**UNIVERSITÉ PARIS NANTERRE**

# Normalisation des structures de la définition : étude empirique à partir des articles de Wikipédia

Mémoire de Master 2

Mention : Traitement Automatique des Langues

Spécialité : Recherche et Développement

**Luigi (Yu-Cheng) LIU**

Sous la direction de

Sylvain KAHANE

Kim GERDES

*Septembre 2017*

# Table des matières

# Liste des symboles

\# Mot vedette

**ADJ** Adjectif (trait syntaxique)

**Circonst** Circonstancielle (relation syntaxique)

**Comp** Complétive (relation syntaxique)

**Copul** Copulaire (relation syntaxique)

**Déterm** Déterminative (relation syntaxique)

**GAdj** Groupe adjectival (trait syntaxique)

**GN** Groupe nominal (trait syntaxique)

**GPrép** Groupe prépositionnel (trait syntaxique)

**GV** Groupe verbal (trait syntaxique)

**IND** Indicatif (trait syntaxique)

**INFINI** Infinitif (trait syntaxique)

**Modif** Modificative (relation syntaxique)

**NOM** Nom (trait syntaxique)

**Obj-dir** Objectale-directe (relation syntaxique)

**Obj-indir** Objectale-indirecte (relation syntaxique)

**Obj-obl** Objectale-oblique (relation syntaxique)

**PL** Pluriel (trait syntaxique)

**PRON** Pronom (trait syntaxique)

**PRÉS** Temps Présent (trait syntaxique)

**Prép** Complétive-prépositionnelle (relation syntaxique)

**SG** Singulier (trait syntaxique)

**Subj** Subjectale (relation syntaxique)

**V** Verbe (trait syntaxique)

# Remerciements

Je tiens à exprimer toute ma reconnaissance à

Sylvain KAHANE

et

Kim GERDES

de m'avoir encadré, orienté et encouragé en tant que directeurs de mémoire.

Je remercie Alain POLGUÈRE et Lucie BARQUE pour l'éclairage sur la question de l'ambivalence sémantique.

Je remercie Marine COURTIN pour la relecture attentive de ce travail.

# 1 Introduction

De nombreuses applications du TAL telles que la traduction automatique, l'extraction d'information et le système de questions-réponses nécessitent entre autres, des traitements importants de nature sémantique. Contrairement à des analyses de surface, qui peuvent souvent s'appliquer sur le texte, l'analyse sémantique a besoin de ressources lexicales pour accéder au sens du texte qu'elle traite.

En TAL, les ressources sémantiques auxquelles on a souvent eu recours peuvent être classées en trois types (L'HOMME 2008) : bases de données lexicales, ontologies de domaine, ressources terminologiques.

Dans l'actuel mémoire, nous nous intéressons en particulier, pour des raisons liées à la couverture et la disponibilité de données, aux ressources terminologiques non-formalisées avec l'objectif de proposer une méthodologie pour normaliser ce type de ressources.

Par ressources terminologiques non-formalisées, nous entendons les définitions dans le dictionnaire et l'encyclopédie. Elles peuvent servir à l'analyse du sens (POLGUÈRE 2003), mais il est souvent insuffisant de connaître les définitions pour pouvoir différencier deux termes qui sont proches ou très similaires afin de rendre plus pertinentes leurs exploitations.

Pour illustrer nos points de vue, prenons comme exemple *menuisier* et *ébéniste*, deux métiers qui travaillent tous en étroite relation avec le bois. Nous citons les définitions suivantes de la Wikipédia francophone :

*Le menuisier est un professionnel qui travaille traditionnellement le bois.*
*L'ébéniste fabrique des meubles et panneaux en bois, [...]*

Nous remarquons que c'est au sein d'une structure copulative du type « NOM est un NOM qui V » que la définition du menuisier est construite. Son activité principale est décrite dans *travaille traditionnellement le bois*. Par ailleurs, elle donne l'information sur le genre prochain du *menuisier*, un hyperonyme proche de lui. Il s'agit ici d'un *professionnel*. Par contre, la seconde définition ne caractérise que l'activité de l'*ébéniste* par *fabrique des* *meubles et panneaux en bois* sans pour autant lui proposer un genre prochain.

