, 'Québec' doit être un actant de 'premier ministre', bien que syntaxiquement, 'Québec' agisse comme modificateur, à l'instar de 'Québec' dans *pomme du Québec*. Ainsi, il s'agit d'une distinction de nature linguistique, qui n'opère pas de façon strictement formelle.

Le seul traitement qui a été fait afin d'améliorer le portrait fut au niveau de l'analyse syntaxique. Pour chaque annotation, la liste des relations syntaxiques qu'entretient chaque participant à la cible d'annotation. Parmi les différentes relations syntaxiques possibles, la relation *head* désigne la relation entre un élément et celui dont il dépend. Toutefois, cette relation est dans la direction inverse des autres relations syntaxiques. De plus, cette relation syntaxique peut créer des cycles au sein de la structure annotée, ce qui peut empêcher de sélectionner le nœud communicativement dominant à l'aide de la syntaxe. En effet, il est possible, en utilisant les données, d'avoir des structures syntaxiques comme :

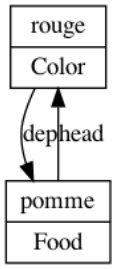


Figure 28. – Syntaxe de *pomme rouge* en sémantique des cadres

Toutefois, une telle structure n'est pas réellement une structure syntaxique, mais plutôt un artefact résultant de l'annotation, puisqu'il ne s'agit alors pas d'un arbre dirigé. En effet, il est possible d'avoir un modificateur comme cible d'annotation d'un cadre pour lequel il existe un

participant qui n'est pas annoté, puisque FrameNet prescrit de ne pas annoter les entités nommées, ces dernières n'étant pas jugées comme étant des éléments pertinents pour une analyse linguistique subséquente. Ainsi, pour un énoncé comme 'le petit Paul', il est attendu d'y retrouver une relation syntaxique entre petit et Paul, même si on ne retrouvera pas, au sein de l'annotation, une relation syntaxique entre Paul et petit, étant donné que Paul est exclu de l'annotation.

Les relations ont donc été renversées, c'est-à-dire qu'une nouvelle relation syntaxique a été créée et des unités lexicales ont été créées ou mises à jour. Ainsi, les participants reliés avec cette relation syntaxique ont été ajoutés ou mis à jour à l'ensemble des annotations, avec seulement une relation syntaxique, *reverse head,* pour permettre aux relations syntaxiques de suivre plus fidèlement la TST, et par le fait même, de sélectionner le nœud communicativement dominant.

4.3.1.7 Lemmatisation

Pour chaque élément annoté comme participant, s'il n'était pas possible de le décomposer (par exemple, lorsqu'il s'agit d'un syntagme pour lequel aucun élément n'est annoté séparément), une lemmatisation a eu lieu. Ainsi, si l'élément de cadre fait également partie des cibles d'annotation, ce qui peut être vérifié par l'utilisation de l'emplacement de l'annotation, le lemme qui a été attribué à la cible d'annotation est utilisé. Toutefois, si la même unité lexicale apparaît plus d'une fois au sein d'une phrase, comme 'pomme' dans 'Paul mange la pomme rouge et Jean mange la pomme verte', alors le lemme associé à l'unité lexicale qui se répète se voit attribuer un numéro afin de distinguer les différentes instances de l'unité lexicale. Pour une unité lexicale qui n'est pas annotée, que ce soit parce qu'il s'agit d'une unité lexicale ne devant pas l'être, ou encore par omission, le texte brut de la phrase est choisi comme lemme, et si la même expression revient à plusieurs reprises en différents endroits, un numéro est accolé.

Il est important de souligner ici qu'il est possible que le même élément soit présent deux fois, mais en l'absence d'indication qu'il s'agisse de la même entité ou du même prédicat, ce qui introduit une erreur dans l'annotation. Fort heureusement, de telles situations ne se produisent presque jamais en discours au sein d'une phrase. En effet, l'analyse de l'utilisation des pronoms et anaphores forme l'assise empirique de la théorie du gouvernement-liage de la grammaire

générative (Chomsky, 1980), et bien qu'il ne s'agit pas de la théorie linguistique retenue pour ce travail, cette théorie prédit qu'une expression référentielle ne peut apparaître que si elle n'est pas c-commandée par un antécédent, ce qui devrait arriver la majorité du temps au sein d'une phrase avec deux instances du même sémantème. Une telle situation peut toutefois se produire dans nos annotations.

4.4 Évaluation

Afin d'évaluer la réussite de l'implémentation, différents tests ont été effectués, afin de donner un portrait aussi clair que possible de l'issue de ce travail. Une évaluation manuelle des structures s'avère irréaliste étant donné la quantité de données. Toutefois, une telle évaluation est possible sur un échantillon, pour peu qu'il soit représentatif.

Une évaluation automatisée sur un corpus plus large a également été prévue. Pour ce faire, il a été envisagé d'utiliser les RSém générées par le script afin de produire une sortie textuelle, et la comparer avec le texte, sans annotations.

4.4.1 Évaluation automatique

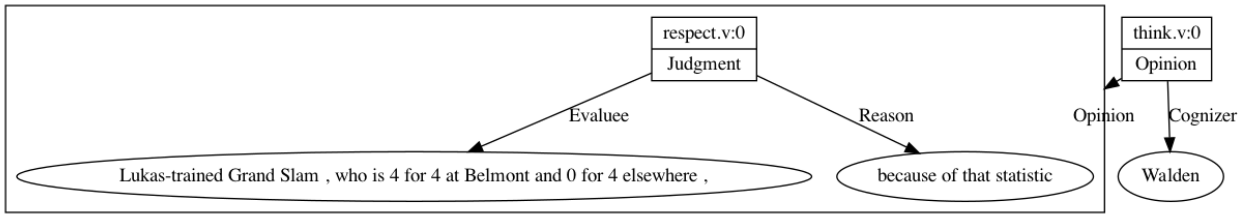
Pour effectuer l'évaluation automatique, il aurait alors fallu utiliser un réalisateur de texte pour produire le texte. Or, GenDR semble être le seul système de génération de texte à partir de RSém, mais son niveau d'avancement n'est pas suffisant pour permettre de générer avec suffisamment de précision des énoncés, sans faire un travail important d'adaptation des différentes unités lexicales, et améliorer certains algorithmes, ce qui est hors de la portée de ce travail. Également, cette approche a été jugée inadéquate, étant donné qu'une telle comparaison entre la sortie du générateur et une phrase cible ne fait pas qu'évaluer la transposition entre annotations de la sémantique des cadres et TST, mais est également très sensible à la capacité de GenDR de produire des énoncés grammaticaux, ainsi qu'à la qualité des annotations initiales. Or, GenDR ne permet pas, dans son état actuel, de produire des énoncés complexes, et les annotations de FrameNet ne sont pas toujours complètes. De plus, une source de biais incontrôlable est la quantité d'information qui n'a pas été annotée au sein des différentes phrases. En effet, il arrive parfois que quelques mots ou syntagmes ne soient pas annotés dans les données originales, ce

qui constitue un biais pour l'évaluation de l'implémentation. Malgré tout, nous pouvons tout de même esquisser un protocole d'évaluation.

Dans le cas hypothétique où GenDR devenait suffisamment mature pour générer sans problème les différentes paraphrases associées aux RSém, il serait alors possible d'attribuer un score basé sur la similarité avec le texte annoté aux différentes sorties textuelles du réalisateur. Étant donné que chaque RSém est associée à un grand nombre de paraphrases, de toutes les sorties possibles d'une même RSém, on ne conserverait que la sortie ayant le meilleur score comme étant une donnée représentative de la qualité de la RSém, le reste pouvant être considéré comme un artifice résultant de GenDR. Aussi, pour éviter d'introduire le biais lié au manque d'annotations, il est possible de retirer les éléments qui n'étaient pas annotés dès le départ afin de s'assurer que la métrique utilisée soit représentative de l'implémentation. Il a été envisagé d'utiliser les métriques BLEU (Papineni et coll., 2002) et ROUGE (Lin, 2004) pour effectuer ce travail. Une présentation de ces métriques se retrouve en annexe.

4.4.2 Évaluation humaine

Finalement, nous avons dû faire une évaluation humaine des différentes structures générées par le système de transformations de structures. Pour ce faire, un échantillon de 242 structures, tirées uniformément au hasard a été analysé par deux étudiants au baccalauréat en linguistique à l'Université de Montréal, familiers avec la TST, ainsi que par l'auteur. Les évaluateurs avaient accès à un protocole pour l'évaluation, disponible en annexe, et pour chaque structure à évaluer, une représentation graphique de la structure d'entrée, ainsi qu'une représentation graphique de la RSém correspondante. À titre d'exemple, les deux figures suivantes représentent une structure d'entrée et sa RSém correspondante (le nœud communicativement dominant de toute la RSém est indiqué en rouge).



Walden thinks Lukas-trained Grand Slam , who is 4 for 4 at Belmont and 0 for 4 elsewhere , must be respected because of that statistic .

Figure 29. – Représentation graphique d'une structure d'entrée pour l'évaluation

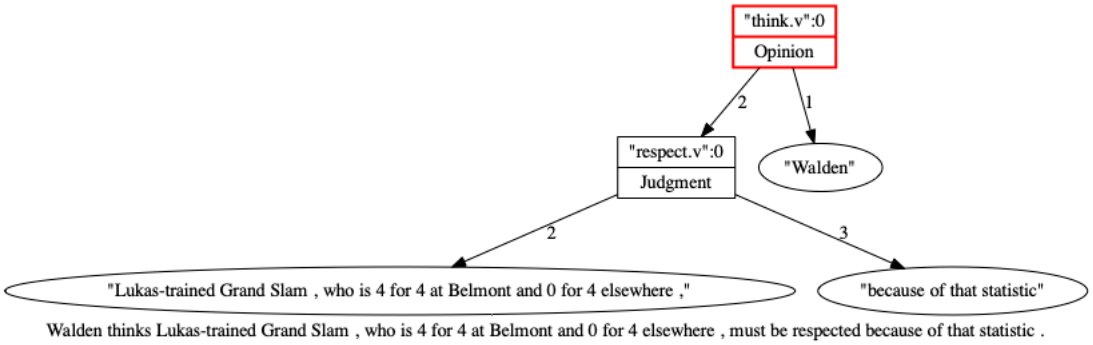


Figure 30. – Représentation graphique de la RSém correspondante à la structure d'entrée

Les différentes structures ont été évaluées sur les aspects suivants : le statut de la RSém correspondante, la conformité au texte, la numérotation des actants, la bonne génération des conjonctions, le cas échéant, et le choix du nœud communicativement dominant.

