# La sémantique de rôle appliquée aux requis logiciels

Alice Breton



# La sémantique de rôle appliquée aux requis logiciels

**Soumis par :**

**Alice Breton**

**Sous la direction de :**

**Pierre André Ménard**

**Tuteur :**

**Claude Ponton**

**UFR LLASIC**

**Département Informatique intégrée en Langues, Lettres et Langage (I3L)**

---

MÉMOIRE DE MASTER 2 MENTION SCIENCES DU LANGAGE - 30 CRÉDITS

PARCOURS INDUSTRIES DE LA LANGUE

**Université Grenoble Alpes, Saint-Martin d'Hères, France**

**Année universitaire 2021 - 2022**

# Remerciements

En premier lieu, je tiens à remercier Pierre André, mon superviseur de stage, pour m'avoir guidée tout au long du projet. Merci d'avoir été présent et fortement impliqué. Tes nombreux conseils avisés m'ont permis d'enrichir mon cheminement académique.

J'aimerais également remercier l'ensemble du CRIM et plus particulièrement l'équipe TALN de m'avoir transmis un savoir immense.

Je remercie également l'équipe pédagogique du master SDL parcours IDL et plus particulièrement Claude. Merci d'avoir partagé ta passion et ta bienveillance tout au long de ces deux années de Master.

J'aimerais particulièrement remercier Margaux qui m'a motivée, questionnée et soutenue. Merci d'avoir tant sacrifié pour m'accompagner de l'autre côté même si tu n'aimes étrangement pas voler à 900km/h à 11 000 mètres au-dessus de l'océan.

Merci à la gang du Master IDL pour cette merveilleuse ambiance de classe.

J'aimerais remercier ma mère Claire de m'avoir encouragée tout au long de mon parcours académique.

**Résumé**

Il existe peu de corpus de documents techniques annotés en sémantique de rôle (SRL). Les corpus CoNLL05 et CoNLL2012 sont couramment utilisés comme données d'entraînement, de développement et de référence pour des modèles d'apprentissage neuronaux. Toutefois, ces deux corpus contiennent une faible quantité de données spécialisées et techniques. Nous proposons le corpus de requis logiciels CTeTex SRL annoté en sémantique de rôle. Le guide d'annotation de Proposition Bank a été suivi et enrichi de nouveaux rôles pour traiter adéquatement ce corpus hors domaine. Ce corpus est composé de 196 requis logiciels annotés en SRL avec un taux d'accord inter-annotateurs de 82 % pour la structure argumentale du verbe. Nous proposons une adaptation de la convention d'annotation pour tenir compte des particularités linguistiques de CTeTex SRL. Ce dernier peut être utilisé comme données de référence pour juger de l'applicabilité d'un modèle neuronal sur ce type de données spécialisées. Nous espérons que d'autres travaux suivront pour agrandir CTeTex SRL qui est le premier corpus de requis logiciel annoté en SRL.

Sémantique de rôle ; Proposition Bank ; CTeTex SRL, Données spécialisées, Accord intercodeur ; Données de référence

**Abstract**

Out of domain corpora in semantic role labeling (SRL) are quite rare in the literature. Frequently used corpora are CoNLL05 and CoNLL2012 as input and evaluating gold data for neural network models but they lack specialized data. We propose CTeTex SRL, the first, to our knowledge, software requirement specification (SRS) corpora annotated in SRL. We followed the Proposition Bank's guidelines and applied new conventions to better treat linguistic features of an out of domain dataset. CTeTex SRL is composed of 196 SRS each manually annotated in SRL with an interannotator agreement of 82% for argument labeling. We hope this first gold SRS corpora can be used to evaluate neural network driven semantic role labeling models and we hope our enriched annotation convention is applied to new SRS data to be used as a training dataset.

Semantic role labeling; Proposition Bank; CTeTex SRL, Out of domain, Interannotator agreement; Gold dataset

# Table des matières

# Table des figures

# Liste des tableaux

# Liste des abréviations, sigles et acronymes

| | |
|---|---|
| ADJ | Adjectival |
| ADV | Adverbials |
| ARG | Argument |
| ARGM | Argument modificateur |
| CAU | Cause |
| COM | Comitative |
| CRIM | Centre de Recherche Informatique de Montréal |
| CXN | Construction |
| DIR | Directional |
| DIS | Discourse |
| DMS | Dynamic Message Sign |
| DSP | Direct Speech |
| EPL | Expression polylexical |
| EXT | Extend |
| FE | Frame Elements |
| GOL | Goal |

| | |
|---|---|
| IAA | Accord inter-annotateurs |
| IVVES | Industrial-grade Verification and Validation of Evolving Systems |
| LIDILEM | Laboratoire de linguistique et didactique des langues étrangères et maternelles |
| LOC | Locative |
| LU | Unité lexicale |
| LVB | Light Verb |
| MNR | Manner |
| MOD | Modal |
| NEG | Negation |
| PRD | Secondary Predication |
| PropBank | Proposition Bank |
| PRP | Purpose |
| REC | Reciprocals |
| REL | Roleset |
| RL | Requis logiciel |
| SRL | Sémantique de rôle |
| SRS | Software requirement specification |
| TMP | Temporal |
| UD | Universal Dpendencies |

# Chapitre 1

# Introduction

Ce mémoire s'inscrit dans le cadre du *Master Sciences du langage parcours industries de la langue* de l'Université Grenoble Alpes. Pour clôturer ce parcours académique, un stage de recherche d'une durée de quatre à six mois devait être effectué en laboratoire universitaire ou en industrie. Ce mémoire propose de mettre en lumière mon cheminement de recherche en tant que Stagiaire en Traitement automatique des langues naturelles au Centre de Recherche Informatique de Montréal (CRIM).

Depuis 1985, le Centre de Recherche Informatique de Montréal est un acteur majeur dans le domaine de l'intelligence artificielle au Québec. En plus d'être un organisme à but non lucratif, il propose des solutions concrètes aux entreprises et organisations québécoises. Le CRIM est un centre de recherche appliquée et d'expertise. Il est un acteur clef dans différents domaines tels que le traitement automatique des langues naturelles autant du côté de l'écrit que de l'oral, la vision et la science des données. Ce centre de recherche est donc un pilier important dans le domaine de la recherche en intelligence artificielle au Québec (CRIM, 2022).

Mon stage de recherche est en lien avec le projet international : *Industrial-grade Verification and Validation of Evolving Systems* (IVVES). Ce projet a débuté en 2019 et prendra fin en décembre 2022. Plusieurs partenaires de recherche situés dans différents pays d'Europe ainsi qu'au Canada y participent dont le CRIM (ITEA3, 2022). Ce stage est dirigé par Pierre André Ménard, chercheur en Traitement automatique des langues naturelles. Du côté de l'Université Grenoble Alpes, mon enseignant en informatique et en Traitement automatique de la langue et chercheur au Laboratoire de linguistique et didactique des langues étrangères et maternelles (LIDILEM), Claude Ponton, supervise ce stage de recherche.

L'objectif de recherche est la constitution d'un corpus de référence de documents techniques annotés en sémantique de rôle. Plus précisément, ces documents techniques sont des requis logiciels (RL) annotés manuellement en sémantique de rôle en segment ou en *span-based* comme montre la Figure 1.1. Ce mémoire propose de mettre en lumière le cheminement de recherche dans le but d'atteindre une telle annotation. Ce corpus de référence permettra, entre autres, d'évaluer les performances de systèmes neuronaux utilisant des modèles de langues pour la prédiction des annotations en sémantique de rôle (SRL).

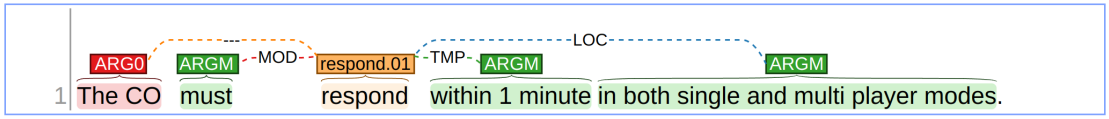



FIGURE 1.1 – Exemple d'attendu d'annotation d'un requis logiciel en sémantique de rôle

## 1.1 Problématique de recherche

Le but de IVVES, au plus haut niveau, est d'instaurer un projet d'envergure international regroupant plusieurs laboratoires et industries pour permettre de focaliser la recherche sur les activités de vérification et de validation des systèmes logiciels critiques en industrie. Ces systèmes logiciels sont généralement accompagnés de données textuelles langagières qui montrent leurs fonctionnements et leurs aptitudes. En effet, on y retrouve des requis fonctionnels ou non fonctionnels écrits par ses concepteurs et conceptrices. Les requis exposent les différentes attentes de comportements ou d'aptitudes auxquelles un système ou une partie du système, de la machine ou d'un logiciel doit répondre selon différentes conditions externes ou internes. En somme, ces requis ont pour but d'évaluer les aptitudes et les performances d'un système industriel.

— The system **shall do** *The requirement* (Le système **doit faire** *le requis.*)
— The system **shall be able to** *The requirement* (Le système **devrait pouvoir faire** *le requis.*)

La classe fonctionnelle montre les attentes comportementales. Elle définit si une réponse est obtenue selon différents facteurs ou critères. Elle peut être une réponse donnée par le système, mais cette attente peut aussi être qu'aucune réponse ne doit être donnée. Alors, si un système ne répond pas à l'ensemble des requis fonctionnels, ce dernier ne fonctionne pas complètement ou même intégralement. Un système sera donc jugé comme *valide, selon les normes,* ou bien, *non conforme* par son respect intégral ou non intégral des requis fonctionnels.

Pour ce qui est d'un requis non fonctionnel, il vise plus large en analysant la qualité de l'intégralité ou d'une composante d'un système par ses capacités ou ses per-

formances. Un requis non fonctionnel définit comment un système répond selon un contexte d'opération composé de différents facteurs externes ou internes. Il permet généralement d'évaluer si un système est intuitivement utilisable, sécuritaire ou effectif (Puzhevich, 2021).

Afin de vérifier si un système répond correctement à ces attentes de comportements ou de qualités, nous devons le tester sous un ensemble de conditions diverses. Prenons l'exemple d'une montre ayant une alarme programmée par l'utilisateur ou l'utilisatrice à sept heures du matin. L'attente est qu'une sonnerie sera l'action de la montre à sept heures. Pour tester si oui ou non la montre sera capable de répondre à cette attente, nous pouvons effectuer plusieurs tests :

— S'il est 6h55, la montre ne devrait pas produire de sonnerie.
— S'il est 6h59, la montre ne devrait pas produire de sonnerie.
— S'il est 7h, la montre devrait produire une sonnerie.
— S'il est 7h01, la montre ne devrait pas produire de sonnerie.

On peut constater avec l'exemple qu'il semblerait simple pour un humain natif de la langue utilisée par le requis (généralement l'anglais) de générer ces tests. Cependant, pour atteindre ce jeu de tests, plusieurs raisonnements complexes et logiques sont nécessaires. Plusieurs questions peuvent donc être soulevées. Quelles sont les étapes effectuées par l'humain avant de produire ces tests ? Est-ce qu'une machine ou un programme peut être capable d'effectuer ces étapes pour générer des tests ? Comment passer d'un requis logiciel à un jeu de tests ? Ces questions soulèvent l'importance de bien comprendre la faculté du langage dont l'humain est doté. Ainsi, avant de comprendre comment générer des jeux de test, il est crucial de comprendre comme un humain analyse sémantiquement un requis logiciel.

## 1.2 Objectifs de recherche

Ce mémoire de recherche propose de mettre en avant la structuration logique d'un requis depuis une analyse sémantique. Celle qui sera utilisée est la sémantique de rôle ou appelée *semantic role labeling* en anglais. L'acronyme SRL sera quelquefois utilisé pour respecter son utilisation fréquente dans la littérature. La SRL est couramment utilisée pour plusieurs tâches de compréhension de la langue (*Natural language understanding*) en tentant de répondre à la question suivante : qui à fait quoi à qui, et comment, quand et où? (*Who did What to Whom, and How, When and Where?*) (Palmer et al., 2010). Le but principal de cette recherche est de construire un corpus de requis logiciels annoté manuellement en SRL. Ces données annotées constitueront un corpus de référence qui permettra d'évaluer les performances de prédictions d'un modèle neuronal qui effectuera la tâche d'annotation en SRL.

En constituant un corpus de référence annoté en SRL, certaines problématiques seront parcourues dans le cadre de ce mémoire. Ce mémoire tentera d'exposer la difficulté d'appliquer certaines conventions du guide d'annotation proposée par Proposition Bank (PropBank) aux requis logiciels tout en faisant référence au but industriel de générer des jeux de tests à partir de ces requis. En effet, certaines caractéristiques linguistiques de ces documents techniques entraînent certains défis. Dans certains cas, cette convention sera non-applicable comme c'est également le cas de la convention d'*Universal Dependencies* pour de l'analyse syntaxique en dépendance (Hassert et al., 2021). Nous constatons que certains sens verbaux ne s'appliquent pas au domaine de l'informatique ou de l'ingénierie de systèmes (domaine auquel les requis logiciels sont rattachés). Ainsi, certaines propositions seront avancées pour pallier ces problématiques et pourront, entre autres, être appliquées à d'autres données textuelles de domaines spécialisés.

## 1.3 Une introduction à la sémantique de rôle

La sémantique de rôle propose une analyse sémantique d'une phrase. Cette analyse met en évidence l'action de la phrase qui est exprimée par le verbe dans certaines langues, dont la langue anglaise et française. L'élément central de la phrase dans cette théorie est donc le verbe. Ensuite, selon le sens qu'aura le verbe au sein la phrase, il est susceptible d'accepter des arguments porteurs de rôles. Les arguments sont généralement des syntagmes nominaux, propositionnels, adjectivaux ou adverbiaux. En d'autres mots, ce sont des groupes d'un ou de plusieurs mots-formes qui sont régis par le verbe. Enfin, chaque sens d'un mot-forme est appelé la lexie. Elle définit quels arguments du verbe seront possibles dans une phrase. En effet, chaque mot-forme d'une langue peut généralement être utilisé pour exprimer plusieurs sens. Par exemple, le mot-forme *clef* peut être utilisé dans différents contextes et il aura un sens distinct (proche ou éloigné).

(1) Je dois sortir mes **clefs** pour débarrer ma porte d'appartement.

(2) Ce projet sera la **clef** de notre succès.

(3) Sors-moi la **clef** anglaise du coffre à outils.

(4) J'utilise la **clef** de sol pour jouer cette partition.

(5) Elle a fait une **clef** de bras à son adversaire.

Tous les sens du mot-formes *clef* sont différents dans chacune de ces phrases. La phrase (1) contient le mot-forme *clef* défini comme un outil ayant une forme qui coordonne avec une serrure dans le but d'activer son mécanisme. L'exemple (2) contient le même mot-forme, mais son sens est qu'il donne accès à quelque chose et ici la *clef* donne accès au succès. L'exemple (3) utilise la *clef* comme un outil permettant de serrer ou de desserrer un écrou. L'exemple (4) contient le mot *clef* dans son sens musical qui définit une portée selon la hauteur des notes inscrites sur celle-

ci. Le dernier exemple (5) contient le mot *clef* comme une prise en sport de combat. De plus, on peut retrouver d'autres sens du mot *clef* qui sont utilisés dans certains domaines spécialisés comme l'informatique, la médecine, la biologie ou bien la construction. On peut le constater par les 14 entrées différentes de la ressource Terminium Plus (Gouvernement du Canada, 2022b). On peut donc conclure qu'en utilisant le même mot-forme, plusieurs sens peuvent lui être associés pour l'utiliser dans différents contextes. Plus précisément, on parle du vocable *clef* qui porte plusieurs lexies et ces lexies sont ses différents sens associés.

Les verbes peuvent aussi être utilisés pour exprimer plusieurs sens. Il est donc crucial de bien les distinguer. En effet, les arguments susceptibles d'être acceptés dépendront du sens du verbe qui les régit. Dans le cadre de ce mémoire, le contexte du verbe comprend les mots-formes qui l'entourent dans la phrase, mais aussi le contexte externe à la phrase. Pour rendre compte du sens qui lui est associé, son contexte d'utilisation est autant important que les mots qui l'entourent.

Les informations pragmatiques sont tout autant importantes puisqu'elles permettent de donner des indices qui guideront quel sens sera ciblé par le ou la locutrice qui utilise son système de langue pour rédiger un requis logiciel. Par exemple, prenons la phrase *I should run it* (Je devrais le courir/Je devrais le faire tourner). Son sens dépendra de son contexte. Si la personne qui l'a produit est entrain de songer à participer au Marathon de Montréal, son sens est que la personne croit qu'elle devrait participer au marathon. Par contre, si la personne qui produit cette phrase est derrière un écran, son sens est possiblement que la personne songe à faire tourner un script sur sa machine. Dans la phrase tirée de *Frames and the semantics of understanding* de Fillmore (1985), "We never open our presents until the morning" (Nous n'ouvrons jamais nos cadeaux avant le matin), en partageant la même culture, on

peut comprendre que les *presents* sont des cadeaux de Noël même sans l'avoir produit dans la phrase (Fillmore, 1985).

Tout compte fait, un même mot peut avoir plusieurs sens associés et c'est ce que la ressource verbale de Proposition Bank tente d'illustrer. Leurs sens distincts permettent de mettre en évidence ses arguments possibles. Par exemple, le verbe *to meet* a cinq sens selon cette ressource.

TABLEAU 1.1 – Description du prédicat *MEET* selon PropBank

| RoleSet ID | Sens | Role | Exemples |
|---|---|---|---|
| **meet.01** | Arrive at, achieve | **Arg0-PAG** : achiever, agent<br>**Arg1-PPT** : goal | Ford should meet the deadline easily.<br>**Arg0** : Ford<br>**Argm-mod** : should<br>**Rel** : meet<br>**Arg1** : the deadline<br>**Argm-mnr** : easily |
| **meet.02** | Come upon, become acquainted with initially | **Arg0-PAG** : meeter<br>**Arg1-COM** : person, entity, object being met | When hawk meets hawk<br>**Argm-tmp** : When<br>**Arg0** : hawk<br>**Rel** : meets<br>**Arg1** : hawk |
| **meet.03** | Get together (with), come together spatially | **Arg0-PAG** : one party<br>**Arg1-COM** : other party | Argentine negotiator Carlos Carballo will meet with banks this week.<br>**Arg0** : Argentine negotiator Carlos Carballo<br>**Argm-mod** : will<br>**Rel** : meet<br>**Arg1** : with banks<br>**Argm-tmp** : this week |
| **meet_up.04** | Get together | **Arg0-PAG** : first party (or only mentioned plural)<br>**Arg1-PPT** : other party (when separate mention) | so we met up just after work and got there early.<br>**Argm-dis** : so<br>**Arg0** : we<br>**Rel** : [met] [up]<br>**Argm-tmp** : just after work |
| **meet.05** | Answer, respond to | **Arg0-PAG** : agentive answerer<br>**Arg1-PPT** : thing met/answered<br>**Arg2-MNR** : with what | And they met the challenge with a brilliant and multifaceted campaign that garnered acclaim from many sources.<br>**Argm-dis** : And<br>**Arg0** : they<br>**Rel** : met<br>**Arg1** : the challenge<br>**Arg2** : with a brilliant and multifaceted campaign that garnered acclaim from many sources |

Le Tableau 1.1 montre les différentes descriptions de Proposition Bank pour le verbe *to meet* tirées de leur ressource verbale [1]. On peut constater cinq sens différents (dont une expression polylexicale *meet up*), une définition du sens associé, ses arguments possibles ainsi qu'une phrase exprimant l'utilisation de l'entrée du verbe en contexte avec ses arguments associés selon Proposition Bank.

Dans le cadre de ce mémoire, cette description de l'ensemble des verbes de l'anglais offerts par Proposition Bank est utilisée pour définir les sens des verbes et pour cibler ses arguments.

La multitude de sens associés à un verbe est l'un des exemples qui exposent l'ambiguïté de la langue. Ce mémoire explore cette particularité de la langue à travers l'annotation manuelle en SRL de requis logiciels. Cette théorie tente de comprendre le langage naturel. Bien qu'une brève introduction vient d'être proposée sur la SRL, cette notion sera approfondie subséquemment dans la suite du mémoire.

## 1.4 Le plan du mémoire

Ce **premier chapitre** 1 du mémoire propose de mettre en évidence les problématiques entourant le sujet de cette recherche. Une brève introduction au sujet des requis logiciels ainsi que de la sémantique de rôle est proposée.

Le **deuxième chapitre** 2 de ce mémoire propose un état de la littérature qui met en lumière les différentes approches théoriques du passé jusqu'à aujourd'hui. En débutant par la syntaxe structurale principalement proposée par Tesnière (1959), nous discuterons de la sémantique de rôle de Palmer et al. (2010), de corpus disponibles

1. https://verbs.colorado.edu/propbank/framesets-english-aliases/meet.html

et des requis.

Le **troisième chapitre** 3 est consacré à exposer une vue d'ensemble des corpus Pure et CTeTex ainsi que leurs caractéristiques linguistiques qui les limiteront des autres corpus provenant de journaux, de littérature ou de parole orale spontanée transcrite. Ces différences seront importantes dans le cadre de l'annotation.

Le **quatrième chapitre** 4 de ce mémoire propose une méthodologie d'annotation d'un corpus de référence par un guide d'annotation et un accord inter-annotateurs (IAA).

Le **cinquième chapitre** 5 met en lumière une discussion autour du corpus CTeTex SRL. Elle porte sur les résultats obtenus depuis la métrique de Kappa et les limites de la convention d'annotation.

Le **sixième chapitre** 6 expose un retour sur la recherche ainsi que certaines pistes d'évolution sur le domaine de l'analyse en sémantique de rôle sur un corpus de requis logiciel.

Finalement, ce mémoire de stage se positionne au sein de différents domaines d'étude. Le premier est la linguistique puisqu'il est le domaine analysant l'objet d'étude (le texte ou plus précisément la phrase). Ensuite, le traitement automatique des langues naturelles est un autre domaine que nous toucherons dans le cadre de ce mémoire puisqu'il tente de réduire la distance entre la faculté du langage et une machine.

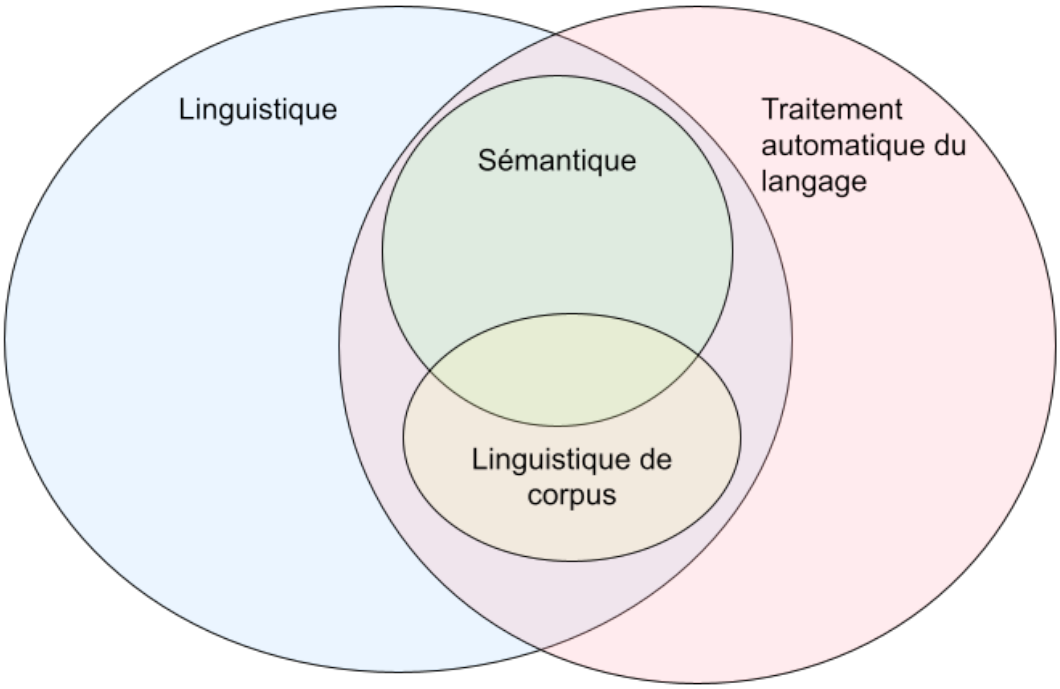



FIGURE 1.2 – Champs disciplinaires dans lesquels s'inscrit le mémoire

# Chapitre 2

# État de la littérature

Cette partie propose de mettre en évidence certaines théories de la linguistique qui serviront de bases pour étudier la constitution d'un corpus de référence. Nous passerons par le signe linguistique, au mot et enfin à l'objet d'étude : la phrase. Ce regard plus granulaire, en commençant par le signe linguistique de Saussure, nous guidera dans la compréhension et la limitation de l'unité plus grande qu'est la phrase (Saussure, 1916). De plus, cette étape est cruciale puisque Ferdinand de Saussure est connu comme un auteur clé du passage d'un sujet d'étude (la linguistique) non défini vers un domaine d'étude scientifique rigoureux par la définition de son unité d'étude et par le rôle qu'un linguiste aura afin d'étudier le langage. Par la suite, la sémantique de rôle sera abordée. En passant par la valence verbale, la ressource de FrameNet et de VerbNet et enfin Proposition Bank, la sémantique de rôle sera définie.