Pour le second exemple ci-dessous, nous comparons le *boulanger* au *pâtissier* sur la base de leurs définitions Wikipédia :

*Le* *boulanger* *est un* *professionnel spécialiste* *de* *la fabrication du pain* *[...]*
*Le* *pâtissier* *est un* *artisan* *spécialisé dans* *la fabrication des pâtisseries*.

La première définition tout comme la seconde présente une structure copulative similaire à celle employée pour définir *menuisier*. Dans les deux cas, le genre prochain est effectivement donné.

Cependant, les genres prochains que les deux définitions contiennent ne sont pas égaux sur le plan de granularité : un *professionnel spécialiste* par rapport au *boulanger* est bien plus général, donc moins spécifique qu'un *artisan* par rapport au *pâtissier*.

De plus, si les deux métiers sont définis par leur spécialisation dans la fabrication de certains produits alimentaires, le sens de spécialisation est réalisé par un nom commun *spécialiste* pour le *boulanger*, et fait partie du genre prochain *professionnel spécialiste* tandis que pour le *pâtissier*, ce sens est réalisé par un participe passé suivi d'une préposition régie *spécialisé dans* et fait partie des marqueurs linguistiques de relation utilisés souvent pour signaler les traits spécifiques dans une définition.

D'où on constate les difficultés de comparer deux termes directement selon leurs définitions encyclopédiques et l'importance de les formaliser pour une meilleure comparabilité entre définitions pour permettre ensuite une étude fine de la différence entre termes ou d'autres applications ayant recours aux ressources lexicales, tel que la désambiguïsation lexicale par ressources lexicales (Brun, Jacquemin et Segond 2005 ; Tchechmedjiev 2012) et la construction automatique d'ontologies à partir de texte (Bourigault et Aussenac-Gilles 2003 ; Kamel et Aussenac-Gilles 2009).

La formalisation des définitions ou leur normalisation s'inscrit dans un cadre plus large, celui des recherches sur la normalisation de texte. La normalisation de texte est un processus qui consiste à transformer un texte en une forme standard. Dans ce mémoire, afin de mieux guider la tâche de notre objectif, la normalisation des structures de définition, nous proposons une grammaire de normalisation.

La grammaire de normalisation est une méthode symbolique (Brun et Hagege 2003 ; Blanc 2006), le module de traitement qui lui est associé est souvent placé après l'analyseur syntaxique mais avant les traitements d'analyse sémantique (Hagège et Roux 2003). Cette grammaire regroupe un ensemble de règles de normalisation, chaque règle précise les conditions qu'il faut remplir pour son application, et le type de transformation qui peut être appliqué une fois l'ensemble de ces conditions vérifiées.

L'étude présentée dans ce mémoire se structure de la manière suivante : La section 2 présente la problématique de ce mémoire. La section 3 explique le choix du corpus, détaille comment le corpus a été constitué. La section 4 présente les 5 types de définitions relevées dans notre corpus d'étude. Une grammaire de normalisation pour ces 5 types de définitions encyclopédiques est proposée dans la section 5. La section 6 décrit comment ce normalisateur de définitions a été conçu, détaille ensuite le résultat d'une expérimentation réalisée avec notre normalisateur de définitions sur un corpus d'évaluation. Après quoi, nous donnons une conclusion générale dans la section 7.

# 2 Problématique

L'objet d'étude de ce mémoire porte sur les définitions dans le dictionnaire ou dans l'encyclopédie. L'objectif consiste à rendre les définitions comparables. Pour y parvenir, trois approches sont possibles :

— Ressources lexicales normalisées (L'HOMME 2008)

— Règles de reconnaissance de paraphrases (POLGUÈRE 2011)

— Normalisation (ou standardisation) des définitions

La première consiste à créer ou importer des contenus lexicaux dans un modèle lexicographique formalisé et dans un langage descriptif et explicatif contrôlé. Le deuxième courant d'approche consiste à reconnaître les paraphrases ou les quasi-paraphrases d'une définition. Les règles de reconnaissance de paraphrases ou quasi-paraphrases permettent d'établir une correspondance entre les éléments lexicaux d'une définition et ceux d'une autre sans pour autant transformer leurs structures.