Pour le statut de RSém, les évaluateurs ont dû déterminer si la structure présentée était une RSém conforme à la TST. Pour ce faire, ils ont dû déterminer si la RSém était bien formée dans l'absolu, et, s'il ne s'agissait pas d'une bonne RSém, si la mauvaise formation de la RSém était attribuable à la structure d'entrée (qui pouvait être incomplète, ou contenir des erreurs).

Pour la conformité au texte, les évaluateurs ont dû déterminer si la RSém fournie pourrait représenter l'énoncé, en tenant compte de la partie annotée du texte. En effet, la plupart des annotations étaient incomplètes, avec quelques unités lexicales manquantes, ce qui peut interférer avec la production de l'énoncé.

Pour la numérotation des actants, les évaluateurs ont dû déterminer si l'ensemble des actants étaient numérotés adéquatement, c'est-à-dire si les numéros qui ont été attribués aux différents actants d'un même sémantème étaient compatibles avec une définition générale du sémantème. Ainsi, les évaluateurs ont dû se prêter à un exercice lexicographique visant à attribuer un numéro à chaque actant, en s'appuyant sur les principes de la TST. Dans le cas où la numérotation n'était pas parfaite, les évaluateurs ont dû déterminer si l'ordre relatif des différents actants était

approprié, en omettant les éléments manquants. Ainsi, si un actant avait une numérotation inadéquate, mais suivant un ordre relatif possible, la structure a été notée comme tel.

Par exemple, la figure 31 est un cas de bonne numérotation absolue, tandis que la figure 32 est une numérotation relative correcte, puisqu'il est attendu d'avoir, en termes de numérotation, le mangeur< l'aliment < l'instrument, alors que la figure 33 est une mauvaise numérotation, étant donné l'inversion de l'aliment et de l'instrument.

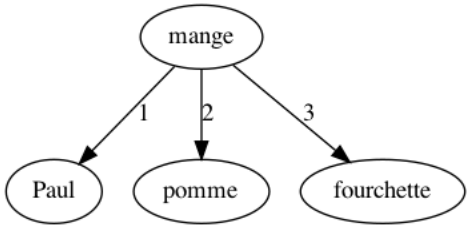


Figure 31. – *Paul mange une pomme avec une fourchette*

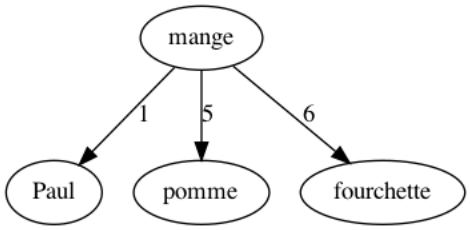


Figure 32. – Numérotation relative de *Paul mange une pomme avec une fourchette*

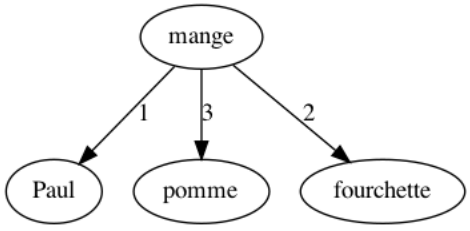


Figure 33. – Mauvaise numérotation de *Paul mange une pomme avec une fourchette*

Pour les conjonctions, les évaluateurs ont eu comme consigne de valider qu'en présence de l'unité lexicale *and* ou *or*, les différents arguments de la conjonction y sont reliés adéquatement.

Pour le choix du nœud communicativement dominant, les évaluateurs ont dû déterminer si ce choix était adéquat, considérant les annotations en sémantique des cadres. Pour ce faire, un peu de latitude a été laissée aux évaluateurs, étant donné qu'en théorie, plusieurs nœuds différents pourraient jouer le rôle de nœud communicativement dominant d'une RSém, parmi lesquels un seul doit être choisi.

Finalement, les évaluateurs ont eu comme consigne de signaler toute anomalie non répertoriée plus haut.

Les différents résultats sont répertoriés ci-après, pour chaque évaluateur, avec le coefficient kappa de Fleiss (Fleiss, 1971), afin de déterminer l'accord interévaluateurs. Nous renverrons le lecteur à (Landis & Koch, 1977) pour une interprétation qualitative du coefficient kappa, oscillant dans notre cas entre *Slight* [faible] (0,1-0,2) et *Fair* [passable] (0,41-0,6).

Statut de RSém					
	Évaluateur 1	Évaluateur 2	Évaluateur 3	Moyenne	Kappa de Fleiss
Bonne en absolu	36 %	47 %	22 %	35 %	0,18
Bonne compte tenu de l'entrée	49 %	45 %	26 %	40 %	
Mauvaise	15 %	7 %	52 %	25 %	

Tableau 9. – Évaluation du statut de RSém

Pour le statut de RSém, il ressort de l'évaluation qu'environ 35 % des structures analysées ont été jugées comme étant bien formées, dans l'absolu, tandis qu'environ 40 % des structures ne sont pas bien formées, en raison de la structure d'entrée.

Conformité au texte					
	Évaluateur 1	Évaluateur 2	Évaluateur 3	Moyenne	Kappa de Fleiss
Conforme	74 %	69 %	64 %	69 %	0,11
Non-conforme	26 %	31 %	36 %	31 %	

Tableau 10. – Évaluation de la conformité au texte

En ce qui a trait à la conformité au texte, 69 % des structures semblent conformes au texte, tandis que 31 % des structures ne sont pas conformes au texte cible.

Numérotation des actants					
	Évaluateur 1	Évaluateur 2	Évaluateur 3	Moyenne	Kappa de Fleiss
Bonne en absolu	48 %	69 %	19 %	42 %	0,14
Bon ordre relatif	27 %	24 %	29 %	26 %	
Mauvais ordre	25 %	17 %	52 %	32 %	

Tableau 11. – Évaluation de la numérotation des actants

En ce qui a trait à la numérotation des actants, 42 % des structures avaient une numérotation parfaite des actants des différentes unités lexicales présentes au sein d'une structure annotée.

Conjonctions					
	Évaluateur 1	Évaluateur 2	Évaluateur 3	Moyenne	Kappa de Fleiss
Non applicable	54 %	76 %	54 %	61 %	0,53
Bien fait	10 %	10 %	14 %	11 %	
Mal fait	36 %	14 %	33 %	28 %	

Tableau 12. – Évaluation des conjonctions

En ce qui a trait aux conjonctions, il semble que la qualité de l'implémentation est plutôt piètre. En effet, sur les 39 % des structures présentant une conjonction ciblée par l'algorithme, moins de

la moitié des structures ont vu leurs conjonctions être implémentées adéquatement dans le passage de FrameNet à la TST.

Choix du nœud communicativement dominant					
	Évaluateur 1	Évaluateur 2	Évaluateur 3	Moyenne	Kappa de Fleiss
Bien fait	74 %	81 %	57 %	71 %	0,46
Mal fait	26 %	19 %	43 %	29 %	

Tableau 13. – Évaluation du choix du nœud communicativement dominant

Quant au choix du nœud communicativement dominant, il s'agit de la tâche pour laquelle l'évaluation a démontré le meilleur résultat, avec 71 % de bonnes sélections du nœud communicativement dominant.

Finalement, parmi les erreurs jugées pertinentes à mentionner par les évaluateurs, un élément rapporté à quelques reprises concerne le traitement des modificateurs. En effet, la direction de la relation sémantique entre un élément et son modificateur était fréquemment inappropriée. En revanche, étant donné que cet aspect n'était pas présent dans le protocole d'évaluation, il n'est pas possible de quantifier l'effet des incongruités au niveau du traitement des modificateurs.

En somme, l'évaluation humaine permet de constater que, pour la plupart des critères, environ 70 % des structures produites sont adéquates, les autres n'étant pas conformes.

Malgré la relative faiblesse de l'accord interévaluateurs, il ne semble pas y avoir d'évaluateur qui ait fait une évaluation radicalement différente de celle des deux autres. Il semble que l'évaluateur 1 est celui qui a fait l'évaluation la plus consensuelle, étant donné que les coefficients kappa entre l'évaluateur 1 et les deux autres, pour chaque critère sont systématiquement supérieur à l'accord entre les évaluateurs 2 et 3, tel qu'on peut apercevoir dans le tableau 14.

Évaluateurs	Statut de RSém	Conformité au texte	Numérotation des actants	Conjonctions	Nœud dominant
1 et 2	28,19 %	4,12 %	19,38 %	46,58 %	53,21 %
1 et 3	12,15 %	25,19 %	14,91 %	69,19 %	53,10 %
2 et 3	5,54 %	2,87 %	4,52 %	40,07 %	29,92 %

Tableau 14. – Coefficients Kappa de Fleiss par paire d'annotateurs

Ce désaccord relatif entre les évaluateurs démontre que la tâche d'évaluation comporte une part de subjectivité. Toutefois, bien qu'il y ait une forte variabilité entre les différents annotateurs, les résultats globaux restent comparables d'un évaluateur à l'autre. Ainsi, bien qu'il y ait un effet évident du choix d'annotateur sur les résultats d'évaluation, cet effet est attendu, comme pour toute mesure où une certaine subjectivité existe.

5 Discussion

5.1 Résultats de l'évaluation

Pour l'analyse de la conformité au texte, où 70 % des structures ont été jugées conformes, un élément rapporté à quelques reprises par les évaluateurs pouvant nuire à la conformité du texte est celui du traitement des modificateurs, tel que rapporté plus haut. En effet, la direction de la relation entre un modificateur et l'élément modifié fait en sorte que la structure est non conforme au texte.