## 2.1 Le signe linguistique

Une langue est un système de signes et de règles. Un signe est une idée qui est associée à une forme ou un contenu (Polguère, 2016). Par exemple, le feu vert de circulation est un signe qui signifie le droit d'avancer sur une route empruntée par

plusieurs personnes. Ferdinand de Saussure s'est intéressé à la définition du sujet d'étude d'un linguiste. Il croit, entre autres, que c'est le système de signes linguistiques d'une langue qui doit l'être. Il le définit avec deux faces indissociables, d'un côté le signifiant et de l'autre le signifié. Cette association est arbitraire. Le côté du signifiant représente l'image acoustique ou l'abstraction du son qui est présent dans la phonologie (Saussure, 1916). En d'autres termes, c'est la représentation mentale d'un son qui est propre au système de la langue. Par exemple, une représentation mentale du mot arbre /aʁbʁ/ sera prononcée différemment selon différents facteurs. Par exemple, un enfant qui acquiert sa langue maternelle peut le prononcer [ab] puisque la consonne fricative uvulaire visée (ʁ) est l'une des dernières consonnes perçue et performée par l'enfant avec la langue maternelle française (Eimas et al., 1971). Ou simplement les différences phonétiques des variations régionales comme le verbe [nage] ou le nom [kʁɑb] du Bas-Saint-Laurent & Gaspésie versus [nɑge] & [kʁab] de la région montréalaise.

De l'autre côté, il y a le signifié qui représente sa conceptualisation. Cette conceptualisation se rapproche de son sens ou de sa signification. On peut le visualiser comme un nuage de concepts qui est propre au signifiant. Le mot /aʁbʁ/ peut signifier le concept de la plante, mais aussi d'un arbre de décisions (Saussure, 1916). On peut retrouver 20 entrées différentes du mot *arbre* sur la banque de données terminologiques et linguistiques Termium (Gouvernement du Canada, 2022a).

Le signe linguistique proposé par Saussure (1916) sera réutilisé, entres autres, dans le cadre de la sémantique de rôle et par la sémantique de cadres de Fillmore (1982). Le signe linguistique sera par la suite étendu à son rôle et sa combinatoire dans une phrase. En somme, Saussure a été un acteur qui a catalysé l'avancée scientifique de la linguistique en proposant l'étude approfondie du signe linguistique.

## 2.2 L'ambiguïté et les limites du mot "mot"

Suivant la conceptualisation du signe linguistique, le mot doit être défini puisqu'un ensemble de mots est présent dans le corpus de référence. Bien qu'un mot dans la langue est un signe linguistique puisqu'il est une forme (écrite, orale ou signée) associée à une idée ou une conceptualisation, cette section propose de le définir en introduisant les termes mot-forme et lexème.

En traitement automatique de la langue naturelle, il est courant de ne pas parler de *mot* lorsque nous traitons un corpus. C'est plutôt le *token* qui est couramment utilisé dans la littérature lorsqu'il est question de sous-parties d'un corpus. Chaque *token* d'un texte est ciblé à la suite d'une analyse lexicale ou d'une tokenisation d'un texte (segmentation). Ce traitement du texte est généralement utilisé avant toute tâche de traitement automatique de la langue comme la traduction, la classification d'informations, la reconnaissance vocale, le résumé de texte ou encore la détection d'entité nommée. L'analyse lexicale agit au niveau des mots d'un texte puisqu'elle tente de les cibler, c'est alors important de bien définir un *mot* et de comprendre ses limites lorsqu'il est question de décrire un corpus de référence. Peut-on le définir comme les éléments d'une phrase qui sont limités par un espace de chaque côté ? Comment tenir compte du premier mot d'un paragraphe, des acronymes ou des expressions polylexicales ? Une ponctuation peut-être entourée par des espaces, est-elle alors un *token* donc conséquemment un mot ?

Pour comprendre ce qu'est un *mot*, prenons simplement une phrase en exemple (6) et essayons de la découper :

(6) "On" "utilise" "ces" "pommes" "de" "terre" "parce" "qu'""elles" "ont" "plus" "d'""amidon".

Suivant cette découpe, on peut considérer qu'il y a 13 mots sans considérer la ponctuation finale de la phrase. Pourtant, certains éléments de cette phrase sont sémantiquement indissociables comme le nom *pomme de terre* et la conjonction *parce qu'*. Ce sont des locutions ou autrement appelées des expressions polylexicales (EPL). Ces éléments soulèvent une ambiguïté de ce qu'est un *mot* au sein d'une phrase. Conséquemment, un *mot* provoquera un défi d'identification par un système traitant une langue naturelle. En effet, pour justifier que ces éléments ne sont pas séparables, tentons d'intégrer un élément à l'intérieur et de remplacer la locution par un autre mot-forme ayant les mêmes propriétés :

(7) * On utilise ces pommes **énormes** de terre parce qu'elles ont plus d'amidon.

(8) * On utilise ces pommes de terre parce **de** qu'elles ont plus d'amidon.

(9) On utilise ces **patates** parce qu'elles ont plus d'amidon.

(10) On utilise ces pommes de terre, **car** elles ont plus d'amidon.

Dans la langue française, certaines expressions polylexicales sont attachées par des traits d'union comme *arc-en-ciel, chou-fleur, là-bas* ou encore *par-dessus*. Ces EPL sont contiguës et inséparables, elles ne forment pas une combinaison de leur ensemble, mais bien un ensemble fixe (Gross, 1982). Les locutions peuvent être nominale : *oeil de boeuf, mise en scène, mise au jeu* - adverbiale : *à tue-tête, à la volée, jusqu'alors* - verbale : *rendre l'âme, monter la garde* - prépositive : *grâce à, autour de* - conjonctive : *bien que, avant que*. Puisque les locutions se comportent en un tout indissociable, elles peuvent constituer un enjeu dans un système de traitement automatique du langage. En anglais, ce sont des *multi-word token*. La phrase (2.2), devrait plutôt être découpée en 10 mots-formes comme suit :

(6) "On" "utilise" "ces" "pommes de terre" "parce qu'" "elles" "ont" "plus" "d'""amidon".

Dans le cadre de la constitution d'un corpus de référence, le terme *mot* ne sera pas utilisé puisqu'il relève d'ambiguïté comme nous avons pu le voir avec les expressions polylexicales. Le terme *mot-forme*, introduit par Mel'cuk (1993), sera celui utilisé dans le cadre de ce mémoire. De plus, les mots-formes peuvent être fléchis en genre et en nombre, comme le mot-forme pomme**s** de terre. On parle de lexème lorsqu'on rapporte un mot-forme à sa forme canonique (pomme de terre). Ce terme est aussi introduit par Mel'cuk (1993) et par Quellet (1998). Les lexèmes sont généralement représentés en majuscule POMME DE TERRE. Les différentes entrées d'un dictionnaire sont généralement des lexèmes.

Dans le cadre du mémoire et dans le contexte du traitement automatique des langues, le terme *token* sera aussi utilisé. Un token est un élément qui peut porter plusieurs définitions selon la tâche finale. Il est possible de tokeniser un texte en mot-formes et en ponctuations, mais il est aussi possible de tokeniser un texte et de représenter chaque token en vecteur qui sera associé a une partie d'un mot, un mot ou un mot ainsi que ses informations contextuelles.

Maintenant que nous avons exploré les signes linguistiques, les mots-formes ainsi que les lexèmes, on peut agrandir notre fenêtre à la phrase. La suite de ce chapitre se consacre à son analyse par la syntaxe structurale Tesnière (1959), la sémantique de cadres (Fillmore, 1982) et ensuite notre cadre théorique la sémantique de rôle (Palmer et al., 2010).

## 2.3 Le structuralisme

Suivant la publication du Cours de linguistique générale en 1916, le domaine de la linguistique a connu un courant que l'on appelle le structuralisme. La langue est un

système de signes linguistiques organisés (Saussure, 1916). Cette organisation est régie par une structure et c'est ce que le courant du structuralisme tente d'étudier. La phrase est l'objet d'étude principal de ce courant puisque la syntaxe structurale présente les mots-formes comme ayant des relations hiérarchiques entre eux. Bien que l'objectif de recherche est de constituer un corpus de référence de documents techniques annotés en SRL, il est important de bien comprendre les théories antérieures pour comprendre ses enjeux.

### 2.3.1 Propositions de Lucien Tesnière

Lucien Tesnière est l'un des premiers linguistes importants dans l'analyse de la sémantique. Même s'il est connu pour avoir fondé l'analyse syntaxique en dépendance ou la grammaire de dépendance, son objet d'étude est la phrase. Elle est un ensemble organisé de mots-formes et de connexions. Les mots-formes d'une phrase cessent d'être limités à eux-mêmes, ils sont maintenant porteurs de connexions externes qui respectent des règles syntaxiques et sémantiques. Ces connexions sémantiques que portent les différents mots-formes respectent un rôle précis qui rend une phrase grammaticale. Pour obtenir une phrase grammaticale, le respect des règles syntaxiques est aussi important que les règles sémantiques. Selon Tesnière, ces deux domaines ne sont pas dissociables. Exposons tout d'abord la théorie structurale de Tesnière (1959).

Selon la théorie du structuralisme, la phrase est hiérarchique. Pour qu'il y ait des connexions entre les mots d'une phrase, elle doit être hiérarchique. L'action d'une phrase est le point de départ de son analyse. Par exemple, dans la phrase "Alfred parle.", en opposition aux présuppositions pragmatiques, il n'est pas question de penser qu' "il y a un homme qui s'appelle Alfred", ensuite que "quelqu'un parle". L'analyse structurale de Tesnière propose plutôt que nous pensons "Alfred fait l'ac-

tion de parler" ou "celui qui parle est Alfred" (Tesnière, 1959).

Le verbe est le centre de son analyse structurale. En d'autres mots, cet élément grammatical est la racine d'une phrase. Selon le structuralisme, les mots-formes respectent une structure profonde, ils sont donc porteurs de rôles. Le plan sémantique est la raison d'être de la structure syntaxique d'une phrase, sans son sens pour quelle raison une phrase serait exprimée ? Pourtant, Tesnière finit par définir la syntaxe structurale comme de la grammaire tandis qu'il définit la sémantique comme la pensée et l'abstraction de la langue qui ne relèvent pas de cette grammaire, mais de la psychologie et de la logique (Tesnière, 1959). Il exclut finalement la sémantique de son analyse de la phrase.

Il propose la distinction entre les mots vides (souvent appelé mots outils) et les mots pleins. Les mots vides sont les mots grammaticaux qui n'ont pas de charge sémantique, mais qui indiquent, précisent ou transforment la catégorie grammaticale d'un mot. Par exemple, les mots comme : *le, la, de, comme ou que* sont des mots vides. Les mots pleins ont une charge sémantique, c'est-à-dire qu'une forme est associée à une idée. On peut ici directement faire référence au signe linguistique de Ferdinand de Saussure. Suivant ce discours, si la sémantique doit être utilisée pour juger d'une charge sémantique, et donc, juger qu'un mot est vide ou plein, ne serait-ce pas important d'accepter la sémantique sur un même stade que la syntaxe ?

### 2.3.2 Le noeud verbal de Lucien Tesnière

Comme nous l'avons précédemment mentionné, le verbe est l'élément central de la théorie de Tesnière. Il l'est tout autant dans le cadre de cette recherche d'application de la sémantique de rôle sur des requis logiciels. Selon Tesnière, la structure principale d'une phrase est le nœud verbal dans plusieurs langues comme le français.

Il mentionne ‘Le petit drame’ (aussi appelé le *stemma*) comme l’expression d’une phrase, c’est-à-dire, un drame qui comporte un procès (le verbe) qui possèdent des acteurs (les actants) et des circonstants (Tesnière, 1959). Cette proposition se rapproche de la théorie de sémantique de rôle. Par exemple, dans la phrase *Jean mange une pomme.*, le procès est exprimé par l’action de manger. Ses actants ou les éléments participants à cette action sont *Jean* et *une pomme*. Tesnière les appelle les substantifs ou les actants du verbe. Pour ce qui est des circonstants, elles expriment le temps, le lieu ou la manière d’une action d’une phrase. On constatera par la Figure 2.1 qu’il définira ensuite la phrase comme un stemma.

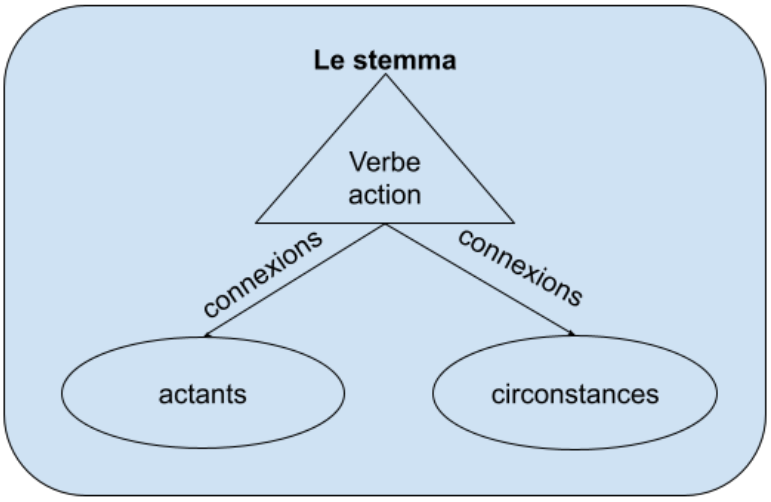



FIGURE 2.1 – La structure d’une phrase représentée par le Stemma

Bref, le verbe est l’aspect central d’une phrase et il est susceptible de régir un ou plusieurs actants. Tesnière introduit alors le concept de la valence verbale par le fait qu’un verbe puisse régir des actants qui participent au *petit drame*. Par exemple, le verbe *lire* peut régir une personne qui lit (actant) et ce qui est lu (instrument).

Cette valence se rapproche de la sémantique de rôle. En effet, les actants seront plutôt considérés comme des arguments et cette valence sera la *proposition* du verbe.

### 2.3.3 Le verbe

Même si le verbe est une notion couramment connue par les locuteurs et locutrices d'une langue, mettons en lumière la définition de Lucien Tesnière. Il exprime un procès. Il suggère de définir un procès comme toute forme qui englobe à la fois les verbes d'état et les verbes d'action. La distinction entre un verbe d'état et un verbe d'action est importante puisque l'analyse de la phrase implique des particularités propres aux verbes d'état ou aux verbes d'action dans la théorie du structuralisme. Cette distinction sera présente dans le choix des actants accompagnant le verbe. Nous verrons plus tard des exemples concrets. Il est aussi important de mentionner que les verbes d'état ne sont pas forcément des verbes intransitifs et que les verbes d'action sont transitifs.

### 2.3.4 Les actants

Les actants présents dans une phrase dépendent intégralement du verbe. Toutefois, certains verbes n'ont pas d'actants, comme les verbes météorologiques. Dans la phrase 'il pleut', le terme 'il' ne participe pas à l'action de pleuvoir. Le verbe pleuvoir sera donc défini comme avalent. Les actants sont les éléments qui participent à l'action de la phrase ou le procès dans Tesnière (1959). L'ensemble des verbes ne comporte pas toujours le même nombre d'actants.

Certains verbes comportent un seul actant, ils sont monovalents. Par exemple, les verbes comme *tomber, dormir* ou *courir*. Dans la phrase *Alain tombe*, l'actant du verbe *tomber* est *Alain*. Cet actant peut être rempli par tout autre être ou objet ayant

la capacité de tomber.

Certains verbes comportent deux actants, ils sont bivalents. Les verbes comme *frapper, manger* ou *lire*. Dans la phrase "Éric frappe Bernard", les deux actants du verbe *frapper* sont Éric et Bernard. Chacun de ces actants peut être rempli par tout autre être ou objet ayant la capacité de frapper quelque chose ou quelqu'un.

Certains verbes comportent trois actants, ils sont trivalents, comme le verbe "donner". Dans la phrase Jean donne ce livre à Bernard, les trois actants du verbe donner sont Jean, ce livre et Bernard.

Dans Tesnière (1959), il y a jusqu'à trois actants possibles pour les verbes du français, mais il en existe jusqu'à six dans la langue anglaise. Nous le constaterons plus tard dans la théorie de la sémantique de cadres de Fillmore (1982).

### 2.3.5 Les circonstants

Chaque verbe à ses actants, mais une phrase peut être composés de circonstants. Contrairement aux nombres d'actants qui sont limités par le verbe dans la phrase, les circonstants n'ont pas de limite stricte. Ils permettent d'exprimer des informations de localisation dans l'espace et dans le temps et de marquer des relations. Ils sont nommés *Adverbe* selon Tesnière (1959), par contre, ces éléments sont généralement considérés comme des arguments modificateurs. Nous en ferons référence entre autres dans la théorie de la sémantique de rôle.

### 2.3.6 La valence verbale

Nous avons vu précédemment qu'un verbe peut régir une quantité X d'actants. Ce nombre d'actants régis dépend de la valence du verbe. Tesnière parle de susceptibilité du verbe. Cette notion propose de mettre en évidence la nature intrinsèque ou même psychologique du verbe selon Tesnière. On parle de susceptibilité puisqu'une phrase peut comporter un verbe bivalent comme dans l'exemple 2.2 employé dans Tesnière (1959) : *Alfred chante* vs *Alfred chante une chanson.*

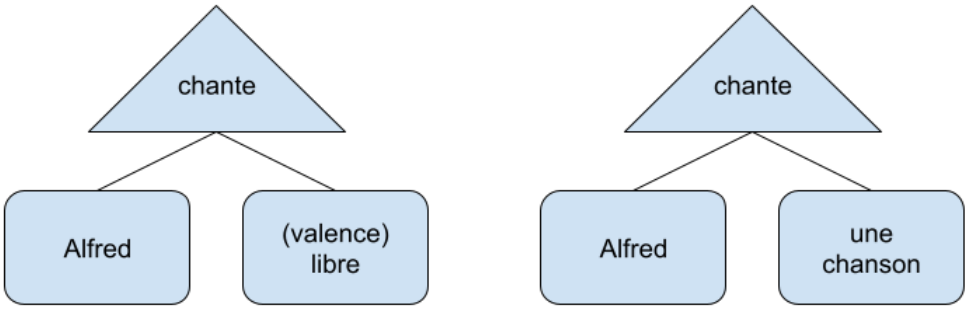



FIGURE 2.2 – La valence libre

La Figure 2.2 montre le verbe chanter qui est susceptible de régir deux actants, mais la première phrase *Alfred chante* [*valence libre*] n'en contient qu'un. C'est ce qu'est la valence libre selon Tesnière. Prenons l'exemple du verbe *manger* dans la phrase *Julie mange.*, même s'il y a qu'un actant *Julie*, nous savons qu'il y a forcement quelque chose de mangée par Julie. Cette information n'est pas présente dans la phrase en elle-même, mais le verbe permet quand même de la produire. Cette valence libre s'apparente aux maximes de Grice (Grice, 1975), théorisé quelques années plus tard. Il définit l'acte de langage comme une activité coopérative qui a un but. Il propose quatre maximes : de quantité, de qualité, de relation et de manière (Grice, 1975). En somme, cette valence libre ou la susceptibilité du verbe selon Lucien Tesnière permettra de respecter les maximes de Grice dans l'acte du langage.

Les propositions de Tesnière sont encore acceptées dans l'étude de la sémantique et plus précisément en sémantique de rôle. Sa théorie qui est exprimée par son livre *Éléments de syntaxe structurale* Tesnière (1959) est fondatrice de la grammaire en dépendance, de l'analyse de la phrase par le stemma qui régit les actants et la valence verbale. Même s'il s'inscrit dans le domaine de la syntaxe, les actants et la valence verbale seront réutilisés dans plusieurs théories sémantiques comme la sémantique de cadres et de rôles. Il sera critiqué plus tard par d'autres linguistes puisque la valence verbale dépend du sens du verbe au sein de la phrase. En d'autres mots, les actants sont régis par le sens du verbe. Pourtant, le sens verbal n'est pas mentionné dans la théorie du structuralisme.

## 2.4 La grammaire de cas

Pour étudier les langues, la sémantique est autant importante que la syntaxe. Ce sont les travaux, entre autres, de Charles J. Fillmore qui catalysent le besoin d'inclure le domaine de la sémantique dans l'étude des langues. Il propose une définition de la phrase comme suit : "The sentence in its basic structure consists of a verb and one or more noun phrases, each associated with the verb in a particular case relationship" (Fillmore, 1982). Nous tenterons d'expliciter cette citation dans cette présente section.

La grammaire de cas propose de mettre en lumière les règles qui sont régies par le verbe (Fillmore, 1968). Le centre de l'analyse est donc le verbe. Depuis *The Case for Case* publié en 1968, il expose son étude sémantique sur les verbes de l'anglais. Selon Fillmore, le verbe a des propriétés internes de cas tels que l'Agent, le Patient ou l'Instrument.

> "I find helpful in distinguishing the operation and the goals of frame semantics from those of standard views of compositional semantics is between a grammar and a set of tools - tools like hammers and knives, but also like clocks and shoes and pencils. To know about tools is to know what they look like and what they are made of - the phonology and morphology, so to speak - but it is also to know what people use them for, why people are interested in doing the things that they use them for, and maybe even what kinds of people use them." (Fillmore, 1982)

On constate par cette citation que Fillmore propose de limiter son objet d'étude en le séparant des analyses de la sémantique compositionnelle standard. Elle propose, dans une certaine mesure, que la syntaxe soit au cœur de l'échelle de l'analyse de la phrase et que la sémantique doive donc être utilisée en second lieu. Au contraire, Fillmore propose une analyse plus ouverte des règles (outils) en étudiant les personnes qui performent leur langue en utilisant des outils. Il s'inspire des travaux de Tesnière sur la grammaire en dépendance et la théorie de la valence sémantique, précédemment exposées, pour développer son étude. Rappelons-nous que la valence sémantique propose, selon Tesnière, de mettre le verbe ou le procès au centre de l'analyse de la phrase et que ce dernier régit des actants. L'auteur de la grammaire de cas, propose deux paramètres aux verbes : *case frames* & *rule features* ou autrement appelé *frame features*. Le premier paramètre est défini par les cas qui entourent un verbe : un **agent** qui est l'acteur ou l'actrice du verbe, le **patient** qui est l'objet auquel l'acteur ou l'actrice influence son état et finalement l'**instrument** qui permet cette influence d'état de l'objet. Enfin, ces trois cas sont : l'agent, le patient et l'instrument. Le deuxième paramètre propose de mettre en évidence les possibilités des cas qu'un verbe accepte. Par exemple, le verbe "ouvrir" peut régir deux cas : une personne (agent) ouvre quelque chose (patient). Cette théorie proposée par Fillmore

est le début du dictionnaire de valence sémantique qui sera appelé FrameNet (Baker et al., 1998).