Nous nous intéressons ici à la troisième approche. Elle vise à reformuler les définitions dans une même structure, une même construction sans en perdre l'information, tout en conservant la fonction de définir. Or, il semble, qu'à première vue, il existe autant de constructions possibles pour définir que le nombre de définitions qui figurent dans un dictionnaire. Nous sommes donc confrontés à un problème de complexité et de diversité linguistique dans cette recherche sur la normalisation des structures de définition.

Pour voir de plus près ce problème, nous proposons de reprendre les premières définitions que nous avions considérées dans l'introduction. Elles sont reproduites ci-dessous :

*Le* *menuisier* *est un* *professionnel* *qui* *travaille traditionnellement le bois*. (1)
*L'* *ébéniste* *fabrique des meubles et panneaux en bois* *[...]* (2)

Déjà, on constate que leurs structures syntaxiques ne sont pas identiques. Visiblement, il n'y a pas lieu de les considérer comme issu d'un même type de définitions. Cependant, nous pouvons remarquer que la définition (2) possède quasiment la même structure syntaxique que la proposition subordonnée de la définition (1).

D'après des observations similaires, nous émettons l'hypothèse que les définitions simples ou simplifiées présentent des structures syntaxiques de base, qui pourraient servir à mieux analyser, factoriser les définitions plus complexes, et une méthode permettant de

normaliser les définitions simplifiées fournirait des avantages d'accès à d'autres méthodes dérivées qui servent à normaliser un nombre plus grand de définitions. Ensuite, il s'agit de distinguer trois types différents de relations entre deux définitions : soit l'une fait partie de l'autre ou inversement, soit elles peuvent être classées dans le même type de définitions, soit elles correspondent à deux types distinctifs.

Pour exploiter au mieux les deux derniers types de relation, nous nous proposons de situer cette recherche dans une perspective empirique, et étudier des modèles lexico-syntaxiques de définition que nous avons soulevés à travers une étude de corpus des définitions courtes. Sur la base de ces modèles de définition, nous nous proposons de construire une méthode symbolique qui sert à faire correspondance entre modèles par des règles de transformation.

Ainsi, nous divisons le processus de normalisation en deux étapes : la première étape consiste à identifier le type auquel une définition correspond [1], la seconde étape consiste à appliquer les règles de transformation sur le corpus dans le but de convertir les définitions dans un nombre plus réduit de modèles ou dans un seul modèle dans la mesure où cela est possible.

1. Lorsqu'une définition est susceptible de correspondre à la combinaison de plusieurs types, l'identification consiste à trouver la bonne combinaison de bons types afin de rendre compte de la structure de définitions.

# 3 Constitution du corpus

Pour constituer une ressource terminologique des termes comparables ou quasiment synonymiques, il est commode de choisir un domaine de connaissance.

Le domaine de métier semble approprié pour une comparaison des définitions de terme, car, de manière générale, un terme de métier désigne soit une personne qui l'exerce comme profession, soit une activité professionnelle. Dans les deux cas, le genre (personne ou activité) peut être identifié sans mobiliser des connaissances de spécialité, il en est de même pour l'étude de la structure. Ceci est moins le cas pour le domaine de médecine ou de droit. D'autre part, les termes de métier apparaissent fréquemment dans le fait divers, le reportage des journaux, il est par conséquent plus aisé de confronter les définitions et l'usage de termes dans un texte pour le domaine de métier.

C'est ainsi sur le thème de métier que nous nous proposons de former notre corpus en ne sélectionnant que les entrées pertinentes dans l'encyclopédie Wikipédia de langue française. Wikipédia est un projet d'encyclopédie collective, accessible sous forme de site Web. Il met à disposition de l'encyclopédie en version française 1 892 970 articles produits grâce à la collaboration des contributeurs internautes dont 14 633 sont considérés comme collaborateurs actifs.