Pour la numérotation des actants, il est bon de rappeler que les structures comprenaient plusieurs unités lexicales annotées, et que dès qu'une erreur était rapportée, l'ensemble de la structure se voyait mise dans la catégorie de mauvaise numérotation, ou de bon ordre relatif des actants, si les actants étaient dans un bon ordre relatif (ce qui comptait pour 26 % des énoncés restants). Avec ce fait en tête, la numérotation des actants semble particulièrement efficace, surtout en considérant que la précision du meilleur algorithme pour la sélection des actants homologues était de 68,69 % (voir section 4.3.1.5). Toutefois, il semble opportun de rappeler qu'en sémantique des cadres, les différents cadres incluent habituellement un assez grand nombre de participants, 15 en moyenne, dont une quantité appréciable de participants non obligatoires, ce qui vient limiter la précision de la méthode de numérotation employée. En pratique, il est très rare que l'ensemble des participants soient instanciés, puisque pour tout acte de parole, une relation est établie entre le locuteur et les interlocuteurs potentiels, et selon Grice (1982), nommer des savoirs mutuels serait une violation de la maxime de quantité, étant donné qu'un énoncé serait alors plus informatif que nécessaire, ce qui peut amener les interlocuteurs à être confus.

Ainsi, la précision relativement faible de l'algorithme pour la sélection des actants homologues est à relativiser, considérant que ce n'est pas toute la numérotation qui est sélectionnée par cet algorithme.

Face à l'échec de l'implémentation des conjonctions, avec seulement 11 % des structures avec une implémentation réussie des conjonctions, par rapport à 28 % de conjonctions mal implémentées, deux options s'offrent à nous : soit les structures comprenant une conjonction devraient être éliminées d'emblée, ou encore, l'algorithme de génération des conjonctions pourrait être utilisé, faute de mieux, en sachant que le résultat obtenu est médiocre. Ce choix sera laissé aux personnes désirant utiliser les structures de la TST générées, selon leurs besoins. Toutefois, il est impératif de soit retirer les énoncés comportant une conjonction, soit tenter de corriger les conjonctions, étant donné que d'avoir résolu les conjonctions est crucial pour certaines parties de l'algorithme de conversion, notamment le choix du nœud communicativement dominant et la résolution des constructions relatives.

C'est sans surprise que la sélection du nœud communicativement dominant est le critère pour lequel le résultat est le meilleur, étant donné qu'il y a, théoriquement, plusieurs choix de nœud communicativement dominant qui soient appropriés pour une phrase donnée.

5.2 Modifications nécessaires aux formalismes pour faciliter la conversion

Il est possible de passer d'une représentation du sens depuis la sémantique des cadres vers la TST sans trop de heurts, ce qui démontre que les cadres théoriques sont *a priori* compatibles au niveau de la sémantique. Toutefois, bien que les cadres théoriques soient compatibles entre eux, quelques divergences existent, empêchant la conversion d'être parfaitement triviale.

Ces différences sont de deux types : il y a des différences attribuables à la méthode de développement appliqué des formalismes, et des différences attribuables à la granularité différente des formalismes, la TST ayant une approche plus fine des différentes unités lexicales, alors que FrameNet tente d'extrapoler des faits généraux sur les unités lexicales en les regroupant par cadres. Les différences au niveau du développement appliqué constituent la plupart des sources d'erreur pour la conversion d'annotations depuis le corpus de FrameNet vers des RSém, tandis que les différences attribuables à la granularité des formalismes nous permettent d'anticiper les erreurs probables dans une conversion future de RSém vers des structures

annotées en sémantique des cadres. En effet, pour pouvoir affirmer hors de tout doute que les deux formalismes sont compatibles en ce qui a trait à l'aspect sémantique, il faut pouvoir effectuer le travail inverse[8]. En terminant, il semble que les problèmes liés à la numérotation des participants relèvent des deux types de différences mentionnées plus haut.

5.2.1 Différences d'application

Les différentes difficultés et sources d'erreur lors du passage de FrameNet à la TST proviennent principalement du manque d'information au sein des structures. Ce manque d'information résulte notamment de la structuration des données, ainsi que de certains choix qui ont été faits dans le protocole d'annotation. En effet, le projet FrameNet vise l'élaboration des cadres sémantiques et le classement des unités lexicales, et non la possibilité de générer du texte. En ce sens, les annotations sont ciblées pour la tâche en question, omettant certains détails qui nous sont pourtant cruciaux.

Au sein des structures annotées de FrameNet, lorsqu'il y a l'utilisation d'un pronom, mis à part les pronoms relatifs, il n'est pas fait mention de l'antécédent, et aucune coréférence n'est indiquée. Ainsi, un énoncé comme *Il a dansé et il a joué* n'est pas annoté de sorte que l'on puisse savoir s'il s'agit du même individu qui a dansé et qui a joué, alors que la structure sémantique est très différente qu'il s'agisse du même référent ou non. Dans la même veine, le format des données ne permet pas de savoir si un participant est un pronom ou l'antécédent d'un pronom, sans que la relation ne soit parfaitement explicite en tout temps. Par exemple, dans un énoncé comme en (10),

(10) « Le temps est bon », qui a été chanté par Isabelle Pierre dont c'est le plus grand succès musical […]

Tant *dont* que *c'* seront annotés comme des pronoms, avec comme antécédents *Isabelle Pierre* ainsi que *« Le temps est bon », qui a été chanté par Isabelle Pierre*, sans mentionner explicitement quel antécédent va avec quel pronom.

[8] Il serait alors possible de convertir sans heurts d'un formalisme à l'autre, en préservant les structures, ce qui mathématiquement, s'apparente à un isomorphisme.

Aussi, dans les données de FrameNet, les participants sont indiqués par leur position au sein de la chaîne de caractères, et non par un lemme. Si l'absence de lemmatisation peut être contournée, comme ça a été fait dans ce travail, des participants qui ne sont pas que des mots-formes peuvent poser problème. En effet, lorsqu'il s'agit d'un syntagme, il est nécessaire, pour la TST, de sélectionner un nœud comme étant communicativement dominant. Toutefois, cette sélection suppose, pour être bien faite, que l'ensemble du syntagme soit correctement annoté, ce qui n'est pas toujours le cas, une annotation incomplète créant des distorsions de sens.

Également, le guide d'annotation de FrameNet prescrit que seuls les participants reliés syntaxiquement à l'unité lexicale évoquant un cadre soient annotés (Ruppenhofer et coll., 2016). Toutefois, certains participants peuvent ne pas être reliés syntaxiquement, en vertu, notamment, de constructions syntaxiques. Ainsi, des participants peuvent être omis de l'annotation, comme dans l'exemple suivant :

(11) À Paris, Paul a aimé visiter la mairie.

En effet, 'mairie' est un quasi-prédicat, puisqu'une mairie n'en est pas une sans qu'il y ait une municipalité qui lui soit rattachée. Toutefois, le lien syntaxique entre MAIRIE et PARIS ne permet pas d'annoter 'Paris' en tant qu'élément de cadre pour 'mairie', puisque 'Paris' n'est pas lié syntaxiquement à la projection maximale de 'mairie', qui est la condition nécessaire pour qu'il y ait une annotation, suivant les lignes directrices de FrameNet (Ruppenhofer et coll., 2016). Ainsi, la représentation de cette phrase en sémantique des cadres considère que 'Paris' est un cas d'instanciation nulle par construction [CNI], ce qui empêche d'en extraire l'information.

Aussi, les annotations de FrameNet incluent les verbes supports, sans distinction des différents types (que ce soit un verbe sémantiquement vide, ou qu'il véhicule un certain registre, point de vue, aspect, ou encore une causalité) (Ruppenhofer et coll., 2016). Or ces différences ne sont pas anodines d'un point de vue sémantique, et adopter une distinction au sein même des données permettrait de ne pas échapper ces éléments cruciaux de la sémantique.

Finalement, les annotations de FrameNet n'incluent pas de sens grammaticaux, mais uniquement des sens lexicaux, ce qui ne permet pas d'avoir des RSém véritablement complètes.

Ainsi, pour faciliter la conversion depuis FrameNet vers des RSém, il faudrait bonifier les annotations de FrameNet en ajoutant davantage d'informations qui ne sont pas ou peu utiles, *a priori*, au développement des cadres.

5.2.2 Différences de granularité

Si ces éléments permettaient de rendre trivial le passage d'annotations de FrameNet à des RSém, deux éléments, reliés aux granularités différentes des deux formalismes, manquent pour effectuer le passage inverse, d'une RSém à des annotations de la sémantique des cadres : la dénomination des actants, et les cadres en eux-mêmes. Dénomination des actants et fonctions lexicales dérivés sémantiques nominaux

Par dénomination des actants, il est entendu l'utilisation d'une nomenclature plus générale aux différents actants d'une unité lexicale. Bien que cette nomenclature puisse rendre les définitions plus complexes (section 3.1.5), il s'avère que sans ce mécanisme, le passage depuis des numéros vers des noms de participants pose problème. En effet, il pourrait être tentant de faire exactement l'inverse de ce qui a été fait pour le passage depuis des noms de participants à une numérotation des actants, c'est-à-dire de renverser notre relation entre un nom d'élément de cadre et un numéro d'actant, pour choisir le nom de relation en fonction du numéro d'actant ; toutefois, il est bon de rappeler qu'en contexte d'évaluation, seuls 42 % des structures avaient une numérotation réellement adéquate, tandis que 58 % des structures avaient une numérotation des actants erronée. L'apparente résistance à la dénomination des actants en TST mérite toutefois d'être relativisée. En effet, si les noms d'actants ne sont pas mentionnés dans les définitions lexicographiques des unités lexicales, c'est également parce que le but premier de la TST n'est pas de classer les unités lexicales, mais vise d'abord et avant tout à l'élaboration de modèles Sens-Texte, et peut donc — et doit donc — en faire l'économie. Mentionnons au passage que (Alonso-Ramos, 2003) propose que les actants sémantiques sont tous associés à des éléments de cadre, mais pas l'inverse.