### 2.4.1 La sémantique de cadres

Comme Tesnière, Fillmore croit que l'aspect central de la phrase est son action qui régit des éléments porteurs de rôles sémantiques. C'est pour cette raison que le "sujet" dans la phrase n'a toujours pas été abordé dans le cadre de ce mémoire. Il n'y a pas de limite entre le "sujet", "verbe", "objet" comme dans la grammaire traditionnelle. En fait, le "sujet" et l'"objet" sont plutôt subordonnés au verbe (Fillmore, 1968).

(11) John broke the window.

(12) A hammer broke the window.

(13) John broke the window with a hammer.

Ces exemples sont tirés de Fillmore (1968). Il montre avec l'exemple (11) que le sujet (John) est l'agent du verbe, en (12) le sujet est un instrument et qu'en (13) l'agent et l'instrument sont présents dans la phrase sans qu'ils soient tous les deux sujets du verbe. Les phrases (14) et (15) ne sont pas sémantiquement permises.

(14) [*]John and a hammer broke the window.

(15) [*]A hammer broke the glass with a chisel.

Par ces exemples, on constate que la notion de rôle est plus importante que la notion de "sujet" ou d'"objet". Les phrases (14) et (15) ne respectent pas la structure de cadres du verbe *broke* et c'est la raison qui explique l'agrammaticalité de celles-ci. En d'autres mots, c'est par les cadres ou plus précisément les relations de rôles sémantiques entre le verbe et ses segments qui justifient la grammaticalité ou non d'une phrase. Ces différents rôles sont les *cadres* de Fillmore.

TABLEAU 2.1 – Les *Cases* de Fillmore

| | |
|---|---|
| Agentive (A) | Le cas de l'instigateur de l'action généralement animé. |
| Instrumental (I) | Le cas de la force ou de l'objet inanimé qui est impliqué par raison quelconque dans l'action ou l'état. |
| Dative (D) | Le cas où l'être animé est affecté par l'état ou l'action. |
| Factitive (F) | Le cas de l'objet ou de l'être résultant de l'action ou de l'état identifié par le verbe, ou compris comme faisant partie du sens du verbe. |
| Locative (L) | Le cas qui identifie l'emplacement ou l'orientation spatiale de l'état ou de l'action. |
| Objective (O) | Le cas le plus sémantiquement neutre, le cas de tout ce qui est représentable par un nom dont le rôle dans l'action ou l'état identifié par le verbe est identifié par l'interprétation sémantique du verbe lui-même. |

Le Tableau 2.1 montre l'ensemble des premiers *cadres* de la théorie de la sémantique de cadres tirées de son article fondateur *The Case for Case* de Fillmore (1968). Ils définissent le rôle possible d'un argument du verbe d'une phrase. Par contre, plus tard dans le cadre de ce mémoire, la liste sera actualisée par la suppression et l'ajout de quelques cadres (*Agent, Patient, Recipient, Goal* et bien d'autres).

Tout compte fait, Charles J. Fillmore propose une analyse sémantique de la phrase par sa grammaire de cas qui est par la suite étendue à sa théorie de sémantique de cadres. Il met, comme la sémantique de rôle, le verbe au centre de son analyse. Ce dernier régit des cadres porteurs de rôles précis.

## 2.5 Les ressources verbales en sémantique de rôle

Cette section décrit les deux ressources verbales FrameNet et VerbNet. FrameNet est le produit de la théorie de la sémantique de cadres de Fillmore (Fillmore, 1982). Ensuite, VerbNet s'inspire des classes de Levin (1993) et propose presque le double de

classes verbales (Schuler, 2006). Ces deux ressources inspirerons les travaux du projet Proposition Bank présenté à la Section 2.6 suivante.

### 2.5.1 Le projet FrameNet

Cette partie présente le projet Berkeley FrameNet [1] (Baker et al., 1998). Datant de 1997, le projet s'est déroulé sur trois ans. Trois chercheurs Collin F. Baker, Charles J. Fillmore et John B. Lowe sont les instigateurs du projet. Les buts sont de mettre en évidence une analyse sémantique et syntaxique ainsi que la représentation de la valence pour une multitude de mots-formes de l'anglais. Ils bâtissent la ressource FrameNet. Ce projet est le fruit de la théorie de la Sémantique de cadres de Charles J. Fillmore (Fillmore, 1976, 1977, 1982, 1985; Fillmore and Baker, 2001; Heine and Narrog, 2015). En effet, il écrit même plus de dix ans plus tôt sur cet objectif de bâtir la ressource de sémantique de cadres pour l'anglais (Fillmore, 1982; Baker et al., 1998).

La ressource FrameNet associe 958 *frames* ou cadres à 2500 éléments de cadre ou *Frames Elements* (FE). Par exemple, le cadre *Apply_heat* inclut les FE *Cook, Food, Container, et Heating Instrument.* Les mots-formes qui invoquent ces cadres sont appelés des unités lexicales (LU) (*fry, bake, boil, broil ou grill*). Les membres du cadre *Apply_heat* sont *bake, barbecue, blanch, boil, braise, broil, brown, etc.* La ressource ne présente pas seulement des verbes, mais aussi des noms et des adjectifs. La phrase suivante (16) montre un exemple d'annotation du *Frame Apply_Heat* avec ses *Frame Elements (Cook, Food, Heating_instrument)*. Elle est tirée du site web du projet FrameNet [2].

(16) ... [**Cook** the boys] ... GRILL [**Food** their catches] [**Heating_instrument** on an open fire]

1. `https://framenet.icsi.berkeley.edu/fndrupal/`
2. `https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet`

### 2.5.2 Le projet VerbNet

Bien que FrameNet propose une ressource de plusieurs verbes de l'anglais de cadres sémantiques, il n'est pas question d'évaluation d'annotation d'une telle ressource et il n'y a pas non plus la présence de polysémie. Le projet VerbNet (Schuler, 2006) tente de répondre à ces problématiques en proposant une ressource verbale qui apporte plus d'informations que celle de FrameNet (Kipper et al., 2000). Cette ressource inspire justement les travaux du projet Proposition Bank. Pour rappel, c'est la convention d'annotation de Proposition Bank qui est utilisée dans le cadre de ce mémoire de recherche. Puisque ses auteurs et autrices utilisent VerbNet, regardons plus précisément comment cette ressource a été construite.

Le projet VerbNet s'inspire des travaux de Beth Levin sur les classes verbales (Levin, 1993). Les classes sont un regroupement de verbes qui partagent un comportement syntaxique et sémantique similaire. La ressource de Levin (1993) a un total de 240 classes avec 47 classes de premier niveau et 193 classes de deuxième et troisième niveau. VerbNet a proposé d'étendre à 440 classes pour 1000 lexèmes avec un niveau de classe ajouté pour un total de 4 niveaux. Ces classes sont une représentation de haut niveau et commune à plusieurs verbes. Par exemple, la classe give-13.1 rassemble les verbes *lend, loan, pass, peddle, refund, render* ou bien spray-9.7-1 qui rassemblent les verbes *wash, brush, drizzle, heng, pump* et bien d'autres.

Chaque lexème est associé à une liste finie de rôles sémantiques. Dans la ressource de VerbNet [1], il y a 24 rôles sémantiques différents :

TABLEAU 2.2 – Les rôles thématiques de VerbNet

| | |
|---|---|
| Actor | Lorsque deux arguments sont symétriques (pseudo-agents) |
| Agent | Un humain ou un objet animé |
| Asset | La conversion de devise |
| Attribute | L'attribut du patient ou du thème qui se réfère à la qualité de quelque chose entrain de vivre un changement |
| Beneficiary | L'entité qui bénéficie de l'action |
| Cause | Utilisé par des classes qui relèvent de verbes psychologiques et de verbes impliquant le corps |
| Location | La destination non spécifique, la source ou l'endroit, souvent introduit par un locatif ou une *path preposition (above, behind, over)* |
| Destination | Le destination d'un mouvement ou la direction vers laquelle le mouvement est dirigé |
| Source | Le point de départ d'un mouvement |
| Experiencer | Un participant conscient d'expérimenter quelque chose |
| Extent | La spécification de l'intervalle ou du degré du changement |
| Instrument | L'objet (ou la force) qui est en contact avec un autre objet ou qui cause le changement d'état d'un autre objet |
| Material | L'état d'un produit non transformé |
| Product | L'état final d'un produit transformé |
| Patient | Les patients subissant un changement d'état ou ayant déjà subit un changement d'état |
| Predicate | Complément prédicatif |
| Recipient | La cible d'un transfert |
| Stimulus | Utilisé par les verbes de perception pour des événements ou des objets qui suscitent une réponse d'un expérimentateur |
| Theme | Une localisation ou subissant un changement de localisation |
| Time | L'expression du temps |
| Topic | Sujet de communication qui exprime le thème/sujet de la conversation ou du message |

1. `https://verbs.colorado.edu/~mpalmer/projects/verbnet.html`

Pour conclure, VerbNet de Baker et al. (1998) et FrameNet de Kipper et al. (2000) sont les premières ressources verbales qui traitent de la sémantique de rôle. Ce sont deux outils qui proposent des classes verbales qui regroupent un ensemble de verbes qui partagent des propriétés syntaxiques et sémantiques. La lacune de ces deux ressources est l'exposition claire de la structure argumentale de chacun des verbes. Cet aspect sera traité dans le projet Proposition Bank.

## 2.6 Le projet Proposition Bank et ses corpus

Cette prochaine section du mémoire tente de décrire Proposition Bank. Ce dernier explore de façon plus granulaire la sémantique de rôle par l'instauration de nombreux projets de recherche entourant cette théorie et la constitution d'une ressource verbale, d'un guide d'annotation et de corpus annotés.

### 2.6.1 The Proposition Bank

Le but principal de Proposition Bank est de construire un corpus annoté qui pourra par la suite être réutilisé comme données d'entraînement pour de l'apprentissage machine supervisée. Cette ressource est couramment utilisée dans la littérature lorsque la tâche de prédiction de modèle est l'annotation en sémantique de rôle. Cette partie se concentre sur une description du projet Proposition Bank. La Section 4.1 décrit plus précisément les différentes méthodes d'annotation de leur corpus. Il est également possible de se référer directement au guide d'annotation officiel [1].

L'approche de Proposition Bank est de placer le sens du verbe en avant-plan pour cibler ses arguments. C'est ce que les auteurs et autrices de la ressource appellent le

1. https://github.com/propbank/propbank-documentation/raw/master/annotation-guidelines/Propbank-Annotation-Guidelines.pdf

*roleset ID*. On pourra constater qu'un verbe appelé prédicat aura plusieurs entrées avec un sens différent. Ensuite, chaque entrée d'un verbe aura ses arguments possibles. Ici, la ressource fait directement référence aux lexies (*roleset IDs*) d'un lexème (*predicate*). Un prédicat aura forcément une lexie ou un *roleset ID* avec un sens défini.

PropBank utilise une notation simple pour les arguments du verbe : ARG0 à ARG5. Les deux premiers arguments 0 & 1 respectent généralement un rôle d'agent pour l'ARG0 et de patient pour l'ARG1. Les autres arguments (2-5) respectent moins un rôle spécifique associé à chacun des arguments. Par exemple, un ARG2 ou un ARG3 peuvent tous les deux porter le rôle d'instrument selon le *roleset ID* consulté. Cette approche est utilisée puisque les étiquettes de rôles sémantiques n'ont pas de consensus en littérature. Comme il a été mentionné plus haut, certaines théories proposent seulement trois arguments (agent, patient ou instrument), d'autres en suggèrent 5 ou même 24. Pour répondre à cette problématique, PropBank suggère plutôt d'utiliser des étiquettes qui n'ont pas de rôle défini. Par contre, ce choix peut entraîner certaines problématiques comme la difficulté de faire des inférences ou de généraliser les rôles des arguments. En effet, puisque chaque type d'argument est spécifique à un *roleset ID* et non à un rôle, il sera difficile de déterminer le rôle que joue précisément un argument. On constatera alors que la plupart des modèles de prédiction d'argument ont de moins bonnes performances pour les arguments 2 à 5.

Proposition Bank offre également la possibilité de cibler des arguments modificateurs au sein d'une phrase (ARGM). Ces éléments font référence aux circonstants de la théorie du structuralisme de Tesnière (1959). Les arguments modificateurs sont les éléments facultatifs d'une phrase comme l'adverbe *demain* placé en début d'une phrase. Cet adverbe n'est généralement pas régi par un verbe d'une phrase, mais il apporte une information de *temporalité* à la phrase. Par exemple, selon Propo-

sition Bank l'élément sera annoté comme un *ARGM-TMP*. Cet aspect est différent des autres ressources comme FrameNet et VerbNet. Contrairement aux deux précédentes ressources, Proposition Bank tente de décrire l'ensemble des possibilités sémantiques en partant d'un corpus. Généralement, certains arguments ne seront pas régis par la structure argumentale du verbe de la phrase, ils seront donc des arguments modificateurs (ARGM).

TABLEAU 2.3 – Structure argumentale du *roleset ID bake.01*

| bake.01 : create via heat |
|---|
| Arg0-PAG : baker |
| Arg1-PPT : creation |
| Arg2-VSP : ingredients |
| Arg3-GOL : benefactive |

(17) Today whole grains are freshly ground every day and baked into bread.

**Argm-tmp : Today**

Arg2 : whole grains

Rel : baked

Arg1 : into bread

Dans l'exemple (17) tiré de la ressource verbale de PropBank [1], on constate qu'un argument est noté comme ARGM-TMP par Proposition Bank. En effet, l'élément *today* ne fait pas partie de la structure argumentale du verbe *bake* comme on le constate par le Tableau 2.3. Il est donc noté comme un argument modificateur dans la phrase en (17). Il faut noter également que le verbe *ground* est présent dans la phrase, mais sa structure argumentale n'est pas exposée dans le cadre de cet exemple.

En bref, Proposition Bank est le seul projet qui instaure un guide d'annotation rigoureux qui est utilisé pour la constitution de leur corpus CoNLL05 et CoNLL2012.

1. https://verbs.colorado.edu/propbank/framesets-english-aliases/bake.html

Ces deux corpus sont toujours utilisés pour entraîner des modèles d'apprentissage neuronal qui ont la tâche d'annoter du texte en SRL. Leur guide sera réutilisé dans le cadre de la constitution de notre corpus CTeTex SRL.

### 2.6.2 Une vue d'ensemble des corpus en SRL

Comme précédemment mentionné, le projet Proposition Bank a pour but de construire un corpus annoté en SRL. Cette section se concentre sur une description des corpus CoNLL05 et CoNLL2012 qui proviennent du projet de Proposition Bank. Ces deux corpus sont manuellement annotés et ils peuvent être utilisés comme données d'entraînement, de développement et de référence pour de la SRL. Nous verrons que le genre linguistique de ces trois corpus ne s'apparente pas aux documents techniques comme les requis logiciels.

#### 2.6.2.1 CoNLL05

Le CoNLL05 est un corpus annoté en sémantique de rôle. Il est intégralement en anglais. Les genres de documents annotés proviennent principalement de journaux et aussi de manuels d'ordinateur d'IBM, de notes d'infirmiers et d'infirmières et de paroles orales spontanées transcrites. Le corpus contient un total d'environ 40 000 phrases pour les données d'entraînements.

Les auteurs et autrices du corpus ont procédé à une mesure de l'accord inter-annotateurs (IAA) pour juger la qualité des annotations. Toutefois, il n'y a pas mention du nombre de documents annotés qui ont passé l'accord, il n'y a pas non plus le moment auquel cet accord a été effectué. Si, par exemple, l'accord inter-annotateurs a été effectué à la suite d'une faible quantité d'annotations simples au début du projet, ce résultat ne permettra pas de juger correctement de l'accord sur l'ensemble des annotations finales. Malgré ce manque d'information, Proposition Bank montre

un résultat qui s'approche du taux parfait d'accord avec 91% d'accord en incluant l'ensemble des rôles des arguments modificateur (ARGM-MOD, ARGM-LOC, ARGM-MNR,...). La méthode utilisée est la métrique de Kappa de Cohen (Siegel and Castellan, 1988). Cette mesure calcule la probabilité d'accord entre annotateurs *P(A)* en tenant compte de la probabilité que les annotations soient faites au hasard *P(E)* (en désaccord).

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.1}$$

TABLEAU 2.4 – Vue d'ensemble des données de CoNLL05

| | Entraînement | Développement | refWSJ | refBrown |
|---|---|---|---|---|
| Phrases | 39,832 | 1,346 | 2,416 | 426 |
| Tokens | 950,028 | 32,853 | 56,684 | 7,159 |
| Propositions | 90,750 | 3,248 | 5,267 | 804 |
| Verbes (fréquence) | 3,101 | 860 | 982 | 351 |
| Arguments | 239,858 | 8,346 | 14,077 | 2,177 |
| A0 | 61,440 | 2,081 | 3,563 | 566 |
| A1 | 84,917 | 2,994 | 4,927 | 676 |
| A2 | 19,926 | 673 | 1,110 | 147 |
| A3 | 3,389 | 114 | 173 | 12 |
| A4 | 2,703 | 65 | 102 | 15 |
| A5 | 68 | 2 | 5 | 0 |
| AA | 14 | 1 | 0 | 0 |
| AM | 7 | 0 | 0 | 0 |
| AM-ADV | 8,210 | 279 | 506 | 143 |
| AM-CAU | 1,208 | 45 | 73 | 8 |
| AM-DIR | 1,144 | 36 | 85 | 53 |
| AM-DIS | 4,890 | 202 | 320 | 22 |
| AM-EXT | 628 | 28 | 32 | 5 |
| AM-LOC | 5,907 | 194 | 363 | 85 |
| AM-MNR | 6,358 | 242 | 344 | 110 |
| AM-MOD | 9,181 | 317 | 551 | 91 |
| AM-NEG | 3,225 | 104 | 230 | 50 |
| AM-PNC | 2,289 | 81 | 115 | 17 |
| AM-PRD | 66 | 3 | 5 | 1 |
| AM-REC | 14 | 0 | 2 | 0 |
| AM-TMP | 16,346 | 601 | 1,087 | 112 |
| R-A0 | 4,112 | 146 | 224 | 25 |
| R-A1 | 2,349 | 83 | 156 | 21 |
| R-A2 | 291 | 5 | 16 | 0 |
| R-A3 | 28 | 0 | 1 | 0 |
| R-A4 | 7 | 0 | 1 | 0 |
| R-AA | 2 | 0 | 0 | 0 |
| R-AM-ADV | 5 | 0 | 2 | 0 |
| R-AM-CAU | 41 | 3 | 4 | 2 |
| R-AM-DIR | 1 | 0 | 0 | 0 |
| R-AM-EXT | 4 | 1 | 1 | 0 |
| R-AM-LOC | 214 | 9 | 21 | 4 |
| R-AM-MNR | 143 | 6 | 6 | 2 |
| R-AM-PNC | 12 | 0 | 0 | 0 |
| R-AM-TMP | 719 | 31 | 52 | 10 |

```
WORDS---->  NE--->  POS   PARTIAL_SYNT   FULL_SYNT------>   VS   TARGETS  PROPS------->

The             *   DT    (NP*   (S*        (S(NP*          -    -        (A0*    (A0*
$               *   $        *     *     (ADJP(QP*          -    -           *       *
1.4             *   CD       *     *             *          -    -           *       *
billion         *   CD       *     *             *))        -    -           *       *
robot           *   NN       *     *             *          -    -           *       *
spacecraft      *   NN       *)    *             *)         -    -           *)      *)
faces           *   VBZ   (VP*)    *          (VP*          01   face      (V*)     *
a               *   DT    (NP*     *          (NP*          -    -        (A1*       *
six-year        *   JJ       *     *             *          -    -           *       *
journey         *   NN       *)    *             *          -    -           *       *
to              *   TO    (VP*   (S*        (S(VP*          -    -           *       *
explore         *   VB       *)    *          (VP*          01   explore     *     (V*)
Jupiter     (ORG*)  NNP   (NP*)    *       (NP(NP*)         -    -           *    (A1*
and             *   CC       *     *             *          -    -           *       *
its             *   PRP$  (NP*     *          (NP*          -    -           *       *
16              *   CD       *     *             *          -    -           *       *
known           *   JJ       *     *             *          -    -           *       *
moons           *   NNS      *)    *)            *)))))))   -    -           *)      *)
.               *   .        *     *)            *)         -    -           *       *
```

FIGURE 2.3 – Phrase annotée en SRL du corpus CoNLL05 en format *CoNLL* présentée comme exemple pour le projet de tâche commune (Carreras and Màrquez, 2005).

Le Tableau 2.4 et la Figure 2.3 montrent une vue d'ensemble du premier corpus de PropBank. CoNLL05 est composé de quatre sous corpus; des données d'entraînement, de développement et de tests (WSJ et Brown). Le format du corpus est en *CoNLL*, on constate que la phrase est segmentée avec chaque token par ligne. Ensuite, la colonne 2 montre les entités nommées et les colonnes 3 à 6 décrivent des informations syntaxiques (catégorie grammaticale et type de syntagme). Les dernières colonnes regroupent les informations concernant les propositions, c'est-à-dire, le *roleset ID*, le lexème du verbe et ses arguments. Chaque structure argumentale du verbe est contenue sur une colonne. Donc, pour chaque verbe de la phrase, il y a une colonne qui montre sa structure argumentale.

#### 2.6.2.2 CoNLL2012

Le corpus CoNLL2012 est l'un des derniers corpus produits dans le cadre du projet Proposition Bank. Il est à ce jour encore utilisé pour de la sémantique de rôle (Zhang et al., 2021). Ce corpus fait aussi partie du corpus Ontonotes version 5.0 (Weischedel et al., 2013). Ce dernier est annoté manuellement avec plusieurs informations linguistiques dont la structure argumentale en SRL ainsi que des informations ontologiques et de coréférence.

Les langues du corpus sont l'anglais, le chinois et l'arabe standard moderne. Il contient plusieurs genres textuels comme des articles de journaux, des téléjournaux transcrits, des émissions télévisées transcrites, des blogues et des conversations téléphoniques transcrites. Il contient 1,5 million de mots-formes de l'anglais, 800 000 mots-formes du chinois ainsi que 300 000 mots-formes de l'arabe standard moderne.

TABLEAU 2.5 – Vue d'ensemble du corpus CoNLL2012

| Argument-(rôle) | Nbr. de token |
|---|---|
| ARG0 | 404106 |
| ARG1 | 1320184 |
| ARG2 | 442815 |
| ARG3 | 26396 |
| ARG4 | 20077 |
| ARG5 | 191 |
| ARGA | 57 |
| ARGM-ADJ | 7910 |
| ARGM-ADV | 140076 |
| ARGM-CAU | 42542 |
| ARGM-COM | 1879 |
| ARGM-DIR | 11679 |
| ARGM-DIS | 30094 |
| ARGM-DSP | 66 |
| ARGM-EXT | 4482 |
| ARGM-GOL | 5277 |
| ARGM-LOC | 75770 |
| ARGM-LVB | 555 |
| ARGM-MNR | 61356 |
| ARGM-MOD | 25684 |
| ARGM-NEG | 13465 |
| ARGM-PNC | 7043 |
| ARGM-PRD | 25826 |
| ARGM-PRP | 40571 |
| ARGM-PRR | 6 |
| ARGM-PRX | 5 |
| ARGM-REC | 145 |
| ARGM-TMP | 163405 |
| R-ARG0 | 12300 |
| R-ARG1 | 7406 |
| R-ARG2 | 568 |
| R-ARG3 | 69 |
| R-ARG4 | 14 |
| R-ARG5 | 1 |
| R-ARGM-ADV | 28 |
| R-ARGM-CAU | 43 |
| R-ARGM-COM | 5 |
| R-ARGM-DIR | 17 |
| R-ARGM-EXT | 12 |
| R-ARGM-GOL | 8 |
| R-ARGM-LOC | 981 |
| R-ARGM-MNR | 132 |
| R-ARGM-MOD | 1 |
| R-ARGM-PNC | 6 |
| R-ARGM-PRD | 2 |
| R-ARGM-PRP | 9 |
| R-ARGM-TMP | 610 |
| C-ARG0 | 2085 |
| C-ARG1 | 25544 |
| C-ARG2 | 1732 |
| C-ARG3 | 106 |
| C-ARG4 | 179 |
| C-ARGM-ADJ | 2 |
| C-ARGM-ADV | 112 |
| C-ARGM-CAU | 17 |
| C-ARGM-COM | 4 |
| C-ARGM-DIR | 3 |
| C-ARGM-DIS | 3 |
| C-ARGM-DSP | 21 |
| C-ARGM-EXT | 183 |
| C-ARGM-LOC | 61 |
| C-ARGM-MNR | 263 |
| C-ARGM-MOD | 2 |
| C-ARGM-NEG | 6 |
| C-ARGM-PRP | 4 |
| C-ARGM-TMP | 129 |

Le Tableau 2.5 montre une vue d'ensemble du corpus CoNLL2012. La colonne de gauche montre les arguments ainsi que leur rôle avec leur nombre de tokens étiquetés dans la colonne de droite. On constate de nouvelles étiquettes depuis le corpus précédent CoNLL05. Ce sont huit nouveaux rôles attribués aux ARGM : PRP, GOL, COM, ADJ, PRR, PRX, DSP, LVB. Dans le Chapitre 4, ces différents rôles seront décrits. Cette section propose plutôt une vue d'ensemble du corpus pour exposer le type linguistique des données écrites ainsi que leur distribution. Par ailleurs, on constate également une nouvelle étiquette qui s'attache à n'importe quel argument : C-. Cette étiquette est utilisée pour les arguments disjoints, c'est-à-dire que certains éléments de la phrase sont placés à l'intérieur de l'argument, comme on peut le constater dans l'exemple (18). Ces éléments disjoints représentent environ 1% du corpus. Même si celle-ci est présente dans le corpus CoNLL2012, aucune mention de notation n'est présente dans la dernière version du guide d'annotation de Proposition Bank (Bonial et al., 2015).