Dans cette section, nous commençons par détailler les moyens d'accès aux données. Ensuite nous présentons la structure de l'encyclopédie Wikipédia dans la section 3.2 et 3.3. Quant à la technique que nous avons mise en oeuvre pour identifier et extraire les textes pertinents, elle est ensuite détaillée dans la section 3.4.

## 3.1 Moyens d'accès aux données

La fondation Wikimédia qui est à l'origine du projet Wikipédia et qui l'a accueilli dans sa plate-forme en ligne propose deux moyens d'accès aux articles de l'encyclopédie libre.

La première façon, la plus classique est de visualiser un article sous forme de page Web sur le site https ://fr.wikipedia.org/. En effet, chaque page Web consacrée à un article est pré-générée par un logiciel nommé MediaWiki à partir d'un texte enregistré en wikicode. Wikicode est un langage minimaliste de mise en page qui a été conçu pour faciliter la rédaction de textes pour Wikipéda. Un rôle de MediaWiki est donc d'assurer la conversion du format wikicode d'un texte vers le format HTML / CSS.

La seconde manière est d'accéder à une copie locale de Wikipédia que Wikimédia a mise en ligne dans le site Web (*Wikimedia Downloads* 2017) destiné à présenter des copies archivées de Wikipédia. Les archives pour chaque langue y sont classées par date.

Dans chaque archive, nous retrouvons systématiquement des sauvegardes intégrales de l'ensemble des articles sur Wikipédia de la version concernée, les copies des méta-données tels que les titres d'articles, les liens inter-pages, les liens de référence externes, et celle de l'historique d'édition par article et utilisateur, etc. Dans les fichiers des articles, la structure arborescente de l'encyclopédie a été formatée en XML, tandis que le contenu de chaque texte associé à un nœud particulier de la structure XML, est encodé en wikicode.

Par rapport aux moyens d'accès au contenu de Wikipédia, le choix s'est porté sur les copies locales pour les raisons suivantes : d'abord, nous nous intéressons aux textes bruts que les rédacteurs produisent, il s'agit dans ce cas-là plus des textes en wikicode que ceux du format Web, ensuite d'un point de vue pratique, un temps au parcours plus court et une meilleure stabilité d'accès nous ont incité à privilégier une interrogation locale des données.

## 3.2 Structure globale de Wikipédia

Wikipédia que nous étudions par l'accès local, comme nous l'avons mentionné plus tôt possède une structure arborescente que nous pouvons représenter par un arbre simple comme dans la figure 1. Il présente les détails sur la page intitulée *Menuisier*. Chaque nœud y est caractérisé par deux attributs obligatoires, à savoir, tag et text. L'attribut tag identifie la fonction ou le type du contenu textuel que le nœud contient, tandis que text est une variable où se trouve le contenu textuel du nœud.

Dans la figure 1, un nœud étiqueté `title: Menuisier` étant un descendant direct d'un nœud étiqueté **page** signifie que le **titre** de cette **page** actuelle est **Menuisier**. Pour l'exemple de *menuisier*, sa page a notamment des enfants comme **title** qui permet d'intituler la page, le nœud **révision** qui permet de rattacher à la même page plusieurs révisions d'un texte [2].

Par rapport à un nœud **révision**, les descendants qui vont nous intéresser sont **model**, **format** et **text** que nous commentons un par un : **model** décrit l'encodage du contenu

2. Dans la figure 1, nous remarquons que la page de Menuisier ne possède qu'un nœud intitulé **révision**, cela est du au fait que le fichier que nous avons obtenu de la part de Wikipédia est une copie allégé de la version intégrale dont seul la dernière révision de chaque page a été conservée.