Selon Mel'čuk et collègues (1995), il existe une famille de fonctions lexicales S_1, S_2, S_3, … associant à un sémantème le nom typique donné à ses actants. Ces fonctions lexicales sont appelées les

dérivés sémantiques nominaux actanciels. Par exemple, les fonctions S_1, S_2, S_3, appliquées à l'unité lexicale PARLER ont comme valeur respective *locuteur*, *propos*, et *destinataire*.

Il est ainsi possible de tracer un parallèle entre ces fonctions lexicales et l'opposition entre numérotation et dénomination des actants. En effet, tel que mentionné plus haut (3.1.5, p.36), la pratique de la TST, contrairement à celle de la sémantique des cadres, est de ne pas associer un nom aux actants, mais plutôt, de les numéroter. Toutefois, les valeurs des fonctions lexicales de dérivés sémantiques nominaux actanciels sont encodées au sein du lexique des locuteurs, ce qui selon nous est équivalent à attribuer un nom aux différents numéros. Bien qu'il y ait un rapprochement, celui-ci est à nuancer légèrement, étant donné que ce n'est pas pour tous les sémantèmes que les dérivés sémantiques nominaux actanciels ont une valeur non nulle. En effet, pour un sémantème comme 'tenir', réalisé linguistiquement par l'unité lexicale TENIR$_I$[9], dénotant qu'une personne serre quelque chose pour éviter qu'il ne tombe ou ne s'échappe, il n'existe pas, à notre connaissance, de dérivé sémantique pour ce qui est tenu. Ainsi, il n'y a pas de valeur pour S_2(tenir).

Malgré cela, rien ne nous empêche de considérer un nom plus général exprimant l'actant sémantique 2, par exemple CHOSE, qui n'est toutefois pas à TENIR ce que PROPOS est à PARLER, puisque dans un cas la relation est plutôt générale, tandis que dans l'autre cas, la relation est véritablement régie par la langue.

5.2.2.1 Existence des cadres

En ce qui a trait aux cadres en eux-mêmes, il est entendu que pour associer des unités lexicales à un cadre, il faut que ceux-ci existent ; or, la TST ne prévoit pas l'existence de cadres. Toutefois, il pourrait être tentant de suppléer la notion de cadre avec le genre prochain, tels que présents au sein des définitions lexicographiques, lesquels, conceptuellement, pourraient s'approcher de la notion de cadre. En effet, la principale différence entre la sémantique des cadres et la TST vient de la granularité de l'analyse, la sémantique des cadres ayant une approche macroscopique, tandis que la TST, une approche microscopique de l'analyse des unités lexicales. En lexicologie, la définition d'une unité lexicale est classiquement divisée en deux parties : le genre prochain et les

[9] Numérotation du petit Robert

différences spécifiques, c'est-à-dire un classement sommaire, et ce qui distingue l'unité lexicale en question des autres membres du classement, respectivement. Le genre prochain ayant lui-même un autre genre prochain, on peut facilement imaginer une arborescence entre les différents sens, de façon analogue aux cadres sémantiques.

5.2.3 Numérotation des participants

En ce qui a trait aux participants, les cadres en comptent en moyenne 15. Or, une bonne partie de ceux-ci sont des circonstants et modificateurs, et pour les raisons évoquées plus haut (section 3.2), ces participants répondent à des impératifs reliés à l'élaboration des cadres, mais au fil des révisions, il reste possible qu'il y ait des raffinements. De plus, les participants d'un cadre ne sont pas ordonnés entre eux, ce à quoi nous avons tenté de remédier en leur attribuant un ordre. Bien que les participants ne soient pas ordonnés, un certain classement est effectué entre obligatoire, périphérique et extrathématique, ce qui laisse entrevoir une certaine relation d'ordre, mais très imparfaite.

Toutefois, la différence ne vient pas seulement de l'application, mais aussi de la granularité du formalisme, étant donné que pour regrouper un ensemble d'unités lexicales au sein d'une même catégorie, il n'est pas envisageable de pouvoir tenir compte de toutes les spécificités des unités lexicales, la distinction entre actant et circonstant, en TST n'étant d'ailleurs pas triviale. Ainsi, cette différence est due à la fois à l'application du modèle de développement et à la granularité relative des deux formalismes.

5.3 Parallèles entre les deux formalismes

Comme mentionné précédemment, les deux formalismes s'accommodent relativement bien l'un de l'autre, et certains rapprochements peuvent être faits entre ces derniers, permettant d'enrichir une discussion future entre les deux formalismes.

5.3.1 ConceptR

La notion de ConceptR, qui n'est pas explicitement définie dans le cadre de la TST, est un élément crucial au sein de la théorie, étant donné que sans modélisation du monde à communiquer, il n'y a pas de possibilités d'énonciation, la notion même de sens reposant sur une conception du

monde schématique. En soi, la notion de ConceptR n'est pas de nature linguistique, toutefois, de la même façon que l'on ne peut faire de chimie sans toucher à la physique, ou encore de biologie sans chimie, les ConceptR, même en appartenant à une autre discipline, méritent d'être développés. Faute de conceptualisation appropriée pour le modèle Sens-Texte, le formalisme de la sémantique des cadres peut être utile pour pallier ce manque, en tant que modélisation schématique du monde. Puisque la sémantique des cadres est, à la vue de nos travaux, compatible avec la TST, il est possible de supposer qu'elle pourrait s'inscrire au sein du modèle Sens-Texte, au niveau des ConceptR.

En pratique, la sémantique des cadres est déjà utilisée pour ajouter une conceptualisation schématique du monde à des travaux lexicographiques, par exemple, avec les bases de données terminologiques telles que le *Framed DiCoEnviro*, qui met à profit à la fois les concepts lexicographiques de la TST, et une méthodologie héritée de la sémantique des cadres pour fournir un arrière-plan conceptuel (L'Homme et coll., 2020).

5.3.2 Relations entre cadres ou entre unités lexicales

Il est également possible de tracer un parallèle entre certaines relations entre cadres (section 2.2) et diverses fonctions lexicales de la TST (cf. Mel'čuk et coll., 1995 ; Polguère, 2008 et section 2.3.1.2.2). En effet, en passant en revue les différentes relations entre cadres, l'ombre de diverses fonctions lexicales émerge. Ces parallèles sont tracés par rapport aux relations entre deux cadres, et non entre les unités lexicales associées aux cadres en relation.

La relation d'héritage entre deux cadres permet de définir un cadre en fonction d'un cadre plus général, ce qui nous permet de l'apparenter à la fonction lexicale Gener, qui permet d'associer à une unité lexicale, une unité lexicale plus générique (Jousse, 2010).

La relation de perspective peut être mise en relation avec la fonction lexicale $Conv_{ij}$, qui permet l'inversion d'actants entre deux entités. Par exemple, le $Conv_{3214}$ d'ACHETER est VENDRE, (Mel'čuk et coll., 1995), puisque les actants représentant l'acheteur et le vendeur sont intervertis entre les deux cadres. Ce rapprochement a d'ailleurs déjà été discuté dans (Coyne & Rambow, 2009), où une procédure est décrite pour utiliser les données de FrameNet pour obtenir une liste de paires de verbes reliés par la fonction lexicale $Conv_{ij}$.

La relation de causativité peut être reliée à la fonction lexicale Caus, qui retourne, pour une unité lexicale donnée, l'unité lexicale représentant le fait de causer l'unité lexicale correspondante, comme dans les phrases (1) et (2), (p.23) où l'unité lexicale COULER a deux acceptions, l'une étant intransitive et l'autre, transitive.

La relation d'inchoativité quant à elle peut être associée à la fonction lexicale Incep, désignant le début d'une action ou encore à Germ, qui en désigne la source.

Toutefois, d'autres relations entre les cadres semblent résister à ces parallèles. En effet, la relation de précédence, désignant un événement venant logiquement avant un autre, pourrait *a priori*, être mise en relation avec Incep, ou encore avec Caus, deux fonctions lexicales impliquant une suite temporelle. Cependant, Incep désigne le début d'une action, alors que lorsqu'un cadre en précède un autre, l'action n'est pas entamée, et Caus désigne un lien de causalité, alors que la relation de précédence n'en est pas une de causalité, mais détermine plutôt un contexte temporel nécessaire au cadre suivant.

Également, les relations de sous-cadre et d'utilisation, n'ont non plus pu être mises en parallèle avec une fonction lexicale standard. En effet, la relation de sous-cadre, définie entre un cadre dénotant un événement complexe et ses parties, bien qu'elle puisse être schématiquement décomposée en une relation entre un tout et ses parties, ne semble pas correspondre à une fonction lexicale précise. Tout au plus, un rapprochement peut être fait entre les sous-cadres et les fonctions lexicales phasiques (Incep, Cont, Fin), qui désignent essentiellement le début, le milieu et la fin. Or, le nombre de sous-cadres d'un cadre complexe n'est pas limité à trois, ce qui empêche d'associer les fonctions lexicales phasiques à la notion de sous-cadre.

La relation d'utilisation ne semble pas pouvoir être mise en parallèle avec une fonction lexicale, étant donné que l'utilisation d'un cadre par un autre ne correspond pas à une fonction lexicale standard. Toutefois, on peut comparer la relation d'utilisation à la décomposition d'une définition lexicographique en ses différents sémantèmes. En effet, dans le cas de la relation d'utilisation, un cadre est utilisé pour en définir un autre, comme dans une décomposition de sémantèmes en sémantèmes sémantiquement plus simples.