(18) [**A1** This place] said [**A0** Chirstopher] [**C-A1** is wonderful].

```
#begin document (nw/wsj/07/wsj_0771); part 000
...
...
nw/wsj/07/wsj_0771 0   0          ``   ``   (TOP(S(S*           -    -  -   -         *        *      (ARG1*          *          *      -
nw/wsj/07/wsj_0771 0   1 Vandenberg  NNP        (NP*           -    -  -   -  (PERSON)   (ARG1*          *          *  (8|(0)
nw/wsj/07/wsj_0771 0   2         and   CC          *           -    -  -   -         *        *          *          *      -
nw/wsj/07/wsj_0771 0   3     Rayburn  NNP          *)          -    -  -   -  (PERSON)        *)         *          *  *(23)|8)
nw/wsj/07/wsj_0771 0   4         are  VBP       (VP*        be   01  1   -         *     (V*)          *          *      -
nw/wsj/07/wsj_0771 0   5      heroes  NNS    (NP(NP*)          -    -  -   -         *  (ARG2*          *          *      -
nw/wsj/07/wsj_0771 0   6          of   IN       (PP*           -    -  -   -         *        *          *          *      -
nw/wsj/07/wsj_0771 0   7        mine   NN    (NP*))))          -    -  5   -         *        *)         *          *    (15)
nw/wsj/07/wsj_0771 0   8           ,    ,          *           -    -  -   -         *        *          *          *      -
nw/wsj/07/wsj_0771 0   9           ,    ,          *)          -    -  -   -         *        *          *)         *      -
nw/wsj/07/wsj_0771 0  10         Mr.  NNP       (NP*           -    -  -   -         *        *      (ARG0*     (ARG0*         *     (15
nw/wsj/07/wsj_0771 0  11       Boren  NNP          *)          -    -  -   -  (PERSON)        *          *)         *)        *     15)
nw/wsj/07/wsj_0771 0  12        says  VBZ       (VP*        say   01  1   -         *        *        (V*)         *          *      -
nw/wsj/07/wsj_0771 0  13           ,    ,          *           -    -  -   -         *        *          *          *          *      -
nw/wsj/07/wsj_0771 0  14  referring  VBG     (S(VP*       refer   01  2   -         *        *  (ARGM-ADV*       (V*)         *      -
nw/wsj/07/wsj_0771 0  15          as   RB     (ADVP*          -    -  -   -         *        *          *  (ARGM-DIS*         *      -
nw/wsj/07/wsj_0771 0  16        well   RB          *)          -    -  -   -         *        *          *          *)         *      -
nw/wsj/07/wsj_0771 0  17          to   IN       (PP*          -    -  -   -         *        *          *      (ARG1*         *      -
nw/wsj/07/wsj_0771 0  18         Sam  NNP    (NP(NP*          -    -  -   -  (PERSON*        *          *          *         *     (23
nw/wsj/07/wsj_0771 0  19     Rayburn  NNP          *)          -    -  -   -         *)       *          *          *         *      -
nw/wsj/07/wsj_0771 0  20           ,    ,          *           -    -  -   -         *        *          *          *         *      -
nw/wsj/07/wsj_0771 0  21         the   DT    (NP(NP*          -    -  -   -         *        *          *          *     (ARG0*      -
nw/wsj/07/wsj_0771 0  22  Democratic   JJ          *           -    -  -   -    (NORP)        *          *          *         *      -
nw/wsj/07/wsj_0771 0  23       House  NNP          *           -    -  -   -     (ORG)        *          *          *         *      -
nw/wsj/07/wsj_0771 0  24     speaker   NN          *)          -    -  -   -         *        *          *          *         *)     -
nw/wsj/07/wsj_0771 0  25         who   WP  (SBAR(WHNP*)         -    -  -   -         *        *          *          *  (R-ARG0*)      -
nw/wsj/07/wsj_0771 0  26  cooperated  VBD     (S(VP*  cooperate   01  1   -         *        *          *          *        (V*)      -
nw/wsj/07/wsj_0771 0  27        with   IN       (PP*          -    -  -   -         *        *          *          *     (ARG1*      -
nw/wsj/07/wsj_0771 0  28   President  NNP       (NP*          -    -  -   -         *        *          *          *         *      -
nw/wsj/07/wsj_0771 0  29  Eisenhower  NNP *))))))))))         -    -  -   -  (PERSON)        *          *)         *)        *)    23)
nw/wsj/07/wsj_0771 0  30           .    .          *))         -    -  -   -         *        *          *          *         *     -

nw/wsj/07/wsj_0771 0   0          ``   ``    (TOP(S*          -    -  -   -         *        *          *          -
nw/wsj/07/wsj_0771 0   1        They  PRP      (NP*)          -    -  -   -         * (ARG0*)          *        (8)
nw/wsj/07/wsj_0771 0   2     allowed  VBD       (VP*       allow   01  1   -         *     (V*)          *          -
nw/wsj/07/wsj_0771 0   3        this   DT     (S(NP*          -    -  -   -         *  (ARG1*     (ARG1*         (6
nw/wsj/07/wsj_0771 0   4     country   NN          *)          -    -  3   -         *        *          *)         6)
nw/wsj/07/wsj_0771 0   5          to   TO       (VP*          -    -  -   -         *        *          *          -
nw/wsj/07/wsj_0771 0   6          be   VB       (VP*         be   01  1   -         *        *        (V*)       (16)
nw/wsj/07/wsj_0771 0   7    credible   JJ  (ADJP*)))))        -    -  -   -         *        *)     (ARG2*)        -
nw/wsj/07/wsj_0771 0   8           .    .          *))         -    -  -   -         *        *          *          -
#end document
```

FIGURE 2.4 – Phrases annotées en SRL du corpus CoNLL2012 formatées en *CoNLL* (Weischedel et al., 2013).

La Figure 2.4 montre le format des annotations du corpus CoNLL2012. Il est semblable au format de CoNLL05, mais on constate certaines différences. La première colonne montre le chemin du document et les deux colonnes suivantes exposent le numéro de partie du document et l'identifiant du token dans la phrase. On retrouve enfin la structure argumentale de chaque verbe dès la douzième colonne. Comme le CoNLL05, il y a une colonne exposant la structure argumentale pour chaque verbe contenu dans la phrase. La dernière colonne n'est pas pertinente dans le cadre de ce travail, mais elle expose les coréférences dans la phrase.

Tout compte fait, en passant par le signe linguistique de Saussure, la sémantique de cadres de Fillmore, différentes ressources verbales ainsi que des corpus en SRL, nous avons exposé la sémantique de rôle. Cette dernière a longtemps été étudiée entre autres par les auteurs et autrices du projet de Proposition Bank. De nouveaux corpus annotés en SRL avec la convention de Proposition Bank sont toujours active-

ment proposés. En effet, de nouvelles langues [1] et de nouveaux genres textuels [2] [3] sont proposés. Pour rassembler ces nouveaux corpus, un nouveau projet est proposé : *Universal PropBank* (Jindal et al., 2022). Le but de cette présente recherche est également de proposer l'application de la convention d'annotation de Proposition Bank sur des données d'un type rarement abordé, soit les requis logiciels. Les prochains chapitres fournissent une description de notre corpus CTeTex SRL ainsi que la méthodologie utilisée.

1. https://universalpropositions.github.io/#languages
2. https://developer.ibm.com/exchanges/data/all/contracts-proposition-bank/
3. https://developer.ibm.com/exchanges/data/all/finance-proposition-bank/

# Chapitre 3

# Vue d'ensemble des corpus de requis logiciels

Les données tests ou les données de références sont cruciales pour évaluer les performances de modèles entraînés. Elles permettent de tester les prédictions du modèle en les comparant à des données de référence qui sont généralement annotées manuellement. Aucun corpus de requis logiciel n'étant disponible dans le projet IVVES, seul un corpus libre de droits pouvait être utilisé comme source de données pour l'annotation. Cette partie propose de mettre en lumière une description du corpus PURE de Ferrari et al. (2017) qui a été utilisé pour construire le corpus qui regroupe 196 requis logiciels : CTeTex. Une section montrera une vue d'ensemble de ce dernier pour comprendre ses limites. Cette partie propose également d'exposer les caractéristiques linguistiques d'un tel corpus. Ces éléments apporteront un défi lors de l'annotation puisqu'une grande partie n' est pas traitée par le cadre d'annotation de Proposition Bank.

Nous avons premièrement décidé de chercher un corpus de requis logiciel annoté en SRL, mais ce fut sans succès. Il existe peu de corpus libres d'utilisation qui

contient des requis logiciels et aucun de ceux-ci n'est annoté en SRL. En effet, la plupart des industries qui développent des logiciels ne veulent pas exposer d'information publiquement à leur sujet. Cet aspect est compréhensible pour des raisons de compétitivité, de sécurité et de propriété intellectuelle. Malgré la rareté de ce type de données, le corpus PURE (Ferrari et al., 2017) a été choisi. Il sera utilisé dans le cadre de ce mémoire.

Un travail manuel a été effectué dans le but de cibler les RL dans l'ensemble des documents. Cette extraction est importante puisqu'un texte qui documente un système ne présente pas seulement des requis, mais d'autres éléments comme des titres, des sous-titres et des explications génériques. À la suite de cette extraction et de son annotation en syntaxe de dépendance, l'équipe du CRIM a nommé ce nouveau corpus CTeTex (Hassert et al., 2021). Il sera ensuite appelé CTeTex SRL après notre projet d'annotation en sémantique de rôle.

## 3.1 Corpus Pure

Le corpus PUblic REquirements (PURE) a été créé dans le but d'être librement utilisé dans la recherche en traitement automatique du langage. En 2017, une équipe composée de trois chercheurs et chercheuses à l'*Istituto di Scienza e tecnologie dell' Informazione* proposent le corpus PURE. Il est composé de 79 documents en anglais qui sont publiquement accessibles, c'est-à-dire que l'ensemble de ces documents étaient publiquement accessibles sur l'Internet lors de la création du corpus (Ferrari et al., 2017).

Les documents ont été trouvés en inscrivant les mots-clefs suivants sur un moteur de recherche ; *Requirements Documents, Requirements Specification, System Spe-*

*cification, Software Specification & SRS*. Ces documents ont des formats variés : 62 PDFs, 3 pages html, 13 DOCs et 1 format de textes enrichis (RTF). Ces documents sont disponibles dans leur format original et en *xml* sur le site du groupe de recherche [1]. Les auteurs décrivent les types de documents comme des normes de produits industriels, des documents d'institution ou de compagnie publique (armée, sécurité publique & recherche) ainsi que des projets universitaires.

Portons une attention particulière aux documents du corpus PURE. La plupart des documents sont en format PDF avec un total de 62 documents sur les 79. Ces documents PDFs comportent des éléments qui ne sont pas des requis logiciels comme les titres et les sous-titres, les images, la table des matières, la numérotation des pages et les explications générales. Pour distinguer un requis logiciel d'une autre phrase décrivant simplement le système envisagé, il faut des connaissances préalables. Un expert en ingénierie logiciel aura généralement ces connaissances pour effectuer cette extraction. Conséquemment, l'extraction textuelle (automatique) de l'ensemble du contenu des documents pourrait, dans une certaine mesure, être possible, mais elle est non pertinente si le but est d'extraire seulement des requis logiciels pour notre constitution du corpus. La page suivante montre un exemple d'une page de document du corpus PURE, elle nous aidera à exposer ce défi d'extraction des requis logiciels.

1. http ://nlreqdataset.isti.cnr.it/

| PAYLOAD TYPE | UAV |
|---|---|
| EO/IR | Predator |
| SAR | Predator |
| EO/IR | Outrider |
| | Future |

### *3.1 Required States And Modes*

The states of operation of the TCS shall include Startup, Operation, and Shutdown. [SSS014]

The TCS states shall not exist concurrently. [SSS015] Figure 3.1-1 shows the existing states of the TCS.

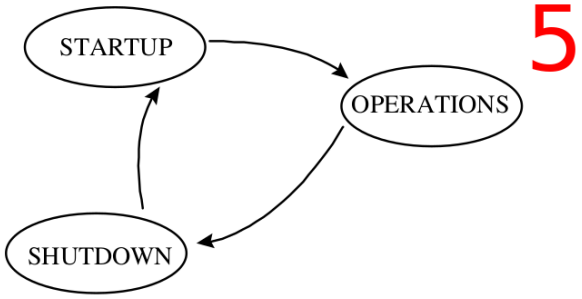



**Figure 3.1-1 TCS State Diagram**

#### 3.1.1 Startup State

Upon application of power the TCS shall enter the Startup State. [SSS016]

The Startup State shall be comprised of the following modes: Normal Startup Mode and Recovery Startup Mode. [SSS017]

Figure 3.1.1-1 shows the modes that exist in the Startup State.



FIGURE 3.1 – Exemple de la page huit du PDF 1999 - tcs du corpus PURE

La Figure 3.1 présentée à la page précédente a été découpée en éléments distincts. Ils ne sont pas intégralement des requis logiciels. Le premier élément de cette page est un tableau qui montre le type de *Payload* ou de charge utile ainsi que son aéronef qui lui est associé. Ce tableau pourrait être considéré comme un requis système de haut niveau sur l'interopérabilité en se référant aux énonciations à la page précédente du document original. Toutefois, dans le cadre de la recherche, un requis doit être exprimé en forme textuelle. Ce tableau ne sera donc pas considéré comme un requis logiciel pour notre projet de constitution du corpus de référence. Le deuxième élément est un sous-titre qui a pour but de guider le ou la lectrice au commencement d'une nouvelle section qui expliquera les états et les modes requis. Il n'est donc pas considéré comme un requis logiciel. L'élément trois peut être considéré comme un RL puisqu'il est une phrase comportant un verbe ou une action, un agent et un patient. De plus, cette phrase expose un requis non fonctionnel puisqu'il est question des états de l'opération qui devrait inclure certains modes (Startup, Operation et Shutdown). Cette phrase montre une attente de qualité et non de comportement que doit avoir le TCS. Les éléments (4) et (5) sont des requis fonctionnels. L'élément (4) exprime une exclusivité des états et l'élément (5) illustre les transitions valides entre les états. Toutefois, ces deux éléments n'ont pas été considérés comme des requis dans le cadre du corpus puisqu'il relève trop du contexte pour être interprété. Ensuite, comme l'élément deux, le sous-titre '3.1.1 Startup State' ne l'est pas non plus. Maintenant, les éléments 7 et 8 sont des requis logiciels fonctionnels :

— Upon application of power the TCS shall enter the Startup State. [SSS016]
— The Startup State shall be comprised of the following modes : Normal Startup Mode and Recovery Startup Mode. [SSS017]

Finalement, la phrase (élément 9) qui explique l'image à la page suivante et la numérotation de la page '8' ne sont pas des requis logiciels. Bref, on peut constater que sur les dix éléments de la page de ce document, seulement trois de ceux-ci sont des requis logiciels. Il est important de mentionner que, même si certains sont considérés comme des requis logiciels au sens large (1, 4, 5), dans le cadre de la constitution du corpus, notre définition d'un requis logiciel a été adaptée à notre objectif. L'objectif est de constituer un corpus de référence de requis logiciels qui peuvent être utilisés pour générer des jeux de tests. Il faut alors qu'un requis ait une forme textuelle et il doit constituer assez d'information pour générer un jeu de tests. Comme nous l'avons exposé, certains requis ne contiennent pas assez d'informations et relèvent trop d'un niveau de contexte externe pour être inclus dans le cadre du corpus CTeTex SRL. C'est pour cette raison qu'un travail de filtration des documents doit être effectué sur les documents pour extraire l'ensemble des RL disponibles au sein du corpus PURE.

## 3.2 Corpus CTeTex SRL

Le corpus CTeTex SRL est le produit d'un travail de filtration du corpus PURE. Un expert en ingénierie logiciel a effectué cette tâche. Pour qu'un requis logiciel soit accepté dans le corpus, il devait respecter deux conditions. La première est qu'il devait être exprimé sous forme de phrase. La deuxième condition était que le requis contienne, dans une certaine mesure, assez d'informations pour générer un jeu de test. Enfin, cette section propose une vue d'ensemble du corpus ainsi que des comparaisons entre les annotations du corpus CTeTex SRL et d'une plateforme d'annotation à la pointe de la littérature.

Une version annotée en Universal Dependencies sera rendue disponible dans le répertoire de *Universal Dependencies* sous le nom de CTeTex et en format *conllu*. En effet, avant le commencement du présent stage de recherche, un travail d'annotation du corpus a été fait en analyse syntaxique en dépendance selon le standard d'*Universal Dependencies* (UD). Comme précédemment mentionné, à la suite du notre projet d'annotation des RL en sémantique de rôle, le corpus est appelé CTeTex SRL.

CTeTex SRL est composé de 196 RL disponibles en fichiers textes individuels (.txt). Ces derniers sont composés de l'intégralité du requis tel qu'il est présenté dans un document. L'ensemble des acronymes, des listes de points, de tirets ou de sauts à la ligne ainsi que les éléments entre parenthèses et crochets sont conservés. Ce choix a pour but d'éviter tout ajout de manipulations de prétraitements et de tenir compte de ces éléments présents dans un tel genre de corpus. Nous pourrons alors juger plus adéquatement de la généralisation ou non d'une convention d'annotation en SRL.

TABLEAU 3.1 – Vue d'ensemble des corpus de références en comparaison avec CTeTex SRL

| **Corpus test** | **Phrases** | **Tokens** | **Moy. de tokens par phrase** | **Nbr. de verbes** | **Moy. de verbes par phrase** |
|---|---|---|---|---|---|
| CoNLL05 (WSJ) | 2 414 | 56 684 | 23 | 5267 | 2.18 |
| Brown | 426 | 7159 | 17 | 804 | 1.89 |
| CoNLL2012 | 9479 | 170 000 (Mots) | 18 (Mots) | - | - |
| **CTeTex SRL (Le Nôtre)** | **276** | **9273** | **33,60** | **730** | **2.64** |

Le Tableau 3.1 montre une vue d'ensemble du corpus utilisé dans le cadre de ce mémoire. Les informations des corpus de CoNLL05, de Brown et de CoNLL2012 ont été tirées de l'article de CoNLL05 & CoNLL2012 shared task (Carreras and Màrquez, 2005; Weischedel et al., 2013). Certaines des informations du corpus CTeTex SRL pro-

viennent de l'article Hassert et al. (2021). Les deux dernières colonnes exposent le nombre de verbes ainsi que la moyenne de verbe qu'on pourrait retrouver dans une phrase. On constate que le corpus CTeTex SRL obtient un score de moyenne de verbe dans une phrase plus élevé que les corpus de référence WSJ et Brown. L'information du nombre de verbes dans le corpus 2012 n'est pas présente dans leur publication, il est donc impossible de le comparer avec CTeTex SRL. Toutefois, il s'apparente aux données de référence de CoNLL05 puisqu'il est son extension avec l'ajout de nouvelles phrases annotées. On peut donc conclure que la moyenne du nombre de verbes par phrase devrait être semblable au résultat de CoNLL05. Enfin, comme précédemment mentionné, chaque verbe a une structure argument et plus son nombre est élevé plus il y a aura d'arguments présents dans une phrase. Ainsi, le corpus de requis logiciels présente une plus grande complexité sémantique que les autres corpus par son nombre de verbes plus élevés.

### 3.2.1 Caractéristiques linguistiques de CTeTex SRL

Cette partie du mémoire montre des particularités linguistiques qui sont généralement plus fréquentes au sein d'un corpus de documents techniques. Ces caractéristiques sont uniquement tirées du corpus CTeTex SRL, mais pourront dans une certaine mesure, être présentes à l'intérieur d'autres documents techniques.

Comme on peut le constater par le précédent Tableau 3.1, une première particularité est qu'un corpus de requis logiciel comporte en moyenne presque le double de tokens par phrase que le corpus de référence Brown et huit tokens de plus que le *test set* de CoNLL05. La longueur corrèle généralement avec le nombre de verbes et le nombre d'énumérations d'éléments dans celle-ci. Si tel est le cas, le nombre de prédicats et d'arguments sera en moyenne plus grand que les autres corpus de référence (CoNLL05 & CoNLL2012).

Les requis logiciels comportent parfois des méthodes scientifiques et mathématiques, des abréviations et acronymes, des listes et énumérations et enfin un vocabulaire spécialisé (Hassert et al., 2021).

TABLEAU 3.2 – Caractéristiques linguistiques de CTeTex SRL

| **Phenomènes linguistiques** | **Nbr. de tokens** |
|---|---|
| Notations scientifiques & mathématiques | 68 |
| Abréviations & acronymes | 579 |
| Listes & énumérations | 2 736 |
| Vocables spécialisés | 1 180 |

La distribution de ces éléments présents à l'intérieur du corpus CTeTex SRL, qui est présenté par le Tableau 3.2, est tirée de l'article du corpus CTeTex annoté en syntaxe de dépendance (UD) (Hassert et al., 2021).

### 3.2.2 Comparaisons entre l'annotation manuelle et un outil d'annotation automatique récent

La suite de cette section met en lumière la comparaison entre une annotation manuelle et une autre venant de la plateforme d'AllenNLP (Gardner et al., 2017). Certains éléments peuvent apporter des défis à des systèmes d'annotation en SRL sur des requis logiciels. Les deux exemples de requis sont tirés du corpus de CTeTex SRL.

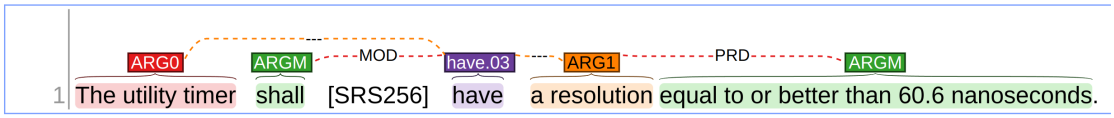



FIGURE 3.2 – Exemple d'annotation manuelle d'un requis logiciel

La **Figure 3.2** montre un exemple d'annotation en SRL d'un requis logiciel com-

portant un vocabulaire spécialisé qui se rapproche d'une représentation mathématique avec le symbole $\leqslant$. Ce requis exprime qu'un minuteur doit avoir une résolution plus petite ou égale à 60.6 nanosecondes. Le verbe de la phrase est *have* qui exprime, dans une certaine mesure, le fait de posséder ou d'avoir quelque chose. Il faut donc faire référence au *roleset ID have.03* de la ressource verbale de Proposition Bank. Les deux arguments qui sont permis par ce *roleset ID* sont un ARG0 qui joue le rôle du *possesseur* et un ARG1 qui joue le rôle de *ce qui est possédé*. Ici, l'ARG0 est *The utility timer* et l'ARG1 est *a resolution*. Bien que la suite de l'ARG1 peut également jouer le rôle de ce qui est possédé puisqu'il clarifie la résolution, il doit être séparé du ARG1. La raison est que cette information est importante pour la génération de test, puisqu'elle est la condition critique auquel le minuteur utilitaire doit répondre. Dans le guide d'annotation qui sera décrit au chapitre suivant, un argument modificateur qui modifie un argument du verbe (ARG0-5), il porte le rôle de *Secondary Predication* (PRD) ou de seconds prédicats. Pour respecter le guide, cette étiquette a été utilisée (ARGM-PRD).