Les différents parallèles présentés s'inscrivent dans une logique comparative entre les deux formalismes présentés. Bien qu'une adéquation entre fonction lexicale et relation entre cadres ne soit pas prévue *a priori*, certains parallèles se dressent facilement entre des relations entre des cadres et des fonctions lexicales. Étant donné que les deux formalismes visent à modéliser une même réalité, ceux-ci devraient utiliser les mêmes concepts et relations. Il est donc pertinent de s'interroger sur les relations entre les cadres pour lesquels le parallèle ne vient pas aussi naturellement, ainsi que les fonctions lexicales n'ayant pas de correspondance claire en sémantique des cadres en veillant à nourrir les deux formalismes. Pour ce faire, il faudrait analyser les relations entre cadres et les fonctions lexicales. Ainsi, il est possible de considérer que les fonctions lexicales phasiques (Incep, Cont, Fin) puissent être plus nombreuses que trois, ou encore que les relations de sous-cadre devraient être raffinées, en supposant l'existence de sous-cadres intermédiaires, qui seraient regroupés par triplets.

Ces parallèles ne sont qu'à un stade d'ébauche ; si les parallèles étaient parfaitement limpides, nous aurions alors la possibilité d'utiliser ces parallèles pour effectuer de l'extraction d'information plus intéressante. En effet, en tenant pour compte le fait que de tels parallèles sont possibles, il serait, dans un second temps, possible de mettre à profit les relations entre cadres pour produire, au sein des RSém, l'expression de fonctions lexicales depuis les annotations de FrameNet, rendant alors les possibilités de génération et de paraphrasage plus riches.

6 Conclusion

En conclusion, bien que la TST et la sémantique des cadres reposent sur des assises théoriques différentes, les deux modèles ne sont pas pour autant incompatibles, bien au contraire. Tel qu'il a été discuté plus haut, les deux formalismes adoptent un point de vue relativement cohérent en ce qui a trait aux réseaux sémantiques, en adoptant des structures en réseau plutôt qu'arborescentes. De plus, le choix de définir les actants par numérotation ou par nomenclature n'est pas *a priori* une différence fondamentale, pour peu qu'il soit possible d'établir une correspondance entre des numéros et des noms, ce qui a été fait dans ce travail. Ainsi, si, pour Mel'čuk, le fait de nommer rend les définitions plus complexes, le fait de numéroter n'est, à notre sens, qu'une façon de rendre implicite cette notion, d'autant plus qu'il existe, au sens de la TST, des fonctions lexicales S_1, … S_n, associant à un actant sémantique son nom typique (Mel'čuk et coll., 1995).

En ce qui a trait aux fonctions lexicales, un des éléments clef de la TST, et, à notre avis, une façon très élégante de structurer le lexique, la sémantique des cadres adopte explicitement la notion de verbe support, et divers sens, qui s'approchent des ceux des fonctions lexicales sont inclus tacitement au sein de la structure même des cadres, via les relations entre les cadres, démontrant ainsi la compatibilité des formalismes, qui sous-tendent des concepts similaires.

La TST, comme toute entreprise scientifique, n'est somme toute qu'une tentative de modélisation scientifique (Wierzbicka, 1996), et en soi, n'est pas complète, en n'ayant pas de modélisation claire de la notion de ConceptR, qui n'est qu'esquissée, tandis que la sémantique des cadres souffre de lacunes en ce qui a trait aux aspects syntaxiques et morphologiques, empruntant les notions syntaxiques à la simple dépendance, et omettant la morphologie. L'intersection de ces deux formalismes est la sémantique, et il semble, à la vue de ce travail, qu'il soit possible d'effectuer une conversion depuis la sémantique des cadres vers la Théorie Sens-Texte. Nous terminerons donc en postulant qu'il est probable, sous réserve d'avoir une modélisation appropriée des ConceptR, qu'il soit possible d'effectuer une conversion inverse : depuis la TST vers la sémantique des cadres. Toutefois, les ConceptR n'étant pas formalisées, il est nécessaire

d’avoir une telle modélisation, qui pourrait s’inspirer — très fortement — de la sémantique des cadres.

7 Références bibliographiques

Alonso-Ramos, M. (2003). Éléments du frame vs. Actants de l'unité lexicale. *Proceedings of the First Meaning Text Theory*, 77-88.

Aristote. (2015). *Topiques ; Réfutations sophistiques : Organon, V-VI* (P. Pellegrin, Éd.; J. Brunschwig & M. Hecquet-Devienne, Trad.). Flammarion.

Baker, C. F. (2015). *FrameNet 1.6*. https://framenet.icsi.berkeley.edu/fndrupal/

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics -*, *1*, 86. https://doi.org/10.3115/980845.980860

Boas, H. C. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography*, *18*(4), 445-478. https://doi.org/10.1093/ijl/eci043

Chomsky, N. (1980). On Binding. *Linguistic Inquiry*, *11*(1), 1-46. JSTOR.

Chomsky, N. (1993). *Lectures on government and binding : The Pisa lectures* (7th ed). Mouton de Gruyter.

Coyne, R. E., & Rambow, O. C. (2009). *Meaning-Text-Theory and Lexical Frames*. https://doi.org/10.7916/D8K93H0R

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. http://arxiv.org/abs/1810.04805

Fillmore, C. J. (1976). Frame Semantics and the nature of language. *Annals of the New York Academy of Sciences*, *280*(1 Origins and E), 20-32. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

Fillmore, C. J. (1982). Frame semantics. Dans *Linguistics in the Morning Calm* (p. 111-137). Hanshin Publishing Co.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Frege, G. (1892). On sense and reference. Dans A. W. Moore (Éd.), & M. Black (Trad.), *Meaning and reference* (p. 23-42). Oxford University Press.

Galarreta-Piquette, D. (2018). *Intégration de VerbNet dans un réalisateur profond* [Mémoire de maîtrise, Université de Montréal]. http://hdl.handle.net/1866/21112

Grice, H. P. (1982). Logic and conversation. Dans P. Cole (Éd.), *Speech acts* (5. ed, p. 41-58). Academic Press.

Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? Dans S. Staab & R. Studer (Éds.), *Handbook on Ontologies* (p. 1-17). Springer. https://doi.org/10.1007/978-3-540-92673-3_0

Jousse, A.-L. (2010). *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales* [Thèse de doctorat, Université de Montréal]. Papyrus. http://hdl.handle.net/1866/4347

Kingsbury, P., & Palmer, M. (2002). From Treebank to PropBank. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. LREC-2002, Las Palmas.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174.

Lareau, F., Lambrey, F., Dubinskaite, I., Galarreta-Piquette, D., & Nejat, M. (2018). GenDR: A Generic Deep Realizer with Complex Lexicalization. *Proceedings of 11th edition of the Language Resources and Evaluation Conference (LREC)*. http://flareau.ca/files/lareau+etal-lrec18.pdf

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley.

L'Homme, M.-C., Lanevile, M.-E., & Azoulay, D. (2014). *DiCoEnviro Le dictionnaire fondamental de l'environnement*. http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf

L'Homme, M.-C., Robichaud, B., & Carlos, S. (2020). *Building Multilingual Specialized Resources Based on FrameNet : Application to the Field of the Environment*.

Lin, C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. *Text summarization branches out*, 74-81. https://www.aclweb.org/anthology/W04-1013.pdf

Macleod, C., Ide, N., & Grishman, R. (2000, mai). The American National Corpus : A Standardized Resource for American English. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. http://www.lrec-conf.org/proceedings/lrec2000/pdf/196.pdf

Markov, A. A. (1913, janvier). *An Example of Statistical Investigation of the Text* Eugene Onegin *Concerning the Connection of Samples in Chains*. https://www.cambridge.org/core/product/identifier/S0269889706001074/type/journal_article

Mel'čuk, I. A. (1988). *Dependency syntax : Theory and practice*. State University Press of New York.

Mel'čuk, I. A. (2001). *Communicative organization in natural language : The semantic-communicative structure of sentences*. J. Benjamins.

Mel'čuk, I. A., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot.

Mel'čuk, I. A.(2004). Actants in semantics and syntax I : Actants in semantics. *Linguistics*, *42*(1), 1-66. https://doi.org/10.1515/ling.2004.004

Mel'čuk, I. A. (2020). Clichés and pragmatemes. *NEO*, *32*, 9-20. https://doi.org/10.31261/NEO.2020.32.01

Mel'čuk, I. A. (1997). Vers une linguistique Sens-Texte. *Leçon Inaugurale*, 78.

Mel'čuk, I. A., & Milićević, J. (2020). *An Advanced Introduction to Semantics : A Meaning-Text Approach* (1re éd.). Cambridge University Press. https://doi.org/10.1017/9781108674553

Milićević, J. (2006). A SHORT GUIDE TO THE MEANING-TEXT LINGUISTIC THEORY. *Journal of Koralex*, *8*, 187-233.

National Institute of Standards and Technology. (2010). *Advanced Question Answering for Intelligence*. https://www-nlpir.nist.gov/projects/aquaint/index.html

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443-453. https://doi.org/10.1016/0022-2836(70)90057-4

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics (ACL), 311-318. https://www.aclweb.org/anthology/P02-1040.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Polguère, A. (1990). *Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte* [Thèse de doctorat, Université de Montréal]. http://olst.ling.umontreal.ca/pdf/PolguerePhD1990.pdf

Polguère, A. (Éd.). (2008). *Lexicologie et sémantique lexicale : Notions fondamentales* (Nouv. éd, revue et augm.2. éd). Presses de l'Univ. de Montréal.

Polguère, A. (2011). Perspective épistémologique sur l'approche linguistique Sens-Texte. Dans *Mémoires de la Société de Linguistique de Paris* (p. 79-114). https://hal.archives-ouvertes.fr/hal-00686461/document

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50.

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems* (1re éd.). Cambridge University Press. https://doi.org/10.1017/CBO9780511519857

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192-233. https://doi.org/10.1037/0096-3445.104.3.192

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Schmidt, T. (2009). 4. The Kicktionary – a multilingual lexical resource of football language. Dans H. C. Boas (Éd.), *Trends in Linguistics. Studies and Monographs [TiLSM]*. Mouton de Gruyter. https://doi.org/10.1515/9783110212976.1.101

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT : Learning Robust Metrics for Text Generation. *arXiv:2004.04696 [cs]*. http://arxiv.org/abs/2004.04696

Shieber, S. M. (2006). Does the Turing Test Demonstrate Intelligence or Not. *proceedings of the twenty-first AAAI conference on artificial intelligence*, 1539-1542. https://www.aaai.org/Papers/AAAI/2006/AAAI06-245.pdf

Singer-Vine, J. (2020). *Pdfplumber* (0.5.23) [Computer software]. https://github.com/jsvine/pdfplumber

Timponi Torrent, T., Ellsworth, M., Baker, C., & da Silva Matos, E. E. (2018). The Multilingual FrameNet Shared Annotation Task : A Preliminary Report. Dans T. Timponi Torrent, L. Borin, & C. F. Baker (Éds.), *Proceedings of the LREC 2018 Workshop International FrameNet Workshop 2018 : Multilingual Framenets and Constructicons*. http://lrec-conf.org/workshops/lrec2018/W5/pdf/book_of_proceedings.pdf#page=70

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. https://doi.org/10.1145/365153.365168

Wierzbicka, A. (1996). *Semantics : Primes and universals*. Oxford University Press.

8 Annexes

Annexe 1 : preuve de la congruence entre une approche par poids pour la sélection du nœud dominant et la sélection du nœud racine.

Par induction sur le nombre de nœuds

Soit un arbre dirigé, défini comme étant un graphe connecté pour lequel chaque nœud N peut avoir au plus un nœud M pour lequel il existe une flèche de M vers N.

Base d'induction :

Soit un arbre dirigé à un nœud, celui-ci comporte un nœud qui est à la fois le nœud dominant, et le nœud choisi par la méthode par poids.

Induction :

Supposons que pour tout arbre dirigé avec N-1 nœuds, les méthodes par poids et par nœud dominant sont équivalentes

Soit un arbre dirigé T à N nœuds, avec n_0 comme nœud dominant.

On retire le nœud dominant pour obtenir une forêt dirigée F' avec N-1 nœuds.

Alors, chaque nœud qui était relié à n_0 est le nœud dominant d'un arbre dirigé avec un nombre de nœuds k <= N-1. Par l'hypothèse d'induction, le nœud dominant de l'arbre dirigé est également le nœud ayant le plus grand poids. Ainsi, en replaçant n_0 dans la forêt dirigée pour retrouver T, on obtient un poids pour le nœud n_0 supérieur à la somme des poids des nœuds dominants de la forêt, et en particulier un poids supérieur à tous les autres nœuds. Ainsi, si pour tout arbre dirigé à n nœuds, la méthode des poids est équivalente à la sélection du nœud racine, alors pour tout arbre dirigé à n+1 nœuds, la méthode des poids est équivalente à la sélection du nœud racine.

Conclusion : Pour tout arbre dirigé la méthode des poids sélectionne le nœud racine comme nœud dominant.

Annexe 2 : Densité de probabilité :

Lemme : Soit A et B, deux ensembles finis et non nuls, |A|=n et |B| leurs cardinalités respectives, avec |B| > n et F, le nombre d'injections de A dans B, alors $|F| = \frac{|B|!}{(|B|-|A|)!}$

Preuve :

Soient a_1, a_2, a_3, … a_n les éléments de A. Par construction d'une fonction f, appliquée à chaque élément de A, on obtient |B| possibilités pour f (a_1), |B|-1 pour f(a_2), … |B|-n +1 pour f(a_n), pour un total de |B|*(|B|-1)*(|B|-2)*…*(|B|-n+1)

=|B|*(|B|-1)*(|B|-2)*…*(|B|-n+1)*(|B|-n)!/(|B|-n)!

=|B|!/(|B|-|A|)! Injections de A dans B

Lemme : (Hockey stick)

$$\sum_{i=k}^{r} \binom{i}{k} = \binom{r+1}{k+1}$$

Preuve, par induction :

Cas de base : r=k

$$\sum_{i=k}^{k} \binom{i}{k} = \binom{k}{k} = 1 = \binom{k+1}{k+1} = \binom{r+1}{k+1}$$

Étape d'induction :

Supposons que la proposition est valide pour une certaine valeur de r, alors, pour r+1

$$\sum_{i=k}^{r+1} \binom{i}{k} = \sum_{i=k}^{r} \binom{i}{k} + \binom{r+1}{k} = \binom{r+1}{k+1} + \binom{r+1}{k} = \frac{(r+1)!}{(r-k)!\,(k+1)!} + \frac{(r+1)!}{(r+1-k)!\,(k)!}$$

$$= \frac{(r+1)!\,(r+1-k)}{(r+1-k)(r-k)!\,(k+1)!} + \frac{(r+1)!\,(k+1)}{(r+1-k)!\,(k+1)(k)!}$$

$$= \frac{(r+1)!\,(k+1+r+1-k)}{(r-k+1)!\,(k+1)!} = \frac{(r+1)!\,(r+2)}{(r-k+1)!\,(k+1)!} = \binom{r+2}{k+1}$$

La proposition est donc vraie pour tout r >=k

Soient deux ensembles, A et B, contenant respectivement n et m > n éléments, I, l'ensemble des injections de A dans B et F, un élément de I,

Définissons X, la variable aléatoire représentant le nombre de valeurs communes entre un élément de I choisi selon une distribution uniforme, et F, une injection prise au hasard.

Soit P(X=k) la probabilité que X ait pour valeur k.

Considérant que $X \in \mathbb{N}$ alors, P(X=k) = P(X>= k) -P(X>= k-1).

P(X>=k)= nombre de fonctions ayant au moins k éléments en commun avec F/ |I|

Pour un nombre d'éléments en commun k avec F, il est possible de dénombrer le nombre de fonctions ayant autant, sinon plus de points en commun.

Parmi les n éléments de F, il est possible d'en fixer k, puis de laisser varier les autres valeurs de fonction, qui correspond à une fonction injective définie sur un ensemble de cardinalité n-k, et m-k. Par le lemme précédent, on dénombre (m-k)!/(m-n)! de ces fonctions. Toutefois, ce dénombrement implique de recompter plusieurs fois certains éléments. Plus particulièrement, pour chaque fonction comprenant i bons éléments, celle-ci est comptée C(i,k), étant donné que pour une fonction à i éléments communs avec F, il y a C(i,k) façons de choisir k éléments parmi les i. ainsi, pour toute valeur de i > =k, il faut retirer du décompte C(i,k) puis rajouter 1, afin de compter la fonction une seule fois.

Au final,

$$P(X \geq k) = \frac{(m-n)!}{m!}\left(\binom{n}{k}\frac{(m-k)!}{(m-n)!} - \sum_{i=k}^{n}\left(\binom{i}{k} - 1\right)\right)$$

En simplifiant la sommation, via le lemme plus haut, on obtient :

$$P(X \geq k) = \frac{(m-n)!}{m!}\left(\binom{n}{k}\frac{(m-k)!}{(m-n)!} + n - k + 1 - \binom{n+1}{k+1}\right)$$

Finalement, l'espérance de X est donnée par :

$$E(X) = \sum_{k=0}^{n} kP(X = k) = \sum_{k=0}^{n} kP(X \geq k) - kP(X \geq k+1)$$

$$= \sum_{k=0}^{n} kP(X \geq k) - \sum_{k=0}^{n} kP(X \geq k+1)$$

$$= \sum_{k=0}^{n} kP(X \geq k) - \sum_{k=1}^{n+1} (k-1)P(X \geq k)$$

$$= 0P(X \geq 0) - nP(X \geq n+1) + \sum_{k=1}^{n} kP(X \geq k) - \sum_{k=1}^{n} (k-1)P(X \geq k)$$

$$= 0 + 0 + \sum_{k=1}^{n} (k-k+1)P(X \geq k) = \sum_{k=1}^{n} P(X \geq k)$$

Ce qui permet d'obtenir la valeur d'espérance suivante :

$$E(X) = \sum_{k=1}^{n} P(X \geq k) = \sum_{k=1}^{n} \frac{(m-n)!}{m!} \left(\binom{n}{k} \frac{(m-k)!}{(m-n)!} + n - k + 1 - \binom{n+1}{k+1} \right)$$

$$= \frac{(m-n)!}{m!} \sum_{k=1}^{n} \left(\binom{n}{k} \frac{(m-k)!}{(m-n)!} + n - k + 1 - \binom{n+1}{k+1} \right)$$

$$= \frac{(m-n)!}{m!} \left(\sum_{k=1}^{n} \binom{n}{k} \frac{(m-k)!}{(m-n)!} + \sum_{k=1}^{n} (n+1) - \sum_{k=1}^{n} k - \sum_{k=1}^{n} \binom{n+1}{k+1} \right)$$

$$= \frac{(m-n)!}{m!} \left(\sum_{k=1}^{n} \binom{n}{k} \frac{(m-k)!}{(m-n)!} + (n+1)n - \frac{n(n+1)}{2} - \sum_{k=1}^{n} \binom{n+1}{k+1} \right)$$

$$\sum_{k=1}^{n} \binom{n+1}{k+1} = \sum_{i=2}^{n+1} \binom{n+1}{i} = \sum_{i=0}^{n+1} \binom{n+1}{i} - \binom{n+1}{0} - \binom{n+1}{1} = 2^{n+1} - 1 - (n+1)$$

$$= 2^{n+1} - n - 2$$

$$E(X) = \frac{(m-n)!}{m!} \left(\sum_{k=1}^{n} \binom{n}{k} \frac{(m-k)!}{(m-n)!} + \frac{n(n+1)}{2} - 2^{n+1} + n + 2 \right)$$

Pour les cas où m=n, le nombre d'injections de A vers B est de n!.