Portons également une attention sur la qualité de la résolution exprimée par l'adjectif *better*. Sans le fait de connaître son contexte d'utilisation ou même le fait qu'un minuteur utilitaire peut avoir une résolution exprimée par un temps *X*, cet adjectif est ambigu puisqu'il relève intégralement du contexte pour comprendre que *better* montre un *plus petit que*. En effet, lorsqu'un minuteur utilitaire à une résolution plus *base*, il est plus rapide et ainsi qualifié de *bon*. Enfin, le but de cette recherche n'est pas de désambiguïser les requis logiciels, mais il est tout de même intéressant de constater que la sémantique ou le sens des mots-formes ne relève pas seulement de la phrase, mais également du contexte d'énonciation.

Prenons la même phrase et testons-la avec l'outil d'annotation AllenNLP qui est publiquement accessible (Gardner et al., 2017). Il est important de mentionner que l'annotation obtenue provient de leur démo [1]. On constate premièrement l'annotation du *shall* comme un verbe, mais qu'il joue plutôt un rôle de modal comme bien étiqueté avec le verbe *have*. Ensuite, pour le verbe *have*, les deux premiers arguments sont correctement identifiés en les comparant à notre annotation manuelle 3.2. Toutefois, les éléments qui suivent le verbe sont annotés de façon différente. Elle diffère pour l'ARG1 et l'ARGM-PRD du verbe *have* qui ne sont pas séparés. Par ailleurs, l'annotation ne permet pas de distinguer le sens du verbe *have* qui porte le *roleset ID have.03*.

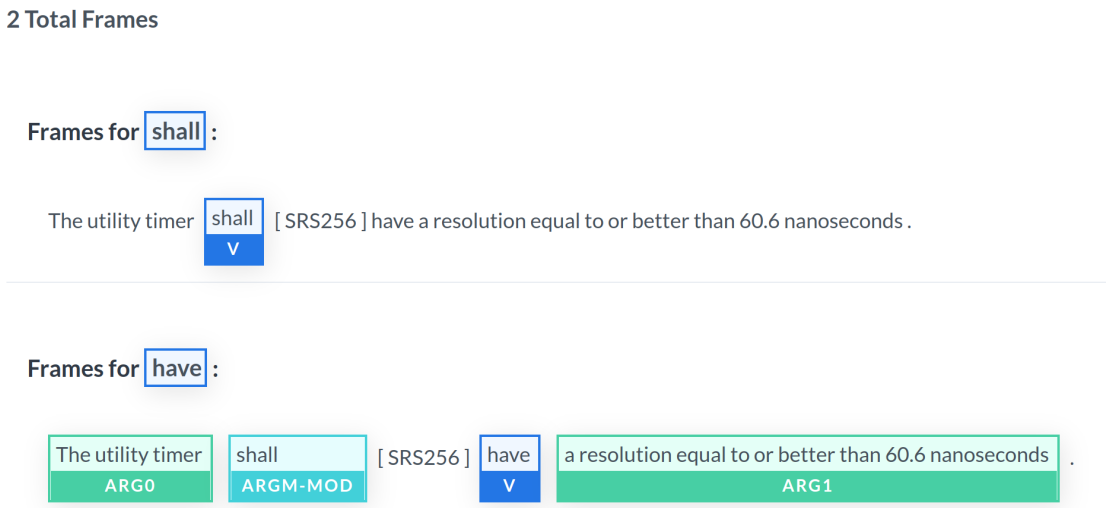



FIGURE 3.3 – Annotation effectuée avec la plateforme AllenNLP

Prenons ensuite un deuxième exemple qui provient du corpus CTeTex SRL. Cet exemple expose les caractéristiques des listes présentes dans un requis logiciel ainsi qu'un acronyme. Il est important de mentionner que les arguments du verbe *capable* ne sont pas identifiés pour focaliser l'attention sur le verbe *displaying*.

1. https://demo.allennlp.org/semantic-role-labeling

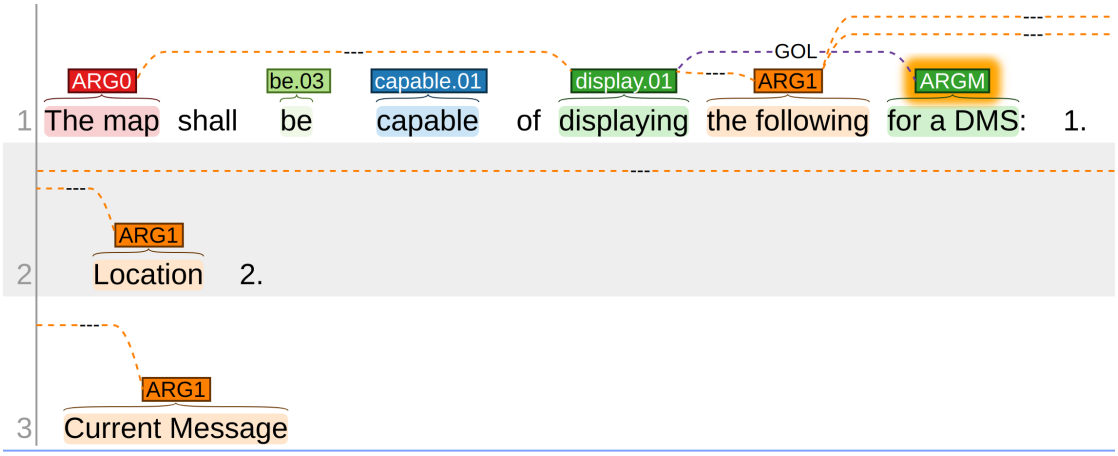



FIGURE 3.4 – Exemple d'annotation manuelle d'un requis logiciel

Ce requis logiciel exprime la capacité d'une carte à afficher selon un DMS (*Dynamic Message Sign* ou panneaux de signalisation routière) choisi, les informations sur la localisation et le message affiché en temps réel sur ce DMS.

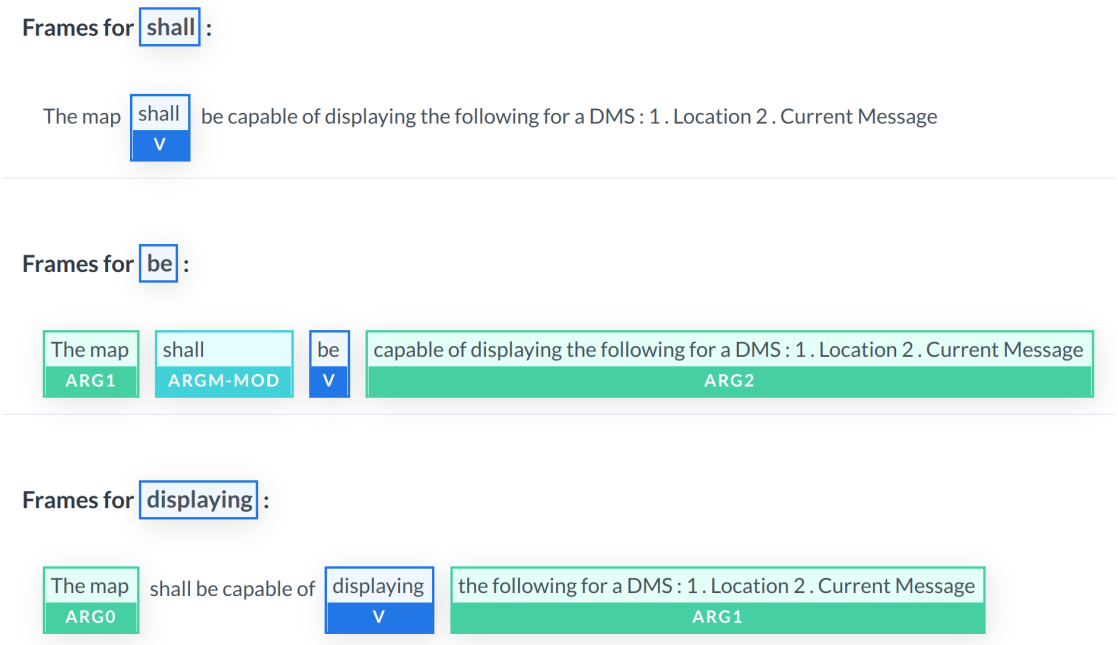



FIGURE 3.5 – Annotation effectuée avec la plateforme AllenNLP

Dans ce requis, nous pouvons constater quatre éléments qui sont différents entre notre annotation 3.4 et celle de la plateforme d'AllenNLP 3.5. La première est l'auxiliaire *be* qui est identifiée comme un verbe. Pourtant, la convention de Proposition Bank suggère de les cibler en les identifiant comme des auxiliaires et de ne pas annoter d'argument. Bien que la raison ne semble pas être exposée par le guide, un auxiliaire semble avoir moins de poids qu'un autre verbe, pour être porteur de l'action d'une phrase. Ensuite, comme dans l'exemple précédent le modal *shall* n'est pas un verbe et le verbe *capable* n'est pas identifié par la plateforme. Ici, le point le plus important est que l'ARG1 du verbe *displaying* présente des éléments qui sont disjoints. L'entité qui est affichée (ARG1) est correctement étiquetée et ciblée par la plateforme. Toutefois, il se retrouve un argument facultatif avant la liste qui dénote du but de l'action. Cet argument n'est pas régi par la structure argumentale du verbe *displaying*, il est donc considéré comme un ARGM-GOL. C'est pour un *DMS* (*Dynamic Message Sign* ou panneaux de signalisation routière) choisi que les informations de la localisation et du message sont affichées sur la carte. Il joue alors le rôle du but de l'action ou du *goal* comme le guide le suggère.

Ces deux exemples mettent en évidence certaines caractéristiques linguistiques des requis logiciels qui peuvent être source de défi pour de l'annotation en SRL. On constate des notations scientifiques, un acronyme, des listes ainsi qu'un vocabulaire spécialisé. Ces divergences n'ont pas comme but de critiquer la plateforme AllenNLP, mais plutôt d'exposer les défis qu'apportent des documents techniques lorsqu'un modèle neuronal tente de l'étiqueter en SRL.

En conclusion, ce chapitre décrit le corpus CTeTex SRL. Il expose le corpus PURE qui est composé de documents qui regroupent des requis logiciels. Un travail de filtration des documents a été entrepris par un expert en ingénierie logiciel en ciblant

les requis logiciels selon deux conditions. Il devait être sous forme de phrase et il devait être composé d'assez d'informations pour générer un jeu de tests. Ensuite, cette partie du mémoire montre les défis que peut apporter un tel corpus par sa moyenne de verbe par phrase plus élevée que d'autres corpus de référence et ses caractéristiques linguistiques. Enfin, une comparaison entre l'annotation manuelle et l'annotation automatique est exposée pour illustrer ces défis.

# Chapitre 4

# Méthodologie d'annotation

Cette partie du mémoire propose une méthodologie pour l'annotation du corpus de référence ainsi que le calcul de l'accord inter-annotateurs. L'annotation manuelle du corpus a été effectuée par l'application de la convention de Proposition Bank en utilisant la plateforme Inception (Version 23.4). Il sera question d'une description de cette convention, de la plateforme Inception ainsi que du processus d'adaptation de la convention pour traiter les requis logiciels en ayant pour but de générer des jeux de test. Enfin, un accord inter-annotateurs sera calculé pour évaluer les annotations.

Le processus d'annotation s'est déroulé en quatre phases. Avant de les débuter, un guide d'annotation a été élaboré en intégrant des informations du guide de Proposition Bank et des modifications apportées à ce dernier pour l'adapter au présent projet. Le projet d'annotation a été découpé en plusieurs phases pour permettre d'évaluer l'application du nouveau guide sur les requis logiciels. Ces phases ont permis d'apporter certaines clarifications au guide et de pointer, tôt dans le processus d'annotation, certaines limites de l'application du guide aux documents techniques. De plus, l'annotation de l'ensemble des requis logiciels a été effectuée par une équipe de deux personnes composée d'un chercheur en traitement automatique des langues

naturelles et ingénieur logiciel et d'une étudiante au master en traitement automatique des langues naturelles. Suivant chaque phase d'annotation, une validation des annotations réalisées par l'équipe d'annotateurs et annotatrices.

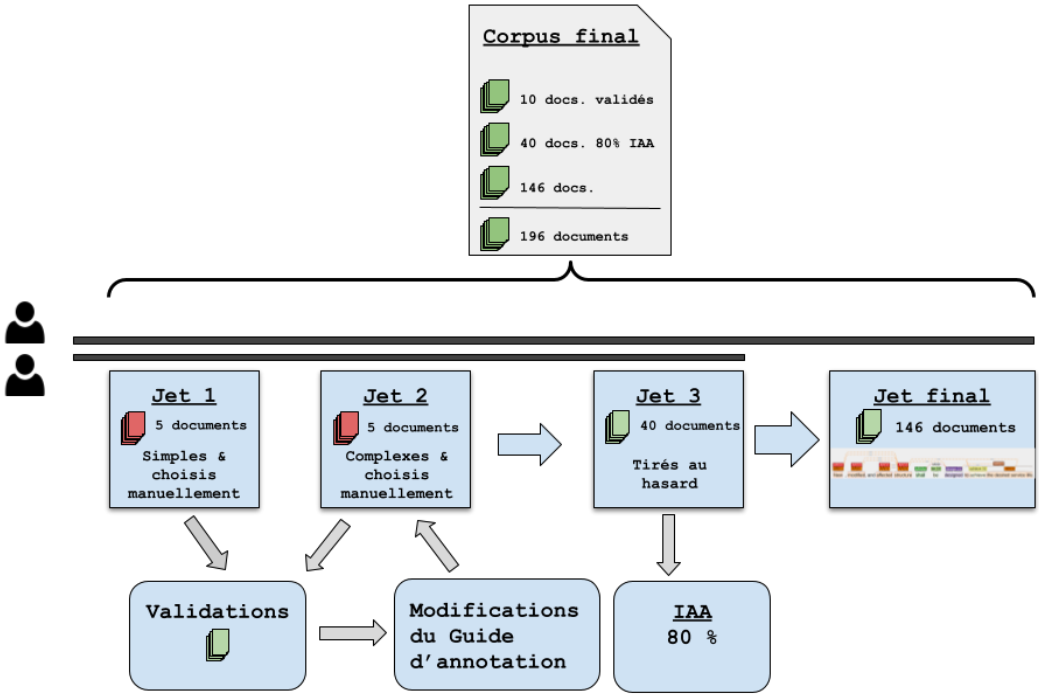



FIGURE 4.1 – Processus d'annotation du corpus CTeTex SRL

La première phase s'est composée de cinq documents du corpus choisi manuellement. Un document représente ainsi un requis logiciel. Ces documents avaient la condition d'être relativement simple à annoter puisque c'était la première tentative d'application du guide. La simplicité d'un requis logiciel était définie par son nombre de verbes et ses phénomènes linguistiques. La deuxième phase s'est composée de cinq autres requis choisis manuellement dans le corpus. Ces derniers présentaient une plus grande complexité que les cinq premiers de par leur longueur, leur nombre de verbes ou la présence de phénomènes linguistiques. Ensuite, la troisième phase de l'annotation s'est composée de 20% des 186 requis restants pour un total de 40

requis à annoter. Cette phase a été importante puisque nous avons évalué l'accord inter-annotateurs (IAA) sur celle-ci, une discussion sur cette évaluation sera proposée à la Section 4.3. Lorsque le taux d'accord était satisfaisant, le projet d'annotation pouvait être conclu par une seule personne de l'équipe d'annotation. Cette décision nous a permis de finaliser les 146 autres annotations plus rapidement. La Figure 4.1 montre les différentes étapes d'annotation qui sont composées de 4 jets.

En somme, le projet d'annotation s'est déroulé sur environ deux mois avec un total du travail de l'annotation et de la validation à plus de 100 heures. Le gabarit de projet d'annotation est disponible à ce lien [1].

## 4.1 La convention d'annotation PropBank

Le corpus CTeTex SRL contient un ensemble de RLs comme nous l'avons vu dans la section précédente. La convention d'annotation de Proposition Bank a été utilisée pour construire un corpus de référence annoté manuellement en sémantique de rôle. Cette convention est utilisée pour la majorité des corpus en SRL présent dans la littérature. Des ressources comme leur page *Github* [2] sont activement utilisées et maintenues par ses auteurs et autrices. La dernière version de la convention date de 2015. Cette partie se concentrera sur la présentation de celle-ci (Bonial et al., 2015).

La méthodologie d'annotation qui est exposée par le guide de PropBank se compose de deux grandes étapes. La première est de choisir d'annoter chaque prédicat d'une phrase avec son sens spécifique et la deuxième est d'annoter les arguments porteurs de rôles sémantiques. L'ensemble des annotations du projet de PropBank sont effectuées sur la plateforme Jubilee (Choi et al., 2009). Comme mentionné pré-

1. https://github.com/alicebret/SRLannotation
2. https://github.com/propbank

cédemment, les phrases présentées aux annotateurs et annotatrices depuis cette plateforme sont déjà préannotées avec les *parts of speech* ou catégories grammaticales de *Penn Treebank*. De plus, la majorité du corpus est annoté en arbre syntaxique (Marcus et al., 1993). Alors, l'ensemble des verbes ou prédicats sont déjà ciblés depuis cette préannotation, la suite est de choisir le sens correspondant au verbe au sein de la phrase. Ensuite la deuxième étape est de choisir les arguments selon le sens du verbe précédemment choisi. Il y a deux grandes classes possibles : ARG0-5 & ARGM. Au sein de la classe ARG0-5, il y a donc les possibilités : ARG0, ARG1, ARG2, ARG3, ARG4 & ARG5. Pour les ARGM, il y a : COM, LOC, DIR, GOL, MNR, TMP, EXT, REC, PRD, PRP, CAU, DIS, ADV, ADJ, MOD, NEG, DSP, LVB & CXN. Au total, il y a 25 classes d'arguments possibles dans une phrase.

Cette partie propose de synthétiser les informations contenues dans la convention d'annotation de PropBank qui, par la suite, sera utilisée et adaptée dans le cadre de l'annotation de requis logiciels. Pour de plus amples informations, veuillez directement vous référer à la version officielle du guide d'annotation [1].

### 4.1.1 Choisir le sens du prédicat

Comme déjà mentionné, l'annotateur ou l'annotatrice doit choisir le sens approprié d'un verbe déjà ciblé par Penn Treebank. Pour ce faire, la forme lemmatisée du verbe doit être associée à l'entrée proposée par la ressource verbale de PropBank [2]. La forme lemmatisée est le lexème d'un verbe ou sa forme verbale sans flexion. Par exemple, les mot-formes *plays, played, playing* ou *play* seront rapportés à son lexème *play*. Pour ce prédicat, disponible à cette page [3], l'équipe d'anno-

1. https://github.com/propbank/propbank-documentation/raw/master/annotation-guidelines/Propbank-Annotation-Guidelines.pdf
2. https://verbs.colorado.edu/propbank/framesets-english-aliases/
3. https://verbs.colorado.edu/propbank/framesets-english-aliases/play.html

tation choisira entre *play.01, play.02, play.08, play.10, play.11, play.12, play_out.03, play_up.04, play_down.07, play_off.05, play_to.06* ou *play_on.09* selon son contexte d'utilisation dans la phrase. La banque PropBank propose un ensemble de 10 685 *roleset IDs.*

TABLEAU 4.1 – Description du prédicat *PLAY* selon PropBank [1]

| Entrée (*RoleSet ID*) | Sens | Roles | Exemples |
|---|---|---|---|
| **play.01** | play a game | **Arg0-PAG** : player<br>**Arg1-PPT** : game<br>**Arg2-MNR** : instrument/equipment used to play game<br>**Arg3-COM** : opponent, play against whom | The backers play a fiscal game of their own.<br>**Arg0** : The backers<br>**Rel** : play<br>**Arg1** : a fiscal game of their own |
| **play.02** | play a role | **Arg0-PAG** : actor<br>**Arg1-PPT** : role | Abbie Hoffman, in this case, is played by Paul Lieber.<br>**Arg1** : Abbie Hoffman<br>**Argm-dis** : in this case<br>**Rel** : played<br>**Arg0** : by Paul Lieber |
| **play.08** | play into : be a factor | **Arg0-PAG** : thing factoring in, subject in active clauses<br>**Arg1-PPT** : thing being factored into | Sunspot variability possibly plays into climate variability.<br>**Arg0** : Sunspot variability<br>**Argm-adv** : possibly<br>**Rel** : plays<br>**Arg1** : into climate variability |
| **play.10** | play a trick on someone | **Arg0-PAG** : Trickster<br>**Arg1-PPT** : mention of trick<br>**Arg2-GOL** : tricked, who trick was played on | Playing clever little tricks on minor issues, but blinded<br>to major issues by trifles are not considered to be elites.<br>**Arg0** : PRO<br>**Rel** : playing<br>**Arg1** : clever little tricks<br>**Arg2** : on minor issues |
| **play.11** | play/perform music | **Arg0-PAG** : performer, player<br>**Arg1-PPT** : thing performed (song, etc)<br>**Arg2-MNR** : musical instrument/style | Javier played the Led Zeppelin record on his grandmother's turntable.<br>**Arg0** : Javier<br>**Rel** : played<br>**Arg1** : the Led Zeppelin record<br>**Arg2** : on his grandmother's turntable |
| **play.12** | play/perform music | **Arg0-PAG** : performer, player<br>**Arg1-PPT** : thing performed (song, play, etc)<br>**Arg2-MNR** : musical instrument/style | The service began with the playing of the national anthem.<br>**Rel** : playing<br>**Arg1** : of the national anthem |
| **play_out.03** | play out : come to completion | **Arg0-PAG** : agent, entity causing something to complete<br>**Arg1-PPT** : thing coming to completion | The battle played out by Tuesday evening.<br>**Arg1** : The battle<br>**Rel** : [played] [out]<br>**Argm-tmp** : by Tuesday evening |
| **play_up.04** | play up : emphasize,<br>make sound better | **Arg0-PAG** : emphasizer, agent<br>**Arg1-PPT** : topic of discussion | Boeing played up the downside.<br>**Arg0** : Boeing<br>**Rel** : [played] [up]<br>**Arg1** : the downside |
| **play_down.07** | play down : deemphasize,<br>make sound less important | **Arg0-PAG** : deemphasizer, agent<br>**Arg1-PPT** : topic of discussion | The Justice Department scrambled to play down the significance<br>of revised guidelines concerning prosecutions<br>under the federal racketeering law.<br>**ARG0** : The Justice Department<br>**Rel** : [play] [down]<br>**Arg1** : the significance of revised guidelines concerning prosecutions<br>under the federal racketeering law |
| **play_off.05** | play off : manipulate | **Arg0-PAG** : manipulator<br>**Arg1-PPT** : one victim<br>**Arg2-PPT** : the other victim | And "shippers are getting the feeling that they have played<br>one trucker off against another as much as they can, " he said .<br>**Arg0** : they<br>**Rel** : [off] [played]<br>**Arg1** : one trucker<br>**Arg2** : against another<br>**Argm-tmp** : as much as they can |
| **play_to.06** | play to : butter up to,<br>try to please | **Arg0-PAG** : causer of pleasing, agent<br>**Arg1-PPT** : entity being pleased | Mr. Ortega's remarks also played to the suspicions of some US officials.<br>**Arg0** : Mr. Ortega's remarks<br>**Argm-dis** : also<br>**Rel** : [played] [to]<br>**Arg1** : the suspicions of some US officials |
| **play_on.09** | play on : manipulate,<br>take advantage | **Arg0-PAG** : manipulator<br>**Arg1-PPT** : one victim<br>**Arg2-PPT** : the other victim | Hamas plays on the Palestinean Authority's failure to produce.<br>**Arg0** : Hamas<br>**Rel** : [played] [on]<br>**Arg1** : the Palestinean Authority's failure to produce. |

1. https ://verbs.colorado.edu/propbank/framesets-english-aliases/play.html

Pour choisir le bon *roleset ID*, la personne qui annote doit se référer à leur sens attribué par la ressource de PropBank. Elle peut aussi se référer à la ressource d'unification des ressources de *FrameNet, VerbNet & PropBank* [1]. Comme nous l'avons vu avec l'exemple du verbe *meet* 1.1 et *play* 4.1, plusieurs *roleset IDs* sont généralement exposés par PropBank. Pour choisir le bon, il faut se référer à l'ensemble des informations proposées par ces ressources.