Si n=1, alors une seule injection existe de A vers B, et donc, toute fonction prise au hasard dans l'ensemble aura exactement un élément congruent à F

Pour n>2, il est possible d'observer que le nombre d'injections …

- à n éléments congruents à est égal à 1,
- à n-1 éléments congruents à I est égal à 0
- à k éléments congruents, pour n>k>0 est égal à C(n,k), multiplié par le nombre d'injections sans qu'il n'y ait d'élément congruent pour un cas où la cardinalité du domaine est de n-k

De plus, la distribution du nombre de fonctions ayant k éléments non congruents à F peut être définie à partir de la suite récursive suivante, définissant le nombre de fonctions élément de I de sorte qu'une telle fonction n'est congruente en aucun point de F :

$$s_0 = 1,\ s_1 = 0,\ s_{\mathrm{n}} = n! - \sum_{k=1}^{n} s_k \binom{n}{k}$$

Cette suite définit donc le nombre de configurations sans congruence pour chaque taille de n. Il est possible de mettre à profit cette suite pour le calcul de l'espérance mathématique, pour une taille n>=1 :

$$E(X) = \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k * \frac{1}{n!}\binom{n}{k} S_{n-k} = \frac{1}{n!}\sum_{k=1}^{n} k\binom{n}{k} S_{n-k}$$

$$= \frac{1}{n!}\sum_{k=1}^{n} k\frac{n!}{(n-k)!\,k!} S_{n-k}$$

$$= \frac{1}{n!}\sum_{i=0}^{n-1} (i+1)\frac{n!}{(n-i-1)!\,(i+1)!} S_{n-1-i}$$

$$= \frac{n}{n!}\sum_{i=0}^{n-1} \frac{(n-1)!}{(n-i-1)!\,i!} S_{n-1-i}$$

$$= \frac{1}{(n-1)!}\sum_{i=0}^{n-1} \binom{n-1}{i} S_{n-1-i} = \frac{1}{(n-1)!}\left(\binom{n-1}{0} S_{n-1} + \sum_{i=1}^{n-1} \binom{n-1}{i} S_{n-1-i}\right)$$

$$= \frac{1}{(n-1)!}(S_{n-1} + (n-1)! - S_{n-1}) = 1$$

Ainsi, l'espérance mathématique pour le nombre d'éléments congruents à I lorsqu'une fonction est choisie uniformément au hasard parmi les différentes bijections de deux ensembles finis et non vides est indépendante de la taille des ensembles.

Annexe 3. Métriques envisagées en cas d'amélioration de GenDR

Étant donné que l'évaluation automatique n'a pu avoir lieu, GenDR n'étant pas encore suffisamment développé pour pouvoir s'effacer derrière le travail, des métriques considérées pour l'évaluation n'ont pas pu être utilisées, mais pourraient l'être dans un travail futur. Ces métriques sont BLEU et ROUGE, décrites ci-après.

BLEU

La métrique BLEU (BiLingual Evaluation Understudy) est une méthode d'évaluation d'abord élaborée pour quantifier la qualité des traductions automatisées. Pour ce faire, un ensemble de phrases cibles sont choisies par des humains, afin d'avoir un étalon de ce qui constitue une bonne traduction. Par la suite, l'énoncé à évaluer est comparé aux cibles de traduction pour évaluer sa ressemblance avec la traduction. La comparaison se fonde sur la précision des N-grammes présents au sein de la phrase à évaluer, c'est-à-dire une suite de N éléments.

Plus concrètement, pour une séquence de *n* jetons, on retrouve (en incluant les répétitions), *n* unigrammes, *n-1* bigrammes, *n-3* trigrammes, et plus généralement *n-N+1* N-grammes. Pour évaluer le score BLEU, l'ensemble des N-grammes présents dans la cible est comparé aux N-grammes réellement produits. Ainsi, une phrase telle que *Ce livre est long* contient les 10 N-grammes du tableau 15, qui peut être comparée à une traduction *Le livre est long*, qui contient les 10 N-grammes du tableau 16, dont 6 sont présents dans la cible. Ainsi, le score BLEU sur l'ensemble des N-grammes possibles donnerait un score de 0,6 (ou 60 %) à l'évaluation.

N-gramme	Liste			
unigramme	Ce	livre	est	Long
bigramme	Ce livre	Livre est	Est long	
trigramme	Ce livre est	Livre est long		
4-gramme	Ce livre est long			

Tableau 15. – Suivi des N-grammes

N-gramme	Liste			
unigramme	Le	livre	est	long
bigramme	le livre	Livre est	Est long	
trigramme	le livre est	Livre est long		
4-gramme	le livre est long			

Tableau 16. – N-grammes de la phrase cible

Les N-grammes présents dans la phrase à évaluer qui sont présents dans la phrase cible contribuent à augmenter le score, avec une pénalité si les N-grammes sont présents plus souvent que nécessaire. Une fois ce score calculé pour les différents N-grammes, leur moyenne géométrique est évaluée, ce qui donne un aperçu de la précision moyenne des n-grammes présents dans la phrase à évaluer. Étant donné que le score ne se fie que sur la précision des n-grammes présents, une pénalité pour brièveté est mise en place : autrement, des énoncés plus courts que nécessaire auraient un score BLEU disproportionné. Ainsi, la précision moyenne est multipliée par un facteur de pénalité pour brièveté, qui se situe entre 0 et 1.

Il est possible de varier la taille des N-grammes et possiblement d'accepter différentes tailles de N-grammes, afin de capturer non seulement des jetons, mais également des locutions, phrasèmes, syntagmes et autres.

Toutefois, lorsque l'ensemble de phrases cibles est grand, la métrique perd en fiabilité, étant donné que la quantité de N-grammes pouvant apparaître dans la phrase à évaluer devient grande, et dès lors, la porte est ouverte à la surgénération.

BleuRT

Bien que BLEU semble être corrélé avec des jugements humains, il existe au moins une métrique plus proche d'une évaluation humaine : BleuRT. Cette métrique se base sur le modèle BERT (Devlin et coll., 2019), qui est un modèle neuronal. Concrètement, le modèle évalue à quel point un énoncé est plausible en se basant sur les énoncés qui ont servi à l'entraînement. Suivant les auteurs, il semble que cette métrique soit mieux corrélée au jugement humain que BLEU. Or, bien que cette métrique soit intéressante d'un point de vue de l'évaluation de traduction automatisée, elle ne semble pas adaptée à la tâche d'évaluation à laquelle l'on souhaite soumettre nos sorties de la génération. En effet, cette métrique requiert d'entraîner un modèle sur « quelques milliers » de phrases d'un sous-corpus (Sellam et coll., 2020). Étant donné que notre corpus ne contient à la base que 5045 énoncés, l'amputer de quelques milliers de phrases le rendrait trop petit. De plus, notre corpus contient notamment des énoncés incomplets (des réponses à des questions, pas nécessairement sous la forme d'un énoncé canonique), ce qui laisse présager qu'une métrique s'intéressant à la bonne formation, et entraîné principalement sur des phrases complètes (tirées de Wikipédia), ne serait pas adapté à la tâche en question.

ROUGE

Pour évaluer l'implémentation, il semble que les métriques de ROUGE [Recall-Oriented Understudy for Gisting Evaluation] (Lin, 2004) soient plus appropriées que le score BLEU. En effet, BLEU a été développé pour entraîner et évaluer des modèles de traduction pour lesquels il existe un petit nombre de phrases cibles qui soient jugées bonnes, tandis que les métriques ROUGE ont été développées notamment pour des tâches de résumé automatique, pour lesquelles on peut accepter davantage de flexibilité. Plus particulièrement, le score ROUGE-L est intéressant pour l'analyse, étant donné qu'il s'agit d'un score se basant sur la plus longue sous-séquence commune [PLSC] entre la sortie et la cible, ce qui assure un certain niveau de flexibilité.

Étant donné deux séquences, peu importe ce que sont leurs éléments, s'il est possible de sélectionner une sous-séquence commune aux deux suites initiales, de sorte qu'aucune sous-séquence n'est plus longue, cette sous-séquence est une plus longue sous-séquence commune.

Plus formellement, soient les éléments de la séquence a : a_1, a_2, … a_i, et de la séquence b : b_1, b_2, … b_j, une sous-séquence s : s_1, s_2, … s_k commune à a et à b est une séquence telle que pour chaque indice k identifiant des éléments de la sous-séquence, il existe un indice i de la séquence a et un indice j de la séquence b pour lequel ai = s_k=b_j, suivant la propriété d'ordre.

En effet, en prenant la PLSC, on évite un problème majeur de BLEU, c'est-à-dire qu'un n-gramme qui n'est pas présent dans la sortie que l'on cherche à tester pénalise, alors que diverses paraphrases sont possibles pour un même texte. Un autre score aurait le même effet, ROUGE-S, mais celui-ci se base sur des *skip-grams*, c'est-à-dire des suites de N éléments dans leur ordre d'apparition, sans qu'ils ne soient nécessairement consécutifs, ce qui a comme désavantage d'inclure des n-grammes composés d'unigrammes fréquents, mais peu utiles du point de vue sémantique, et même au niveau de l'évaluation, comme *le le* ou *le en*. Pour répondre à ce problème, (Lin, 2004) propose d'établir une limite sur le nombre d'éléments qui peut être ignoré dans l'évaluation, afin d'éviter d'avoir tous les n-grammes possibles, mais plutôt seulement ceux qui sont intéressants.

Il est bien de noter que le nombre de 2-skip-grammes possibles, sans borne, sur n éléments ordonnés est n(n-1)/2, soit, pour une phrase de 20 mots, soit l'énoncé moyen dans les structures, 190 skip-2-grammes. Considérant qu'il est possible d'estimer *grosso modo* qu'entre le tiers et la moitié des jetons utilisés sont de classe lexicale fermée, on peut estimer qu'entre le neuvième et le quart des skip-2-grammes sont composés uniquement d'éléments susceptibles de ne pas véhiculer d'information sémantique, mais plutôt d'être composés uniquement de jetons de classes lexicales fermées, ce qui est peu utile pour l'évaluation d'annotations sémantiques.

Toutefois, une telle limite reste dans le domaine de l'arbitraire, ce qui ne permet pas d'avoir une évaluation uniforme permettant une comparaison fidèle avec d'autres travaux. Ainsi, il semble que la méthode de la plus longue sous-séquence commune (ROUGE-L) est la métrique qui semble la plus appropriée pour la tâche d'évaluation en question. De plus, une variante, ROUGE-LU

permet de tenir compte du rapprochement au niveau de la plus longue sous-séquence commune, afin de permettre de résoudre certains problèmes de localité (par exemple, cette métrique attribue un meilleur score à la phrase *I like labradors and poodles* qu'à *I like labradors and hate poodles* si la cible était *I like all labradors and poodles*, étant donné que la plus longue sous-séquence commune dans les trois cas est *I like labradors and poodles*, mais que dans le premier cas, la sous-séquence se présente de façon contiguë, ce qui permet *a priori* de capturer davantage d'éléments de nature sémantique). La méthode utilisée par ROUGE-LU pour attribuer une pondération aux éléments contigus dépend de certains choix humains (entre autres, il faut choisir une fonction telle que $f(x+y)>f(x)+f(y)$, ce qui est relativement arbitraire, mais qui permet néanmoins de conserver une certaine relation d'ordre, la différence se situant au niveau de l'importance relative de la contiguïté.

Bien évidemment, il n'est pas possible d'avoir une métrique parfaite, toutefois, il semble que ROUGE soit relativement bien corrélé avec le jugement humain dans des tâches de génération de texte (Lin, 2004), ce qui permet d'en envisager l'utilisation.

Annexe 4. Protocole d'évaluation

Méthode d'évaluation des RSém générées à partir des annotations de FrameNet :

Le travail accompli

Des annotations sémantiques effectuées dans le formalisme de la sémantique des cadres ont été extraites du projet FrameNet. Ces différentes annotations ont ensuite été transformées en annotations en TST, dans le but d'enrichir la banque de RSém pouvant être ensuite utilisées en recherche en génération de texte, notamment.

Également, le travail de transformation de structures sémantiques d'un formalisme à l'autre permet d'établir une comparaison entre la TST et la sémantique des cadres, en relevant les similarités et différences des deux formalismes.

La sémantique des cadres :

Ce formalisme sémantique postule l'existence de cadres discursifs qui sont évoqués dans l'esprit des locuteurs lorsqu'une unité lexicale associée est mentionnée. Par exemple lorsque confronté à l'unité lexicale « observer », il est postulé par la sémantique des cadres que, dans l'esprit du locuteur, on a affaire au cadre « *perception active* »

> This frame contains perception words whose perceivers intentionally direct their attention to some entity or phenomenon in order to have a perceptual experience. For this reason we call the perceiver role in this frame "Perceiver_agentive"
>
> [Ce cadre contient des mots de perceptions pour lesquels les agents dirigent intentionnellement leur attention vers une entité ou un phénomène dans le but d'avoir une expérience perceptuelle. Pour cette raison nous désignons le rôle de la chose qui perçoit «*Perceiver_agentive*»]

Ainsi, chaque unité lexicale [désambiguisée] évoque un cadre discursif, et différentes unités lexicales peuvent être associées au même cadre, par exemple *toucher* pourrait aussi être associé à *perception active*.

Également, chaque cadre implique l'instanciation de différents participants sémantiques, qui occupent chacun un rôle. Par exemple, le cadre *Perception_agentive* instancie un agent qui perçoit une chose ou un phénomène, la chose ou le phénomène ainsi qu'un sens ou une partie du corps de l'agent percevant la chose ou le phénomène.

Les annotations de FrameNet, pour chaque unité lexicale, précisent quel cadre est évoqué, ainsi que les participants évoqués par l'unité lexicale, tels que présents dans le texte.

La tâche d'évaluation

En ce qui a trait à l'évaluation de la transformation des structures d'un formalisme à l'autre, les différentes structures devront être considérées sous les différents aspects suivants :

Statut de RSém :

- (2) : Est-ce que la structure à évaluer est, dans l'ensemble, une RSém, au sens de Mel'čuk ?
- (1) : Si elle n'est pas une bonne RSém en soi, est-elle bien formée en tenant compte de la structure d'entrée ?
- (0) : Si elle est une mauvaise RSém dans l'absolu

Conformité au texte :

- (1) Considérant la partie annotée du texte (voir la structure d'entrée), est-il possible de relier la RSém à ce qui est annoté ?

Numérotation des actants :

(Voir la page suivante pour davantage d'information)

- (2) Les actants de chaque sémantème sont numérotés parfaitement (il peut y avoir des trous
- (1) L'ordre relatif des différents actants est adéquat, malgré des « trous » dans la structure
- (0) Les actants ne sont pas numérotés adéquatement

Conjonctions :

(Voir la page suivante pour davantage d'information)

- (1) Si un « and » ou un « or » apparaît dans la phrase cible, il est relié aux termes joints de façon adéquate, et si plusieurs conjonctions apparaissent séparément, elles sont effectuées de façon correcte.

- (0) Un « and » ou un « or » apparaît dans la phrase cible, mais il n'est pas annoté correctement
- (NA) Non applicable : il n'y a pas de conjonction dans la phrase cible ** ou les éléments coordonnés ne sont pas annotés individuellement dans la structure et forment un tout.

Choix du nœud dominant :

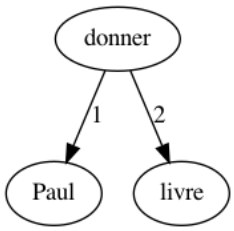
- (1) Le sémantème communicativement dominant est choisi adéquatement
- (0) Le sémantème communicativement dominant est mal choisi.

Autres non-conformités :

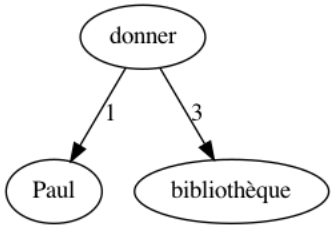
- Insérez une brève description

Précisions

Pour la numérotation des actants, il est possible d'avoir plusieurs actants, sans qu'ils ne soient tous explicitement mentionnés dans la structure, et qu'il y ait donc certains trous. Par exemple, pour un prédicat comme DONNER : Paul donne un livre (à la charité), on peut avoir une telle structure comme la suivante, qui omet le récipient du don, laissé intentionnellement vide :



Dans la même veine, il est possible d'avoir une structure telle que :

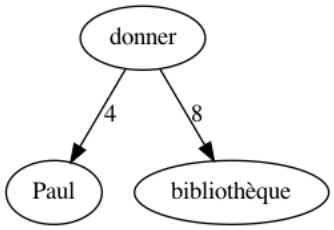


Qui permet de répondre à une question du type :

À quel organisme Paul a-t-il donné ses livres ?

Paul a donné à la bibliothèque

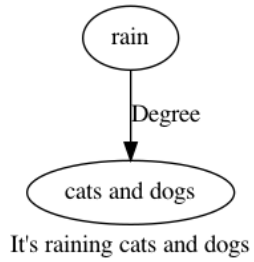
Par contre, il est également possible que les numéros soient bien ordonnés relativement, mais que la numérotation soit jugée comme étant « relativement bien », par la présence de 'trous' au sein de la structure. Ainsi, une annotation comme celle-ci :



est considérée comme relativement bien ordonnée, étant donné que les actants les plus importants ont un numéro plus près de 0 que les actants les moins importants, malgré que les actants soient mal numérotés (il n'y a pas trois actants qui pourraient logiquement s'insérer dans une description lexicographique avant le donneur dans donner)

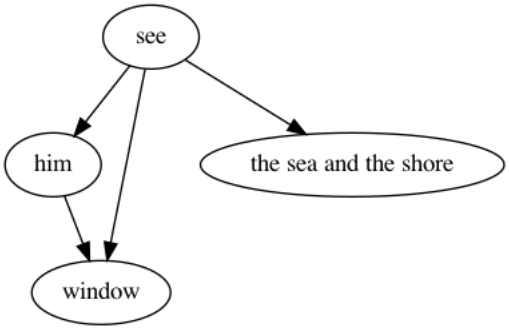
Conjonctions :

Les conjonctions 'and' et 'or' ont été ajoutées à l'annotation afin de relier des éléments. Toutefois, elles ne sont ajoutées que si elles sont nécessaires. En effet, si un sous-ensemble de la phrase est un ensemble d'éléments joints entre eux, pour qu'une conjonction soit ajoutées, il faut que les éléments soient annotés individuellement, soit comme cible d'annotation ou encore comme élément de cadre d'une autre cible d'annotation. Ainsi, une annotation comme :



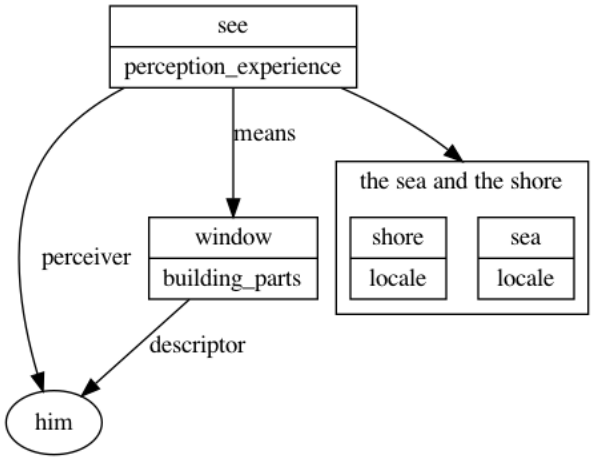
N'implique pas une annotation du 'and', étant donné qu'il apparaît dans une locution (Ici, l'équivalent du Magn() de Rain), et ni cats, ni dogs ne sont des cibles d'annotation. Ce cas de figure ne se limite pas aux expressions figées, dans certains cas, il est contre-productif de considérer

l'ajout d'une conjonction, malgré que la structure soit parfaitement compositionnelle. Par exemple, dans le cas d'une annotation incomplète :



He sees the sea and the shore through his window

Ici, tant la mer (the sea) que la rive (the shore) ne sont pas annotés indépendamment l'un de l'autre, alors qu'une annotation complète ressemblant à la suivante devrait voir les unités lexicale SEA et SHORE être reliées par une conjonction.



He sees the sea and the shore through his window