Il se peut qu'un annotateur ou annotatrice rencontre un verbe ou un prédicat qui n'est pas proposé par la ressource de PropBank. En effet, les verbes proposés par cette ressource sont principalement ceux présents dans une portion annotée de leur corpus. Donc, la liste de verbes reflète leur utilisation selon leur type de corpus. On peut, par exemple, constater plusieurs sens qui sont généralement utilisés dans les domaines de la finance et de l'économie (*advance.03, bleed.02, charge.04*) ou de la politique et de la justice (*antiwar.01, legalize.01*). Ce sont habituellement des domaines présents dans des textes de journaux et de téléjournaux. Si un verbe ou un sens d'un verbe n'est pas proposé, il tient au jugement de chaque annotateur qui rencontre ce problème, de déterminer par lui ou elle-même, les arguments appropriés en se référant à d'autres verbes se rapprochant sémantiquement et syntaxiquement du verbe inconnu de la ressource. Nous verrons plus tard notre proposition lorsque nous rencontrons ce même problème lors de notre annotation du corpus de CTeTex SRL en sémantique de rôle. En effet, ce manque peut montrer certaines limitations dans l'utilisation de la ressource de PropBank pour les documents techniques.

Certaines formes verbales seront des expressions polylexicales. C'est le cas de plusieurs entrées comme *give_up* du prédicat *give, keep_up et keep_up* du prédicat *keep* ou encore *take_away, take_in, take_off, take_on, take_out, take_over, take_back,*

1. `https://verbs.colorado.edu/verb-index/index.php`

*take_down ou take_hold* du prédicat *take*. On retrouve beaucoup plus de ces EPL dans la ressource verbale que celles précédemment énumérées. On les retrouve également au sein du corpus CTeTex SRL. Ces EPL ont leur *roleset ID* et ainsi, chacun leurs arguments possibles dans une phrase. De plus, si une tokenisation est effectuée comme tâche de prétraitement d'un texte, il convient de segmenter correctement les EPL verbales pour les relier aux bonnes entrées de PropBank.

TABLEAU 4.2 – Comparaison du *roleset ID take_out.11* et *take.01* [1]

| take_out.11 | take.01 |
|---|---|
| take out : obtain, draw (out) | take, acquire, come to have,<br>choose, bring with you from somewhere,<br>internalize, ingest |
| **Arg0-PAG** : entity drawing something forth<br>**Arg1-PPT** : thing taken out | **Arg0-PAG** : Taker<br>**Arg1-PPT** : thing taken<br>**Arg2-DIR** : taken FROM, SOURCE of thing taken<br>**Arg3-GOL** : destination |

(19) A company once took out an ad.

a) *Roleset ID :* take_out.11
**Arg0** : A company
**Argm-tmp** : once
**Rel** : took out
**Arg1** : an ad

b) *Roleset ID :* take.01
**Arg0** : A company
**Argm-tmp** : once
**Rel** : took
**Argm-dir** : out
**Arg1** : an ad

1. https ://verbs.colorado.edu/propbank/framesets-english-aliases/take.html

Dans l'exemple 19, on constate une différence qui provient du choix du *roleset ID* pour déterminer la structure argumentale. Ce choix a alors une répercussion sur l'analyse sémantique. Ici, le sens approprié est le take_out.11 selon la ressource Proposition Bank puisqu'une compagnie (ARG0) retire (REL) une publicité (ARG1), cette phrase fait partie de leurs exemples pour ce *roleset ID*. Si le *take.01* avait été choisi, le mot-forme *out* pourrait être considéré comme un argument modificateur de direction puisque l'action de prendre suit une direction ou un mouvement. Ici, il peut même avoir une confusion justifiée avec l'ARG2-DIR du *take.01* puisque l'ARG2 a généralement le rôle de direction. On constate donc que la segmentation des EPL peut-être source de confusion auprès de la structure argumentale d'une phrase.

### 4.1.2 Cibler les arguments

Suivant l'étape du choix du *roleset ID*, il faut cibler les arguments proposés par le sens que porte le verbe dans la phrase présentée. Il y a les arguments (ARG0-5) qui sont susceptibles d'être acceptés dans la phrase par le sens (*roleset ID*) et les autres arguments facultatifs qui sont présents (ARGM). Il est important de mentionner que des rôles sont généralement attachés aux ARG0-5 lors de la description de chacun des *roleset IDs*, toutefois, il n'est pas nécessaire de l'ajouter lors de l'annotation. C'est tout le contraire pour l'argument modificateur (ARGM), ce type d'étiquette montre qu'il y a un argument facultatif qui n'est pas présent dans la structure argumentale du verbe. Il est toujours porteur d'un rôle spécifique et il est important de l'étiqueter. Cette section propose une description du choix d'annotation des ARG0-5 & ARGM.

#### 4.1.2.1 Les arguments régis par le verbe ARG0-5

Il y a six choix d'arguments possibles qu'un verbe est susceptible d'accepter dans une phrase. PropBank priorise l'ordre suivant lors du choix d'argument : ARG0 > ARG1 > ARG2 - 5. Si un argument peut à la fois être considéré comme un agent et un patient, l'ARG0 devrait être utilisé.

TABLEAU 4.3 – Description des types d'arguments du verbe

| Type d'argument | Description | Exemple d'un *roleset ID* | Application SRL |
|---|---|---|---|
| ARG0 | agent | bay.01<br>ARG0-PAG : bay-er (vnrole : 38-Agent) | The hounds bayed at the rabbit.<br>Arg0 : The hounds<br>Rel : bayed<br>ArgM-DIR : at the rabbit |
| ARG1 | patient | crash.01<br>Arg0-PAG : causer of damage, agent<br>Arg1-PPT : entity crashed | his crash of the cymbal<br>Arg0 : his<br>Rel : crash<br>Arg1 : of the cymbal |
| ARG2 | instrument, benefactive, attribute | install.01<br>Arg0-PAG : putter<br>Arg1-PPT : thing put<br>Arg2-LOC : installed where or as | His installment of new gang members in crucial positions...<br>Arg0 : His<br>Rel : installment<br>Arg1 : of new gang members<br>Arg2 : in crucial positions |
| ARG3 | starting point, benefactive, attribute | text.01<br>Arg0-PAG : text sender<br>Arg1-PPT : message<br>Arg2-GOL : recipient<br>Arg3-VSP : subject matter of text (about what) | The wine spoke to me so and I texted him about it?..<br>that I was by no means inviting myself<br>to his b-day, I was just wondering when it was.<br>Arg0 : I<br>Rel : texted<br>Arg2 : him<br>Arg1 : that I was by no means inviting myself<br>to his b-day, I was just wondering when it was |
| ARG4 | ending point | rank.01<br>Arg0-PAG : assigner<br>Arg1-PPT : thing assigned a position<br>Arg2-PRD : rank<br>Arg3-VSP : position relative to other competitors<br>Arg4-VSP : attribute, value | Based on 1988 sales, Georgia-Pacific ranked third at<br>$9.51 billion, behind Weyerhaeuser Co.<br>at $10 billion and International Paper Co. at $9.53 billion.<br>Arg1 : Georgia-Pacific<br>Rel : ranked<br>Arg2 : third<br>Arg4 : at $9.51 billion<br>Arg3 : behind Weyerhaeuser Co. at $10 billion and<br>International Paper Co. at $9.53 billion |
| ARG5 | - | scale.01<br>Arg0-PAG : agent, entity causing change in size<br>Arg1-PPT : thing changing size<br>Arg2-EXT : EXT, amount changed by<br>Arg3-DIR : start point<br>Arg4-GOL : end point<br>Arg5-VSP : direction, up or back | The company's scaling their project back from $350 million<br>to a mere $275 million prevented<br>them from having downsize in other departments.<br>Arg0 : The company's<br>Rel : scaling<br>Arg1 : their project<br>Arg5 : back<br>Arg3 : from $350 million<br>Arg4 : to a mere $275 million |

Dans le Tableau 4.3, le ARG0 est représenté comme un agent du verbe. Généralement, les ARG0 sont le sujet d'un verbe transitif ou "a class of intransitive verbs called unergatives", comme suggéré par les propriétés proto-agent de Dowty (1991). Les ARG0 peuvent être des agents animés ou inanimés qui causent l'action ou le changement d'état. Alors, la condition principale est la raison (*causation*) de l'action ou du changement d'état.

(20) He (ARG0) landed.

De façon générale, le ARG1 est représenté comme des patients du verbe. Ces patients subissent un changement ou ils sont affectés par l'action. Généralement, les ARG1 sont l'objet d'un verbe transitif ou selon les propriétés proto-patient de Dowty (1991) : "the subjects of intransitive verbs called unaccusatives".

(21) (A bullet) ARG1 landed at his feet.

Les arguments 2 à 5 ne respectent pas autant les rôles d'agent ou de patient des ARG0 et ARG1. Bien que PropBank tente de les distinguer par l'association du rôle de l'*instrument* à l'ARG2, ou de *point de départ* pour l'ARG3 et de *point de fin* pour l'ARG4, ce n'est pas toujours le cas. Par exemple, pour *attach.01* ou *avert.01*, l'ARG3 porte le rôle de l'instrument dans la phrase. Pour distinguer si un élément de la phrase est un de ces quatre arguments, l'équipe d'annotation doit se référer à la description de chacun des arguments présentés par PropBank et par les exemples proposés.

#### 4.1.2.2 Les arguments modificateurs ARGM

Les arguments modificateurs (ARGM) sont des arguments qui ne sont pas des ARG 0-5 du verbe, mais ils apportent une information ou un rôle dans la phrase. Ces arguments portent toujours un rôle spécifique de modificateur (MOD, EXT, MNR...). Généralement, ils modifient le verbe, mais dans certains cas, ils modifient un ARG0-5. Selon le guide de PropBank, il faut toujours prioriser les ARG0-5 avant ces arguments modificateurs. Par exemple, le verbe *to travel* accepte comme ARG1 un argument porteur d'une localisation ou d'une direction. Il faut bien l'identifier comme un ARG1 et non comme un ARGM-LOC ou un ARGM-DIR. La Section 4.1.2.3 met en évidence chacun des rôles qu'un ARGM peut avoir.

#### 4.1.2.3 Les rôles modificateurs

Ces modificateurs accompagnent toujours un ARGM. Ils modifient un verbe ou un autre ARG0-5. Voici la liste de l'ensemble des *modifiers* (modificateurs) de PropBank selon Bonial et al. (2015).

TABLEAU 4.4 – Les modificateurs d'ARGM selon PropBank

| | |
|---|---|
| ADJ | Adjectival |
| ADV | Adverbials |
| CAU | Cause |
| COM | Comitative |
| CXN | Construction |
| DIR | Directional |
| DIS | Discourse |
| DSP | Direct Speech |
| EXT | Extend |
| GOL | Goal |
| LOC | Locative |
| LVB | Light Verb |
| MNR | Manner |
| MOD | Modal |
| NEG | Negation |
| PRD | Secondary Predication |
| PRP | Purpose |
| REC | Reciprocals |
| TMP | Temporal |

Le Tableau 4.4 montre les modificateurs ou *modifiers* qui spécifient le rôle de chacun des arguments dans la phrase. Comme précédemment exposés, ces modificateurs peuvent aider l'annotateur ou l'annotatrice à identifier un ARG1-PPT d'un ARG2-GOL puisqu'ici l'ARG1 aura le rôle de *patient* tandis que l'ARG2 aura le rôle du but de l'action. Pour ce qui est du ARGM, le modificateur spécifiera son rôle sans qu'il soit un argument ayant la susceptibilité d'être accepté par le sens du verbe dans la phrase (ARG0-5). En d'autres mots, ce sont des éléments de la phrase qui apportent

une information, une clarification ou une modification, mais qu'il ne sont pas classifiés comme un ARG0-5 de la structure argumentale du verbe.

(22) I sang a song with my sister.

ARG0 : I

REL : sang

ARG1 : a song

**ARGM-COM : with my sister**

TABLEAU 4.5 – Structure argumentale du *roleset ID* sing.01 [1]

| Roles | Description |
|---|---|
| ARG0-PAG | singer |
| ARG1-PPT | song |
| ARG2-GOL | audience |

Ce Tableau 4.5 montre que le *roleset ID sing.01* n'est pas susceptible d'accepter un argument qui porte le rôle de *comitative* ou plus précisément *avec qui l'action est effectué*, même si cette information est présente dans la phrase (22). Pour ce faire, le groupe d'annotation doit l'identifier comme un ARGM et spécifier son rôle (COM) dans la phrase. L'ensemble des exemples qui suivront sont intégralement tirés de la dernière convention d'annotation de Proposition Bank. Pour respecter leur description et pour empêcher de traduire imprécisément leurs exemples, ils seront conservés dans la langue originale qui est l'anglais.

#### 4.1.2.4 Comitatives - COM

Ce modificateur montre avec qui l'action a été faite. Cela peut être une ou plusieurs personnes ou une organisation (entité avec les caractéristiques d'un possible agent

1. https ://verbs.colorado.edu/propbank/framesets-english-aliases/sing.html

de la phrase : être animé et porteur d'une volonté). En anglais, ces modificateurs commencent généralement par la préposition *with*. Un exemple de ce rôle a précédemment été exposé (22).

#### 4.1.2.5 Adverbials - ADV

L'ADV est un modificateur de l'évènement d'un verbe ou d'un adjectif et il ne rentre pas dans un cadre d'un autre rôle que peut porter un argument modificateur. Le guide d'annotation expose de toujours prioriser les autres rôles modificateurs. Ce sont souvent des modificateurs :

- Temporels : Treasures are just lying around, **waiting to be picked up**.
- Intentionnels : probably, possibly.
- Apportant un focus : only, even.
- D'évaluation, d'attitude, de point de vue, de performance : fortunately, really, legally, frankly speaking, clauses beginning with given that, despite, except for, or if.

(23) Happily, she sang.
**ARGM-ADV : happily**

Un argument adverbial peut se rapprocher de la description d'un argument de *manner*. Comme mentionné, il faut prioriser l'ensemble des autres arguments. Le guide suggère une différenciation des deux arguments. L'ARGM-ADV modifiera généralement l'ensemble de la phrase tandis que l'ARGM-MNR modifiera le verbe. Si l'exemple (23) était plutôt exprimé comme *She sang happily*, l'élément *happily* serait un ARGM-MNR puisqu'il modifie l'action de chanter.

#### 4.1.2.6 Adjectival - ADJ

L'ADJ a un comportement similaire à l'ADV. La différence est qu'il se présente sous forme d'un adjectif.

(24) The mayor's shocking abuse of public funds outraged citizens.
ARG0 : The mayor's
**ARGM-ADJ : shocking**
REL : abuse
ARG1 : of public funds

#### 4.1.2.7 Cause - CAU

Le CAU indique les raisons d'une action, ils sont similaires aux PRP. Ces arguments débutent souvent par un *because* ou *due to*. Le guide d'annotation de Proposition Bank suggère de suivre cette règle d'application : Cause Clauses (CAU) > Purpose Clauses (PRP).

(25) However , five other countries - China , Thailand , India , Brazil and Mexico - will remain on that so-called priority watch list because of an interim review , U.S. Trade Representative Carla Hills announced.
ARGM-DIS : However
ARG1 : five other countries - China , Thailand , India , Brazil and Mexico -
ARGM-MOD : will
REL : remain
ARG3 : on that so-called priority watch list
**ARGM-CAU : because of an interim review**

#### 4.1.2.8 Construction - CXN

C'est un argument qui émerge d'une comparaison.

— ex : She is taller **than her sister**.

Ces arguments peuvent être des constructions comparatives *more/less/as-X than/as Y*.

— Hillary Clinton is **about as** damaging to the Dem Party **as Jeremiah Wright**

L'ensemble des éléments de la comparaison doit être annoté. Toutefois, si elle commence par un *more/less X than Y*, seulement la partie *than Y* doit être annotée. S'il y a seulement la première partie d'une comparaison (aucun second élément qui lui est comparé), aucune annotation ne doit être faite.

Ces arguments peuvent être des constructions de degré. Cette construction montre le degré d'une action effectuée ou si un état est *Vrai*, en mentionnant la conséquence du degré si l'état est *Vrai ('X is too/so Y to/that Z' ou 'X is not Y enough to/that Z')*

#### 4.1.2.9 Directional - DIR

Le *directional* montre un mouvement suivant une direction. Par contre, si le trajet n'est pas clairement suivi, l'argument sera plutôt un LOC (ARGM-DIR : 'walk along the road' vs ARGM-LOC : 'walk around the countryside'). L'ARGM-DIR peut être utilisé pour certaines 'particles' comme 'back up'.

(26) That response annoyed Rep. Markey , House aides said, and the congressman snapped back that there had been enough studies of the issue and that it was time for action on the matter .
ARG0 : the congressman
REL : snapped
**ARGM-DIR : back**
ARG1 : that there had been enough studies of the issue and that it was time for action on the matter

#### 4.1.2.10 Discourse - DIS

Les DIS sont des marqueurs de connexion qui connectent une phrase précédente.

— Exemples : also, however, too, as well, but, as we've seen before, instead, on the other hand, for instance, etc.

Pour tester si l'élément est un *DIS*, le guide propose de le supprimer et si la phrase a toujours le même sens, c'est effectivement un *DIS*. Généralement, ces arguments débutent une phrase, si ce n'est pas le cas, ils sont plutôt considérés comme des arguments adverbiaux.

— *Mary reads novels **(in addition ARGM-ADV)** to writing poetry.*

— ***(In addition ARGM-DIS)**, Mary reads novels.*

(27) But for now, they're looking forward to their winter meeting - Boca in February.

**ARGM-DIS : But**

ARGM-TMP : for now

ARG0 : they

REL : looking forward

ARG1 : to their winter meeting - Boca in February

#### 4.1.2.11 Direct Speech - DSP

Ces arguments sont présents avec des verbes qui expriment un discours ou une pensée d'une personne qui parle ou qui pense. Malgré sa présence dans le corpus de Conll2012 comme nous avons pu le constater dans le Tableau 2.5, Proposition Bank propose de les enlever et de les remplacer par des *LINK-DSP* puisqu'ils ne modifient pas un verbe, mais complètent généralement un ARG1. Ces éléments sont une trace de la précédente annotation des arbres syntaxiques effectués par Penn Treebank.

#### 4.1.2.12 Extends - EXT

L'*extend* montre la quantité ou le degré du changement qui est effectué suite à une action. Ces modificateurs sont souvent des :

— Compléments numériques, *raised prices by 15 percent*

— Quantifieurs, *a lot, least, incredibly, extremely, or really*

— Comparaisons, *he raised prices more than she did*

(28) PS of New Hampshire shares closed yesterday at $ 3.75, off 25 cents, in New York Stock Exchange composite trading.

ARG1 : PS of New Hampshire shares

REL : closed

ARGM-TMP : yesterday

**ARGM-EXT : at $ 3.75 , off 25 cents,**

ARGM-LOC : in New York Stock Exchange composite trading

#### 4.1.2.13 Goal - GOL

Le GOL dénote du but de l'action du verbe. Il inclut la destination d'un verbe de mouvement, le bénéficiaire de quelque chose et le modificateur qui indique que l'action du verbe a été produite pour quelqu'un ou quelque chose en leur nom *(on their behalf)*.

(29) We publicized to the masses our determination to fight against evil.

ARG0 : We

REL : publicized

**ARGM-GOL : to the masses**

ARG1 : our determination to fight against evil

#### 4.1.2.14 Locatives - LOC

Les *locatives* indiquent où l'action ou une partie de l'action est entreprise. Cet indice peut être une localisation physique, virtuelle ou abstraite.

(30) In his ruling, Judge Curry added an additional $ 55 million to the commission's calculations.
**ARGM-LOC : In his ruling**
ARG0 : Judge Curry
REL : added
ARG1 : an additional $ 55 million
ARG2 : to the commission 's calculations

#### 4.1.2.15 Light Verb - LVB

Les LVB sont considérés, selon Proposition Bank, comme des verbes légers ou des utilisations de verbes qui ne sont pas sémantiquement *forte*.

(31) He **made** a pie out of fresh cherries and refrigerated dough.

(32) She **made** an offer to buy the company for 2 million dollars.

Les exemples (31) et (32) sont tirés du guide d'annotation de PropBank (Bonial et al., 2015). On constate par l'exemple (31) que le verbe *make* respecte la structure argumentale proposée par Proposition Bank du *roleset ID make.01* qui est décrit comme le fait qu'il y a un ou une créatrice, un objet créé, *created from, thing changed* et un bénéficiaire. Tandis que dans l'exemple (32), le verbe *made* ne porte pas la charge sémantique de l'évènement (plutôt portée par le nom *offer*). La phrase peut être remplacée par :

(33) She offered to buy the company for 2 million dollars.

Proposition Bank propose l'identifiant verbe.**LVB** pour les verbes qui peuvent être utilisés comme des *light verbs*. D'autres verbes *light* sont mentionnés dans la ressource verbale : *pass, get ou take*.

#### 4.1.2.16 Manner - MNR

L'argument qui porte le modificateur MNR montrera la procédure de création d'une action.

(34) The plumber unclogged the sink with a drain snake.
ARG0 : The plumber
REL : unclogged
ARG1 : the sink
**ARGM-MNR : with a drain snake**

#### 4.1.2.17 Modal - MOD

Les arguments modaux sont les mots-formes identifiés comme des modales dans la langue anglaise :

— *can, must, shall, might, should, could and would.*

#### 4.1.2.18 Negation - NEG

Les arguments de négation sont toutes les formes négatives de la langue anglaise. Dans le cas de l'adverbe *never*, il devrait être annoté comme un ARGM-NEG même s'il pourrait dans une certaine mesure jouer un rôle de temporalité. De plus, il ne faut pas confondre l'ARGM-NEG avec des éléments comme *not only* qui ne marquent pas la négation du verbe. Ils ne devraient pas être annotés puisqu'ils sont plutôt des locutions (EPL) conjonctives. Enfin, le guide propose de suivre cette règle lors de l'annotation : ARGM-NEG > ARG-TMP.

#### 4.1.2.19 Secondary Predication - PRD

Comme le modificateur *REC*, il n'est pas relié au verbe, mais à un ARG0-5. Il apporte une information ou une précision à un argument du prédicat et non sur le prédicat. Les exemples ci-dessous montrent une modification apportée à un argument du verbe, en d'autres mots, ils montrent l'état avant et après l'évènement de l'argument et non une modification propre à l'évènement ou au verbe. Alors, il montre une modification d'état.

- Résultat : *The boys pinched them **dead**, She kicked the locker lid **shut***
- Illustre : ***Rosy-cheeked**, Santa came down the chimney*
- Syntagme prépositionel commençant par *as* : *supplied **as security in the transaction***

(35) Pierre Vinken , 61 years old , will join the board as a nonexecutive director Nov. 29.

ARG0 : Pierre Vinken , 61 years old ,

ARGM-MOD : will

REL : join

ARG1 : the board

**ARGM-PRD : as a nonexecutive director**

ARGM-TMP : Nov. 29

#### 4.1.2.20 Purpose - PRP

Le PRP montre la motivation d'une action. Les syntagmes qui commencent par des locutions prépositionnelles comme : *in order to* ou *so that*. Comme mentionné, le guide d'annotation de Proposition Bank suggère de suivre cette règle d'application : Cause Clauses (CAU) > Purpose Clauses (PRP).

(36) More than a few CEOs say the red-carpet treatment tempts them to return to a heartland city for future meetings.
ARG1 : them
REL : return
ARG4 : to a heartland city
**ARGM-PRP : for future meetings**

#### 4.1.2.21 Reciprocals - REC

Ce modificateur n'est pas relié à l'action du verbe, mais plutôt à un ARG0-5. Il inclut les pronoms ou *reflexives* et *reciprocals* comme ; *himself, itself, themselves, each other, own & both*. Ces éléments se réfèrent à un autre argument qui est généralement des ARG1.

(37) But voters decided that if the stadium was such a good idea someone would build it himself, and rejected it 59% to 41%.
ARGM-ADV : if the stadium was such a good idea
ARG0 : someone
ARGM-MOD : would
REL : build
ARG1 : it
**ARGM-REC : himself**

#### 4.1.2.22 Temporal - TMP

Le *Temporal* est l'argument qui montre la temporalité d'une action. Cela inclut les adverbes de fréquence *(often, always, sometimes*, avec l'exception de *never)* ou les adverbes de durée *(for a year, in a year)* ou l'adverbe d'ordre *(first)* ou de répétition *(again)*.

(38) Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.
**ARGM-TMP : recently**
REL : diagnosed
ARG2 : cancer

Cette section décrit les étapes et les étiquettes d'annotation de la convention de Proposition Bank. Dans le cadre de ce mémoire, cette convention a été appliquée, dans une certaine mesure, à des documents techniques. On constate que l'annotation en SRL est une tâche complexe puisqu'il faut bien identifier le bon sens de chacun des verbes d'une phrase pour permettre d'identifier ses arguments, ensuite, d'autres éléments de la phrase peuvent être des arguments modificateurs qui jouent un rôle dans la phrase.

## 4.2 Processus d'annotation du corpus CTeTex SRL

Cette partie propose de montrer l'application de la convention d'annotation de PropBank avec une plateforme d'annotation. On portera une attention particulière aux modifications apportées pour tenir compte des particularités linguistiques des requis logiciels. Enfin, l'évaluation des annotations par un accord inter-annotateurs sera proposée avec l'exposition de ses limites dans une tâche d'annotation en SRL.

### 4.2.1 Plateforme d'annotation Inception

Dans le cadre de l'annotation de requis logiciels et dans le but de construire un corpus de référence, c'est la plateforme Inception [1] qui a été utilisée. C'est un outil d'annotation avec un code source libre (*open source*) (Klie et al., 2018a,b). Inception a été utilisé pour la classification d'évènements sociaux et politiques (Wiedemann et al., 2022), l'extraction d'évènements olfactifs (Menini et al., 2022) & l'extraction d'information de matériels scientifiques (Friedrich et al., 2020).

Cette plateforme permet, entre autres, d'annoter des éléments d'un texte et de les classifier, de les lier entre eux et d'identifier des relations. Par défaut, la plateforme offre déjà un gabarit d'annotation de base pour certaines tâches comme, les coreferences, les entités nommées ou les catégories grammaticales. Ce qui est intéressant est qu'un utilisateur ou utilisatrice de la plateforme peut construire elle-même un projet d'annotation. C'est ce qui a été fait pour l'élaboration du corpus de référence des 196 requis logiciels. Ensuite, un accord inter-annotateurs est proposé par la plateforme avec différentes métriques d'évaluations : Cohen's Kappa (coding), Fleiss Kappa (coding), Krippendorff's Alpha (unitizing) & Krippendorff's Kappa (coding).

Pour bâtir un projet d'annotation avec la plateforme Inception, il faut passer par quelques étapes. La première est d'importer ses documents à annoter. Ensuite, il faut définir les couches d'annotation (*layers*). Elles se réfèrent aux grandes étapes d'annotation. Par exemple, dans le cadre du projet d'annotation de RL, les deux étapes principales sont la sélection et l'identification du sens du verbe et ensuite l'annotation des arguments. Nous avons donc défini 2 grandes couches d'annotation (*layers*) : FrameSet et ARG. De plus, nous avons décidé d'ajouter une couche nommée "Notes"

1. `https://inception-project.github.io/`

qui nous permettra d'identifier des éléments importants qui n'entrent pas dans l'une des deux couches principales. À l'intérieur de ses couches, nous avons spécifié des paramètres.

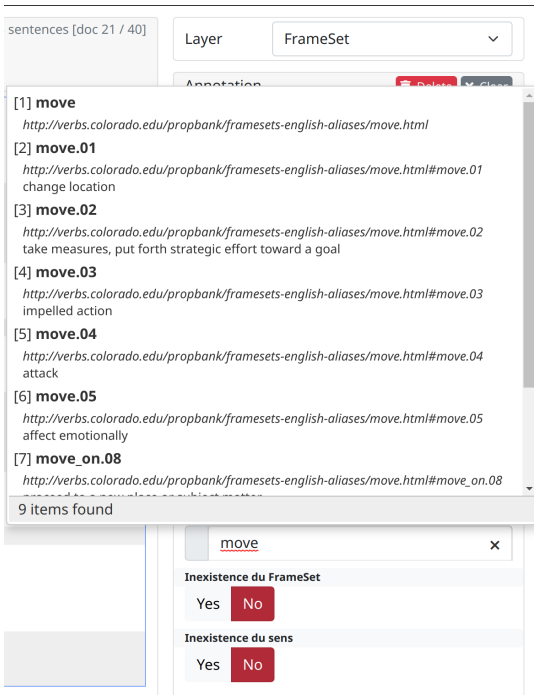



FIGURE 4.2 – Choix des sens du verbe depuis la plateforme d'annotation

Pour la couche du *FrameSet* ou le verbe, ses paramètres sont composés d'une base de connaissance qui a préalablement été construite et importée à la plateforme Inception. Cette base de connaissance rassemble des informations concernant l'ensemble des verbes de la ressource de PropBank. Elle contient l'ensemble des prédicats et chacune de leurs entrées (*roleset ID*). On y retrouve également le sens défini par la ressource ainsi qu'un lien *url* vers la description de chacun des verbes comme nous pouvons le constater par le Tableau 4.2. De plus, nous avions comme hypothèse que certains verbes et certains sens verbaux ne sont pas décrits par la ressource de Proposition Bank, donc nous avons ajouté deux paramètres à la couche d'annota-

tion du verbe : Inexistence du FrameSet (verbe) et Inexistence du sens. Le premier paramètre permet d'annoter si oui ou non (valeur par défaut) le lexème du verbe est présent dans la ressource. Le deuxième permet d'identifier lorsque le verbe est proposé par la ressource, mais qu'aucun sens ne correspond à son utilisation dans le requis logiciel.

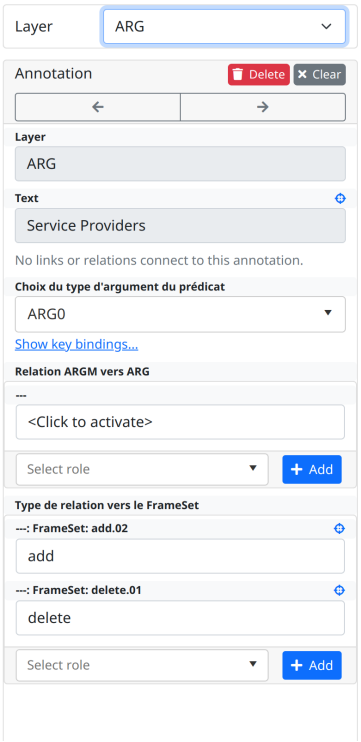



FIGURE 4.3 – Couche d'annotation des arguments

La deuxième couche d'annotation est celle des arguments. Elle permet d'identifier les arguments de verbes et les arguments facultatifs dans la phrase. Les arguments sont composés de trois paramètres. Le premier est le choix du type d'argument (ARG0-5 ou ARGM). Puisque certains arguments ne modifient pas exclusivement un verbe (ARGM-PRD & ARGM-REC), le deuxième paramètre permet de relier un argument à un autre en utilisant un jeu d'étiquettes (*tagset*) proposant ces deux rôles. Le dernier paramètre permet de relier un argument au verbe qui le régit en utilisant un

jeu d'étiquettes qui est composé d'une valeur nulle par défaut pour les ARG0-5 et de l'ensemble des autres rôles modificateurs pour les ARGMs.

### 4.2.2 Application des conventions d'annotation PropBank

#### 4.2.2.1 Choix du sens du verbe

Pour le choix du *roleset ID*, l'équipe d'annotation doit premièrement cibler chaque verbe du requis logiciel puisqu'ils ne sont pas annotés d'avance contrairement à PropBank. Comme précédemment mentionné, il est important de bien sélectionner le verbe et, s'il y a lieu, d'inclure toute forme adverbiale qui peut suivre un verbe comme *take up, act out, bail out*, etc. Ensuite, il faut choisir le bon *roleset ID* à partir d'une base de connaissance (*knowledge base*) qui comprend l'ensemble des prédicats, des *roleset IDs*, leur sens ainsi qu'un lien qui dirige vers la page web du prédicat de la ressource de Proposition Bank.

Cette base de connaissance a été créée pour faciliter les manipulations d'annotations et pour garder en mémoire l'emplacement de chaque verbe annoté. Ensuite, il est plus facile de retrouver certains verbes déjà annotés et ainsi rester cohérent lors des annotations.

#### 4.2.2.2 Choisir le type d'argument

Selon le choix du *roleset ID*, il faut repérer chaque argument régi par le sens du verbe au sein de la phrase. Comme exposé précédemment, il faut faire référence à la ressource verbale de Proposition Bank [1] pour le *roleset ID* en question.

1. `https://verbs.colorado.edu/propbank/framesets-english-aliases/`

Si d'autres éléments de la phrase ne sont pas décrits dans les arguments propres au *roleset ID*, ils peuvent être considérés comme des arguments modificateurs (ARGM) qui portent toujours un rôle de *localisation, temporalité, but, négation*, ou bien un des autres mentionnés à cette Section 4.1.2.3.

En général, l'ensemble du guide d'annotation de PropBank a été respecté et appliqué à la plateforme d'annotation *Inception*. Toutefois, certains éléments ne peuvent pas rendre compte de certaines particularités des requis logiciels. La prochaine section tentera de les exposer en plus de proposer une modification au guide d'annotation.

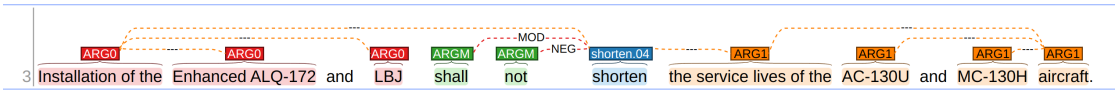
### 4.2.3 Adaptation de la convention

Le guide d'annotation de Proposition Bank a été élaboré dans le but d'annoter en sémantique de rôle un corpus provenant de journaux, de téléjournaux ou de parole spontanée transcrite. Ce type de document présente certaines divergences en comparaison avec notre corpus de requis logiciels. La première différence est l'objectif final de cette annotation et la deuxième est le type de contenu textuel. Cette section met en lumière ces divergences ainsi que les manipulations que nous avons apportées à l'annotation de ces documents.

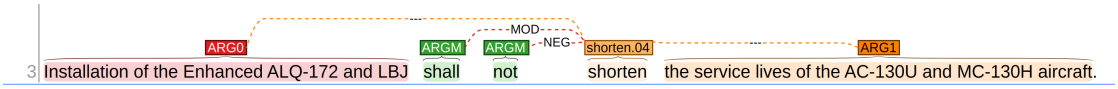
Comme Proposition Bank, le but de notre projet de recherche est de constituer un corpus annoté manuellement en sémantique de rôle. Toutefois, ce qui diffère dans le cadre de notre recherche est son objectif de génération de jeux de test à partir de requis comme nous l'avions mentionné dans l'introduction (Section 1). Pour nous rapprocher de cet objectif tout en respectant, dans une certaine mesure, l'annotation en sémantique de rôle, nous avons décidé de mettre au clair un élément important des requis logiciels avec un expert en ingénierie logiciel. Dans le but de générer des tests,

est-il important d'identifier chacun des éléments d'un argument qui sont séparés par une conjonction, une ponctuation ou un saut à la ligne ? Pour générer des jeux de tests, il est important d'identifier chacun des éléments qui doivent être testés selon une action et selon un ensemble de conditions. Pour illustrer, on retrouve premièrement l'action ou la transition d'état du système visé, ensuite le contexte opérationnel de cette action (l'état du système lors de son déclenchement) et finalement les critères de succès à évaluer. Ces derniers servent à vérifier si l'action est effectuée selon des exigences préétablies. C'est pour cette raison que nous avons décidé de diviser chacun des sous-éléments d'un argument et ainsi d'éviter qu'un système doive les séparer a posteriori.

Au contraire, avec la convention d'annotation de Proposition Bank, ces éléments seraient rassemblés en un seul segment argumental. En suivant cette directive, l'annotation serait certes plus rapide, puisqu'il n'y a pas besoin de sous-diviser les éléments. Toutefois, elle ne permettrait pas de répondre à notre objectif de structurer logiquement un requis logiciel par l'identification de chaque système ou entité, action, état et critère. Les figures 4.4a et 4.4b montrent la comparaison de l'application de notre guide d'annotation et celui de Proposition Bank.




(a)




(b)

FIGURE 4.4 – (a) Application du guide d'annotation de CTeTEx SRL en divisant l'ARG1 du verbe *shorten* (b) Application du guide d'annotation de Proposition Bank qui ne découpe pas les sous éléments de l'ARG1 du verbe *shorten*

Cette annotation proposée dans le cadre de ce mémoire permet de bien distinguer chacun des éléments au sein d'un argument sans avoir recours à un post-traitement de segmentation. On pourrait grossièrement représenter ce requis selon une structure logique :

(39) Installation of the Enhanced ALQ-172 and LBJ shall not shorten the service lives of the AC-130U and MC-130H aircraft.

**Agents :**

- Installation of the Enhanced ALQ-172

- Installation of the LBJ

**Action :**

- shorten (make shorter)

**Choses devenant plus petites :**

- the service lives of the AC-130U aircraft

- the service lives of the MC-130H aircraft

**Modificateurs de l'action** :

- shall (MOD)

- not (neg)

Dans l'exemple (39), on peut donc conclure que les deux agents de la phrase **ne doivent pas** raccourcir la durée de vie des deux avions.

Comme précédemment mentionné, le type de document diverge des corpus de Proposition Bank CoNLL05 et CoNLL2012. Conséquemment, certains rôles modificateurs n'ont pas été proposés dans le cadre de l'annotation pour diminuer les possibles erreurs d'annotation. En effet, les *Direct Speech* et les *Light verb*, qui sont généralement les rôles associés à des arguments modificateurs (ARGM), n'ont pas été proposés comme choix d'étiquette lors de l'annotation. Ce choix vient du type de document et des informations inscrites dans le guide d'annotation. En effet, comme

précédemment mentionné à cette Section 4.1.2.11, PropBank propose de modifier ces arguments pour une autre étiquette et donc de ne plus considérer ces *DSP*. De plus, en aucun cas, un requis logiciel ne propose un discours ou une pensée venant d'une machine, d'un système ou d'un agent humain. Par respect pour la convention d'annotation et le type de document, ce rôle modificateur a donc été enlevé des choix d'annotation. L'autre rôle modificateur enlevé est *Light verb* présenté à cette Section 4.1.2.15. Même s'il est présenté comme un modificateur d'ARGM dans la convention d'annotation, cette étiquette semble plutôt se rapprocher d'une entrée d'un prédicat (*roleset ID*). De plus, un requis logiciel est ambigu puisqu'il est écrit dans une langue naturelle. Cette ambiguïté peut être source de perte de temps et de budget si le requis logiciel est mal interprété (Iqbal et al., 2018). Généralement, un requis logiciel sera donc écrit en étant le moins ambigu possible. Les *light verbs* semblent moins appréciés que les verbes concrets par les locuteurs et locutrices de trois variations de l'anglais (anglais singapourien, hongkongais et britannique) (Mehl, 2017). Alors, cette étiquette ne sera pas proposée lors de l'annotation des arguments modificateurs, mais tout de même permis lors du choix du *roleset ID* d'un verbe.

Une attention particulière a été portée pour les choix des verbes. L'hypothèse est que la ressource verbale de PropBank ne tiendra pas compte de certains verbes ou certains sens verbaux du domaine de l'informatique. Si lors de l'annotation, un verbe ou un sens du verbe n'était pas présent dans la ressource verbale de Proposition Bank, deux choix étaient offerts aux annotateurs et annotatrices comme montre cette Figure 4.2. Ce sont deux annotations booléennes avec la valeur *fausse* par défaut qui montre que le verbe et le sens sont bien dans la ressource verbale. Toutefois, lorsque ce n'était pas le cas, l'annotateur pouvait cliquer sur *Inexistence du FrameSet* si le verbe n'était pas dans la ressource ou *Inexistence du sens* si le sens du verbe

n'était pas proposé. Ces paramètres d'annotation ont pour but de facilement cibler des lacunes de la ressource de Proposition Bank à la suite du projet d'annotation.

## 4.3 Évaluation inter-annotateurs de CTeTex SRL

Pour évaluer les annotations, certaines métriques proposent d'analyser les décisions prises par l'équipe d'annotation. Ces métriques permettent de juger l'accord des annotations. Il est connu qu'un corpus doit être correctement annoté manuellement pour ainsi être réutilisé pour, par exemple, constituer des données d'entrées pour un modèle d'apprentissage neuronal. Ce processus s'appelle : l'accord inter-annotateurs. Cet accord est couramment utilisé lorsqu'un projet est composé de plusieurs phases d'annotation manuelle d'un corpus.

Pour garder une cohérence avec les autres corpus annotés en SRL proposés par le projet de Proposition Bank, nous avons utilisé la même métrique d'évaluation de l'accord inter-annotateurs. La mesure de Kappa de Cohen est celle utilisée dans le cadre de ce mémoire (Siegel and Castellan, 1988). La plateforme utilisée dans le cadre de ce projet, Inception, propose la fonctionnalité d'évaluer l'accord inter-annotateurs. Celle-ci a donc été utilisée pour calculer cet accord.

### 4.3.1 Résultats de l'accord inter-annotateurs

Cette section expose les résultats de l'évaluation de l'accord inter-annotateur par la mesure de Kappa de Cohen. Cette analyse a été effectuée sur environ 20% du corpus de CTeTex SRL qui, au total, se compose de 40 requis tirés au hasard. Deux précédents jets d'annotation ont été effectués sur 10 requis avant d'entreprendre l'évaluation inter-annotateurs.

TABLEAU 4.6 – Éléments pris en compte dans le calcul de Kappa de Cohen

| | | | |
|---|---|---|---|
| A1 : | \|ARG1\| | \|ARG0\| | **\|ARG1\|** |
| A2 : | \|ARG1\| | \|ARG3\| | **\|ARG1\|** |
| | A | D | **NULLE** |

Le Tableau 4.6 met en lumière ce qui est pris en compte dans le calcul de Kappa de Cohen. Les A1 et A2 représentent l'annotateur et l'annotatrice. Les éléments qui se superposent dans leur segmentation ou, en d'autres mots, qui commencent et qui finissent par le même token sont ceux pris en compte dans le calcul de Kappa de Cohen. Par exemple, la dernière colonne de droite représente deux annotations qui ne sont pas incluses (nulle) dans ce calcul.

TABLEAU 4.7 – Évaluation d'accord inter-annotateurs

| | |
|---|---|
| Choix du *RoleSet ID* | 73% |
| Choix du type d'argument | 70% |
| Liaison de l'argument vers le verbe | 92% |
| Choix du rôle de l'argument | 85% |
| **Moyenne de l'accord inter-annotateurs** | **80 %** |

Le Tableau 4.7 présente les différents résultats de l'accord inter-annotateurs pour différentes tâches d'annotation. Le premier résultat expose le choix du sens du verbe avec un pourcentage de 73% de taux d'accord entre les deux annotateurs. Généralement, les verbes de la ressource de Proposition Bank ont plusieurs entrées de sens possibles. Ensuite, le choix du type d'argument obtient un pourcentage le plus bas avec un taux de 70% d'accord. Le choix du type d'argument est l'étiquette de l'argument avec ARG0-5 et l'ARGM. La liaison vers le verbe a le taux le plus élevé avec 92%. Enfin, le choix du rôle de l'argument est de 85%. En somme, les annotations des 40 requis effectués par deux annotateurs et annotateurs ont un taux d'accord de 80% en considérant l'ensemble des types d'informations.

En somme, cette section expose la méthodologie de l'annotation du corpus CTeTex SRL. Une partie de cette section décrit le guide d'annotation de PropBank qui est utilisé dans le cadre de la constitution de ce corpus. Par la suite, ce chapitre met en lumière l'application et l'adaptation du guide avec la plateforme d'annotation Inception. Enfin, cette méthodologie appliquée a permis de constater les forces et les limitations du guide d'annotation de PropBank lors de son application à des RL. Malgré la complexité de la tâche d'annotation impliquant des délimitations de segments, des relations et la catégorisation des éléments selon des rôles prédéfinis, l'accord inter-annotateurs de 80% selon la métrique de Kappa de Cohen permet de conclure sur une bonne qualité des annotations du corpus.

# Chapitre 5

# Discussion

Cette section propose une discussion autour du corpus CTeTex SRL. Elle portera sur les résultats obtenus depuis la métrique de Kappa de Cohencalculée par la plateforme Inception pour évaluer l'accord inter-annotateurs. Par la suite, les limites de la convention d'annotation seront abordées et certaines solutions seront proposées.

## 5.1 Les particularités soulevées par la métrique Kappa de Cohen

Les résultats exposés par le Tableau 4.7 soulèvent des éléments importants. On peut constater que les deux premiers résultats sont très proches avec une différence de 3% entre le choix des *roleset IDs* et des types d'arguments. Souvenons-nous de ces deux tâches, la première consiste à identifier tous les verbes de la phrase. Sur la plateforme Inception, généralement, cette action est effectuée en double cliquant sur le ou les tokens pour qu'un surlignement apparaisse. Selon la ressource verbale de Proposition Bank, il faut choisir la bonne entrée qui correspond au *roleset ID* (sens) du verbe. Par la suite, les arguments du verbe sont identifiés selon le choix de la précédente tâche. Ces tâches d'annotation peuvent sembler simples, mais elles com-

portent certains défis. En effet, accéder à la signification du requis logiciel sans son contexte et sans connaissance experte dans le domaine de l'ingénierie logiciel est une tâche complexe. Par exemple, les acronymes ou les noms de systèmes ne présentent pas d'information transparente sur leur utilisation ou leur description, il fallait, dans une certaine mesure, se référer au document source qui contenait le nom du système pour le comprendre. Ainsi, il était difficile de bien identifier le *roleset ID* qui correspondait au sens exprimé dans la phrase. Nous savions également que ce choix avait un impact sur le choix des arguments propres au verbe.

Regardons plus précisément le résultat pour le choix du *roleset ID* avec un taux de 73% d'accord. Comme précédemment exposé, la tâche est d'identifier tous les verbes d'une phrase et de choisir leur *roleset ID* (sens) approprié. Certaines divergences d'annotation semblent être intéressantes. On constate que la plupart des différences sont en lien avec un désaccord du sens proposé par la ressource de PropBank. En effet, puisque nous partions de l'hypothèse que le domaine de l'informatique comportera certains sens verbaux rarement utilisés ou même jamais utilisés dans le type de corpus de CoNLL05 et de CoNLL2012, nous avions tendance à identifier cet aspect lors de l'annotation par différentes façons. Une des façons consistait à identifier le prédicat du verbe ou la classe supérieure. Par exemple, le mot-forme *executed* serait identifié comme *execute*. L'autre option était de chercher dans la ressource verbale, un autre verbe qui pouvait porter le sens exprimé. Si nous reprenons l'exemple du verbe *executed*, il pouvait être identifié à certains moments comme *run.01*. C'est effectivement le cas de plusieurs divergences d'annotation pour les verbes : [*execute, run*], [*place, put*], [*launch, run*], [*permit, allow*], [*integrate, support*], [*conduct, use*] et [*diagnose, identify*]. Ces derniers composaient la plupart des désaccords entre les annotations. On remarque que l'ensemble de ces verbes sont fréquemment utilisés dans un genre de corpus provenant du domaine de l'informatique. C'est moins le cas

pour le genre de corpus de PropBank. Il n'est donc pas surprenant que les définitions de ces verbes proposés par la ressource verbale de Proposition Bank apportent une ambiguïté d'annotation des verbes et ainsi une difficulté de bien identifier leurs arguments. Bref, ce taux, plus faible, de 73% d'accord pour l'identification des *roleset IDs* montre une certaine difficulté d'application du guide de Proposition Bank. Ce taux devrait être plus élevé si les *roleset IDs* des verbes précédemment exposés étaient acceptés comme un groupe de synonymes.

Pour ce qui est du résultat le plus bas avec 70% qui expose le taux d'accord pour le choix du type d'argument, il dépend du choix du *roleset ID*. Effectivement, si nous nous rappelons de cette tâche, elle consiste à surligner chacun des arguments du verbe et d'identifier son type. Son type peut être un argument propre au verbe (ARG0-5) ou un argument modificateur (ARGM). L'ensemble des arguments dépend intégralement de ceux qui sont acceptés par le *roleset ID* préalablement choisi. Comme nous l'avons exposé précédemment, le taux d'accord du *roleset ID* est de 73%. Il y a une différence de 3% avec le choix du type de l'argument et puisqu'ils sont reliés, cette faible différence n'est donc pas surprenante. De plus, comme la précédente tâche, celle de choisir le type d'argument semble être ambiguë. Premièrement, un argument contient généralement plusieurs tokens, il présente alors une longueur qui est définie par son nombre de token. À certains moments, les limites de début et de fin des arguments n'étaient pas les mêmes pour l'équipe d'annotation, mais les deux personnes avaient choisi la même étiquette. Souvent, certaines prépositions qui intègrent un syntagme nominal à la phrase n'étaient pas incluses dans l'annotation de l'argument. Ce problème provient d'une définition ambiguë des limites des arguments dans le guide d'annotation. Puisque nous avons généralement respecté le guide de PropBank, cet élément n'était pas abordé étant donné que les annotations des corpus de CoNLL05 et de CoNLL2012 sont faites sur des données préanno-

tées en arbre syntaxique (en constituant) depuis la convention d'annotation de Penn Treebank (Taylor et al., 2003). Alors, l'annotation des arguments se faisait au niveau du syntagme prépositionnel. Le corpus CTeTex SRL n'avait pas de préannotation en constituant, alors la limitation des arguments n'était pas donnée d'avance. Ainsi, cibler les limites d'un argument comportait une difficulté et cela peut expliquer, dans une certaine mesure, le résultat de 70%.

De plus, même si l'équipe d'annotation a annoté les deux mêmes arguments en l'identifiant avec le même type, mais qu'une préposition manquait ou qu'une ponctuation finale était ajoutée au segment, l'argument n'était pas inclus dans le calcul de Kappa de Cohen. Cette différence est illustrée par le Tableau 4.6. On pourrait, dans une certaine mesure, conclure qu'il y a un accord puisque l'étiquette est la même et que la majorité des tokens sont inclus dans l'argument. Le calcul de Kappa effectué par la plateforme Inception les annulerait plutôt de son calcul. Cette critique de la mesure d'évaluation de l'accord inter-annotateurs sur une tâche *span-based* a justement été soulevée (Ortmann, 2022). Si ces arguments étaient comptés dans le calcul de Kappa selon le fait que la majeure partie des segments sont annotés de la même façon, le score d'IAA serait plus élevé.

La tâche de liaison de l'argument vers le verbe est celle pour laquelle l'équipe d'annotation est le plus en accord avec un taux qui s'approche de la perfection (92%). Comme il est indiqué dans le Tableau 4.7, la tâche est de relier un argument identifié au verbe qui le régit. Généralement, on constate un bon accord entre les annotations. Les divergences proviennent majoritairement du verbe *to be*. Lorsqu'il est sous forme d'auxiliaire dans la phrase, il ne doit pas avoir d'argument qui lui est relié contrairement à sa forme copulative qui régit des arguments.

Ensuite, la dernière ligne du Tableau 4.7 montre un taux relativement élevé pour le choix du rôle de l'argument (85%). Comme précédemment mentionné, il y a un type d'argument (ARG0, ARG1, ARG2, ARG3, ARG4, ARG5 & ARGM), mais également son rôle dans la phrase. Pour ce qui est de l'argument qui est régi par un verbe (ARG0-5), ils n'ont pas d'étiquette de rôle. Par contre, selon certaines contraintes de la plateforme d'annotation Inception, pour qu'il y ait une relation entre un argument et son verbe, il doit y avoir une étiquette pleine. Alors, pour les arguments du verbe, l'étiquette "—" a été choisie pour représenter un rôle *null*. En contrepartie, les arguments modificateurs (ARGM) sont intégralement porteurs d'un rôle. Même avec une multiclasse ayant un nombre élevé de possibilités de rôle (18 en considérant "—"), le taux d'accord est élevé.

En bref, par ces résultats de l'IAA, ils nous ont permis de cibler les limites de l'application de la convention d'annotation de PropBank à un corpus de documents techniques. On distingue les limites du choix du *roleset ID* qui a un impact direct sur le choix du type d'argument.

### 5.1.1 Comparaison de l'IAA de CTeTex SRL et des corpus de Proposition Bank

L'accord intercodeur du corpus de CTeTex SRL se rapproche de celui du corpus de PropBank. Toutefois, notre IAA ne peut pas être comparé à celui du dernier corpus Conll2012 du projet *Propostion Bank* puisque l'information de l'accord inter-annotateurs n'est pas présente dans l'article présentant ce dernier. Alors, nous ferons une comparaison avec l'ancien corpus CoNLL05. En moyenne l'accord de CTeTex SRL est un peu plus bas que celui du projet du corpus CoNLL05 avec une différence d'environ 11 points de pourcentage. Cependant, le résultat de 91% du CoNLL05 ne tient pas compte du choix des *roleset IDs* (Palmer et al., 2005), l'inclusion peut le mo-

difier à la baisse ou à la hausse selon l'accord des annotations pour cette tâche. Si nous comparons uniquement les annotations des arguments du corpus de CTeTex SRL, le taux d'accord augmente à une moyenne de 82,3%. La différence entre notre corpus et celui du CoNLL05 baisse à environ 8.7 points de pourcentage.

Malgré le fait que le taux d'accord du corpus de CTeTex SRL est plus bas que celui du projet de Proposition Bank, il reste relativement élevé avec un taux qui se rapproche plus de la perfection (100%) qu'un taux qui se rapproche d'un accord de 50%. Dans une certaine mesure, le guide d'annotation avec la ressource verbale de Proposition Bank peut être appliqué pour un corpus hors domaine comme CTeTex SRL. Toutefois, comme nous l'avons pointé, la ressource verbale ne tient pas compte de certains verbes qui on un sens plus spécifique que l'ont retrouve majoritairement en informatique.

## 5.2 Améliorations de la convention d'annotation

Cette section montre quelques pistes d'évolution pour poursuivre le projet en lien avec l'élaboration de notre corpus de référence CTeTex SRL annoté en sémantique de rôle. Nous proposons d'enrichir la ressource verbale de Proposition Bank pour tenir compte du vocabulaire spécialisé des requis logiciels et d'inclure une nouvelle définition de l'ARG0. Pour finir, nous exposons les limites du calcul de Kappa de Cohen proposé par Inception pour une tâche comme de l'annotation en segment. Si notre gabarit d'annotation en SRL est réutilisé, il faudra tenir compte de ces limites lors de l'évaluation de votre accord inter-annotateurs.

Pour tenir compte des sens verbaux souvent utilisés dans le domaine de l'informatique, de nouveaux *roleset IDs* devraient être ajoutés à la ressource verbale. En ef-

fet, comme nous l'avons illustré, le taux d'accord du choix du *roleset ID* (73%) montre que la ressource verbale de Proposition Bank ne couvre pas certains sens verbaux que l'on retrouve dans CTeTex SRL. Puisque la structure argumentale de la phrase dépend intégralement du *roleset ID*, le fait qu'il se retrouve des lacunes dans cette ressource peut rendre difficile la tâche d'annotation en sémantique de rôle. Nous l'avons effectivement constaté par le taux d'accord de 70% pour le choix du type d'argument.

En plus d'ajouter de nouveaux *roleset IDs*, cette ressource doit revoir la définition d'un argument du verbe. Nous avons relevé une ambiguïté du rôle de l'ARG0 pour certains verbes couramment utilisés dans les requis logiciels (*execute, run, integrate* ou encore *support*). Dans le guide, l'ARG0 est l'agent de l'action. Lors de l'annotation, nous avons remarqué que l'ARG0 ne jouait pas totalement le rôle d'agent.

(40) {The system, the interface}**ARG0** shall {execute, run, integrate, support} quelque chose **ARG1**.

On constate dans l'exemple (40) que les éléments pouvant être identifiés comme un ARG0 ne jouent pas intégralement le rôle d'exécuteur, d'agent qui exécute, intègre ou supporte. Généralement, en informatique ce qui exécute, roule, intègre ou supporte quelque chose n'est pas un système ou une interface en entier. C'est plutôt une fonction d'un script préalablement appelé par un stimulus qui déclenche une action quelconque et cette dernière est exécutée sur un système d'opération d'une machine dotée d'un processeur et de plusieurs autres composants. Bref, les précédents ARG0s jouent plutôt un rôle de localisation virtuelle de l'action effectuée. Ils se rapprochent donc d'un argument modificateur de localisation (ARGM-LOC). Pourtant, la ressource de Proposition ne le spécifie pas. Il serait tout de même incorrect de l'annoter comme un ARGM-LOC puisque chacun de ces verbes est susceptible de régir un argument qui produit l'action. Nous proposons donc d'enrichir la définition d'agent de l'ARG0 à la possibilité d'être un agent ayant une localisation virtuellement

sur une machine pour inclure cette particularité des requis logiciels.

Par ailleurs, nous avons soulevé que, dans la plupart des cas, la mesure de Kappa de Cohen utilisée pour de la sémantique de rôle ne permet pas de bien mesurer l'accord d'une annotation en segment. Pour des projets d'annotation en SRL avec la plateforme Inception, il faut en tenir compte.

(41) Annotation 1 : |Continuous BIT **ARG1**| |shall **ARGM-MOD**| |execute| on |the FCP virtual group **ARG0**|.

(42) Annotation 2 : |Continuous BIT **ARG2**| |shall **ARGM-MOD**| |execute| |on the FCP virtual group **ARG0**|.

Les exemples (41) et (42) montrent deux annotations de la même phrase (deux annotations fictives sur la même phrase tirée du corpus CTeTex SRL). On constate un désaccord pour l'annotation du premier argument (ARG1 vs ARG2), ensuite un accord pour l'ARGM-MOD, par contre le dernier argument ARG0 ne serait pas inclus dans le calcul de Kappa puisque les deux annotations n'ont pas la même longueur. La première annotation inclut la préposition *on*, mais pas la deuxième. Pourtant, si nous découpons l'argument en tokens, sur les 5 mots-formes, 4 sont identifiés comme un ARG0. Alors, l'argument possède un accord de 4/5 ou de 80%. Pour mieux calculer le taux d'accord pour de l'annotation en segments, il faudrait utiliser une autre mesure qui n'est pas proposée par Inception et qui peut regarder de façon plus granulaire chacun des segments.

En conclusion, de nouveaux *roleset IDs* et de nouvelles définitions de la structure argumentale devraient être intégrées à la ressource verbale de Proposition Bank. Ces propositions pourraient permettre d'augmenter le taux d'accord. Il faut aussi bien considérer les limites de la métrique de Kappa de Cohen de la plateforme Inception puisqu'elle exclut les arguments qui n'ont pas la même longueur. Cette convention

d'annotation enrichie permettrait d'être appliquée plus correctement à de nouvelles données qui proviennent de documents techniques.

# Chapitre 6

# Conclusion

Ce mémoire propose une méthodologie qui tente de représenter, d'une structure logique, les requis logiciels par une analyse en sémantique de rôle (Palmer et al., 2010). Pour ce faire, un corpus de référence contenant 196 RL a été annoté manuellement. De nouvelles perspectives d'annotations ont été proposées pour répondre aux défis qu'apporte un tel genre de corpus. Nous apportons un enrichissement de la structure argumentale du verbe en découpant ses arguments et nous relevons la non-applicabilité de la ressource verbale pour certains verbes du corpus CTeTex SRL. Ce corpus de documents techniques pourra être réutilisé comme données tests pour évaluer les performances d'un modèle d'apprentissage neuronal qui produit de la SRL.

Pour comprendre la théorie de la sémantique de rôle, nous avons abordé d'autres théories qui la précèdent. Le structuralisme a instauré la valence verbale qui se définit par le fait qu'un verbe peut régir un nombre $X$ d'actants. Par la suite, la grammaire de cas de Fillmore (1968) a permis d'aborder les rôles que jouent les arguments d'un verbe. On constate l'apparition des premiers rôles : l'agent, le patient et l'instrument (Fillmore, 1968). La grammaire de *cas* est par la suite étendue à la sémantique

de *cadres* qui propose d'autres rôles que peuvent jouer les arguments d'un verbe. Nous avons parcouru certaines ressources verbales comme FrameNet de Baker et al. (1998) qui instaurent des classes abstraites qui regroupent une quantité de verbes partageant des propriétés. VerbNet de Schuler (2006) a aussi été une autre ressource illustrée dans le cadre de ce mémoire. Elle s'inspire des travaux de Levin (1993) pour construire des classes abstraites et des niveaux de classes que peuvent porter un ensemble de verbes. VerbNet propose aussi 24 rôles thématiques. Enfin, le projet Proposition Bank de Palmer et al. (2005) a été présenté. Ce dernier a introduit les premiers travaux sur la théorie de la sémantique de rôle, il propose une ressource verbale décrivant les sens que peut porter un verbe et enfin il publie un guide d'annotation en SRL et ses corpus annotés selon ce dernier.

Par la suite, nous avons parcouru le corpus PURE de Ferrari et al. (2017) qui regroupe des documents contenant des requis logiciels. Suivant un travail d'extraction des RL du corpus PURE, CTeTex SRL a été constitué. Les requis logiciels devaient répondre à certaines conditions pour être extraits. La première était qu'il devait être sous forme textuelle et la deuxième était que le RL devait être constitué d'assez d'informations pour, ensuite, générer un jeu de test. Ces requis logiciels comportent des particularités linguistiques qui sont généralement moins présentes que d'autres corpus annotés en SRL. En effet, il comporte des notations mathématiques, des abréviations et acronymes, des listes et énumérations et un vocabulaire spécialisé. En comparant les annotations manuelles de deux exemples à la plateforme de pointe AllenNLP (Gardner et al., 2017), nous avons exposé que ces caractéristiques linguistiques peuvent apporter un défi à un tel système d'annotation automatique.

Une méthodologie d'annotation en sémantique de rôle a été proposée. La convention d'annotation du projet Proposition Bank a été décrite. Elle consiste à appliquer

la théorie de la sémantique de rôle à une phrase. Les tâches d'annotation consistent à choisir un *roleset ID* et ensuite de cibler ses arguments en se référant à la ressource verbale de PropBank et ensuite de cibler les arguments modificateurs et leur rôle. Nous avons proposé une nouvelle méthodologie d'annotation en SRL de documents techniques (requis logiciels). Nous l'avons appliquée à la plateforme d'annotation Inception. Les présentes modifications permettent de tenir compte des sous-éléments au sein d'un argument et d'apporter une attention particulière au choix du *roleset ID*. Finalement, nous avons évalué les annotations avec un accord inter-annotateurs par la métrique Kappa de Cohen. En moyenne, le taux d'accord pour le choix du type d'argument, de son rôle et de sa liaison vers son verbe est d'environ 82.3%.

Ce projet de recherche propose une nouvelle convention d'annotation en sémantique de rôle appliquée à des documents techniques. Cette nouvelle convention permet de tenir compte des particularités linguistiques des requis logiciels. Elle permet aussi de structurer logiquement le requis depuis une annotation en sémantique de rôle dans le but de générer des jeux de tests. Nous exposons enfin que la ressource verbale de PropBank n'est pas appropriée pour certains verbes souvent utilisés dans les requis logiciels.

CTeTex SRL est l'un des premiers corpus composé entièrement de requis logiciels annotés en sémantique de rôle qui suit un guide d'annotation rigoureux. Nous espérons que ce premier corpus de référence suscitera l'intérêt pour d'autres travaux sur ce sujet. Il peut jouer le rôle d'un corpus de référence qui permettra, entre autres, d'évaluer les performances des modèles neuronaux récents qui effectueront de la sémantique de rôle.

# Bibliographie

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, page 86–90, Montreal, Quebec, Canada, Aug 1998. Association for Computational Linguistics. doi : 10.3115/980845.980860. URL https://aclanthology.org/P98-1013.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena Hwang, Martha Palmer, and Nicholas Reese. English propbank annotation guidelines, 2015.

Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task : Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0620.

Jinho D Choi, Claire Bonial, and Martha Palmer. Jubilee : Propbank instance editor guidelines (version 2.1), 2009.

CRIM. À propos : Centre de recherche informatique de montréal, May 2022. URL https://www.crim.ca/fr/a-propos/.

David Dowty. Thematic proto-roles and argument selection. *Language*, 67(3) :

547–619, 1991. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/415037.

Peter D. Eimas, Einar R. Siqueland, Peter Jusczyk, and James Vigorito. Speech perception in infants. *Science*, 171(3968) :303–306, 1971. doi : 10.1126/science.171.3968.303. URL https://www.science.org/doi/abs/10.1126/science.171.3968.303.

Alessio Ferrari, Giorgio Oronzo Spagnolo, and Stefania Gnesi. Pure : A dataset of public requirements documents. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*, page 502–505, Lisbon, Portugal, Sep 2017. IEEE. ISBN 978-1-5386-3191-1. doi : 10.1109/RE.2017.29. URL http://ieeexplore.ieee.org/document/8049173/.

Charles J. Fillmore. The case for case. In Emmon W. Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart & Winston, New York, 1968.

Charles J. Fillmore. Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, 280(1) :20–32, 1976. doi : https://doi.org/10.1111/j.1749-6632.1976.tb25467.x. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1976.tb25467.x.

Charles J. Fillmore. *The Case for Case Reopened*, pages 59 – 81. Brill, Leiden, The Netherlands, 1977. ISBN 9789004368866. doi : https://doi.org/10.1163/9789004368866_005. URL https://brill.com/view/book/edcoll/9789004368866/BP000005.xml.

Charles J. Fillmore. Frame semantics. *Linguistics in The Morning Calm*, page 110–137, 1982. doi : http://brenocon.com/Fillmore\%201982_2up.pdf.

Charles J. Fillmore. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2) :222–254, 1985.

Charles J. Fillmore and Collin F. Baker. Frame semantics for text understanding, 2001.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruscyk, and Lukas Lange. The sofc-exp corpus and neural approaches to information extraction in the materials science domain. In *ACL*, 2020.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp : A deep semantic natural language processing platform, 2017.

Services publics et Approvisionnement Canada Gouvernement du Canada. Arbre [20 fiches] - termium plus - recherche;termium plus, Aug 2022a. URL `https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra&srchtxt=arbre&i=&index=alt&sg_kp_wet=536695&fchrcrdnm=20#fichesauve-saverecord20`.

Services publics et Approvisionnement Canada Gouvernement du Canada. Clef [14 fiches] - termium plus - recherche;termium plus, Aug 2022b. URL `https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra&i=1&srchtxt=clef&codom2nd_wet=1#resultrecs`.

H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics : Vol. 3 : Speech Acts*, pages 41–58. Academic Press, New York, 1975. URL `http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf`.

Maurice Gross. Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, 11(2) :151–185, 1982. doi : https://doi.org/10.7202/602492ar.

Naïma Hassert, Pierre André Ménard, and Edith Galy. UD on software requirements : Application and challenges. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 62–74, Sofia, Bulgaria, December 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.udw-1.5.

Bernd Heine and Heiko Narrog. *The Oxford Handbook of Linguistic Analysis*. Oxford University Press, 02 2015. ISBN 9780199677078. doi : 10.1093/oxfordhb/9780199677078.001.0001. URL https://doi.org/10.1093/oxfordhb/9780199677078.001.0001.

Tahira Iqbal, Parisa Elahidoost, and Levi Lúcio. A bird's eye view on requirements engineering and machine learning. In *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*, pages 11–20, 2018. doi : 10.1109/APSEC.2018.00015.

ITEA3. Home, Jun 2022. URL https://ivves.eu/.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. Universal proposition bank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.181.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *AAAI/IAAI*, page 6, 2000.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform : Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018a. URL http://tubiblio.ulb.

tu-darmstadt.de/106270/. Event Title : The 27th International Conference on Computational Linguistics (COLING 2018).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The INCEpTION platform : Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, pages 5–9, Santa Fe, New Mexico, August 2018b. Association for Computational Linguistics. URL https://aclanthology.org/C18-2002.

B. Levin. *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago Press, 1993. ISBN 978-0-226-47533-2. URL https://books.google.ca/books?id=6wIZWOrcBf8C.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19 (2) :313–330, 1993. URL https://aclanthology.org/J93-2004.

Seth Mehl. Light verb semantics in the international corpus of english : onomasiological variation, identity evidence and degrees of lightness. *English Language and Linguistics*, 23 :55 – 80, 2017.

Igor Mel'cuk. *Cours de morphologie générale : Introduction et première partie : Le mot*. La presses de l'Université de Montréal, CNRS Éditions, 1993.

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenić, and Anja Zidar. A multilingual benchmark to capture olfactory situations over time. In *LCHANGE*, 2022.

Katrin Ortmann. Fine-grained error analysis and fair evaluation of labeled spans. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1400–1407, Marseille, France, 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.150.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank : An annotated corpus of semantic roles. *Computational Linguistics*, 31(1) :71–106, Mar 2005. ISSN 0891-2017, 1530-9312. doi : 10.1162/0891201053630264.

Martha Palmer, Daniel Gildea, and Nianwen Xue. *Semantic Role Labeling*. Springer International Publishing, Cham, 2010. ISBN 978-3-031-01007-1. doi : 10.1007/978-3-031-02135-0. URL https://link.springer.com/10.1007/978-3-031-02135-0.

Alain Polguère. Lexicologie et sémantique lexicale; notions fondamentales. les presses de l'université de montréal, collection paramètres, montréal, 2003, 260p., 2016.

Victoria Puzhevich. Functional vs non-functional requirements : Key differences, 2021. URL https://scand.com/company/blog/functional-vs-non-functional-requirements/.

Jacques Quellet. Igor a. mel'cuk, andré clas, et alain polguère. introduction à la lexicologie explicative et combinatoire. dans la série champs linguistiques. louvain-la-neuve : Éditions duculot/aupelf-uref. 1995. 256 pages. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 43(2) :240–244, 1998. doi : 10.1017/S0008413100020582.

Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Paris, 1916.

Karin Kipper Schuler. *VerbNet : A Broad-Coverage, Comprehensive Verb Lexicon*.

PhD thesis, University of Pennsylvania, 2006. URL http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf.

Sidney Siegel and N. John Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition, 1988.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. *The Penn Treebank : An Overview*, volume 20 of *Text, Speech and Language Technology*, page 5–22. Springer Netherlands, Dordrecht, 2003. ISBN 978-1-4020-1335-5. doi : 10.1007/978-94-010-0201-1_1. URL http://link.springer.com/10.1007/978-94-010-0201-1_1.

Lucien Tesnière. *Eléments de Syntaxe Structurale*. Klincksieck, Paris, 1959.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. OntoNotes Release 5.0, 2013. URL https://hdl.handle.net/11272.1/AB2/MKJJ2R.

Gregor Wiedemann, Jan Dollbaum, Sebastian Haunss, Priska Daphi, and Larissa Meier. A generalized approach to protest event detection in german local news, 06 2022.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Zhenghua Li, Guohong Fu, and Min Zhang. Semantic role labeling as dependency parsing : Exploring latent tree structures inside arguments, 2021. URL https://arxiv.org/abs/2110.06865.