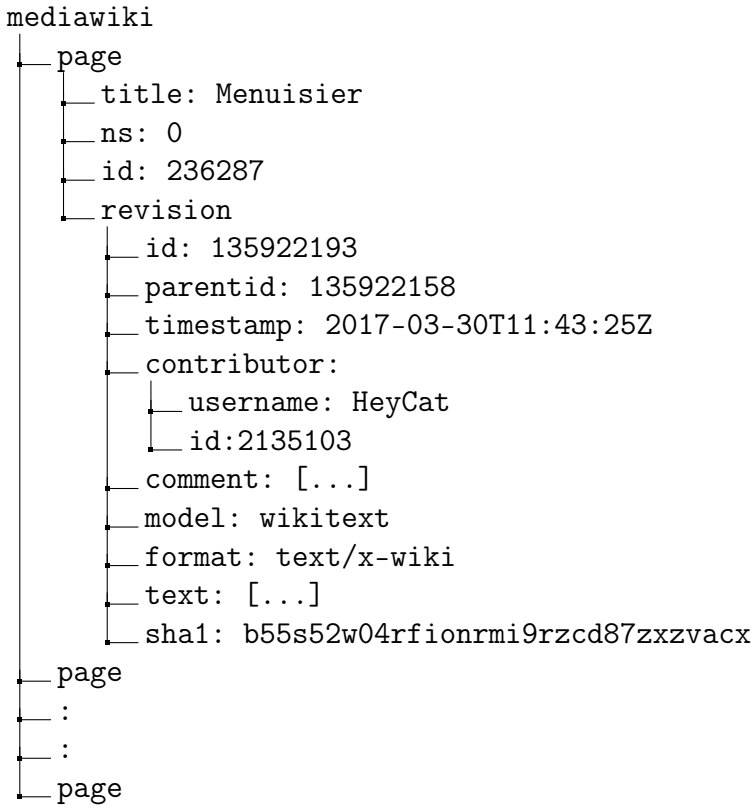


FIGURE 1 – Structure arborescente de Wikipédia illustrée par la page intitulée Menuisier.

dont le format est précisé par le nœud **format**, le nœud **text** contient le contenu textuel de l'article, ce contenu, c'est ce qui nous intéresse principalement dans ce mémoire.

## 3.3 Structure locale de Wikipédia

La syntaxe wikicode (WIKIPÉDIA 2017a) permet non seulement une mise en forme stylistique de l'article, mais aussi de définir un renvoi vers un article de Wikipédia ou en dehors de lui.

L'analyse des expressions wikicode dans un article brut dans Wikipédia nous permet de bien distinguer des mises en formes uniquement esthétiques et des expressions ayant une fonction référentielle. Dans les pré-traitements qui précèdent l'analyse de structure de définitions, deux objectifs principaux peuvent ainsi être fixés : le nettoyage des expressions non-référentielles, l'interprétation des liens référentiels internes à l'article.

### 3.3.1 Sections et inter-titres

Un article peut être organisé en sections. Pour déclarer une section, il consiste à en déclarer le titre ou plus précisément l'inter-titre. Pour définir un inter-titre, on entoure le

nom de l'inter-titre par une balise réalisée avec des paires de signes d'égalité : par exemple, `== Formation ==` permet de déclarer une section intitulée « Formation » de niveau 1, `=== En France ===` permet de déclarer une autre intitulée « En France » de niveau 2 et ainsi de suite.

### 3.3.2 Liens internes

**3.3.2.1 Liens vers un article** Dans un texte sur Wikipédia, une paire de doubles crochets carrés `[[]]` sert à définir un lien. Si les doubles carrés englobent un nom d'article de Wikipédia comme `[[ébéniste]]`, il s'agit d'un lien interne qui renvoie à la page dont le **titre** est *ébéniste*. Il est également possible de renvoyer à une section particulière d'un article sur Wikipédia, dans ce cas-là, nous mettons dans les doubles crochets carrés le titre de l'article et une dièse suivi par le titre de la section à laquelle on souhaite renvoyer comme `[[ébéniste#Formation]]`. Par extension, `[[#Formation]]` fait référence à la section Formation du même article dans lequel l'expression est employée. Pour personnaliser l'affichage du lien, on utilise le symbole de barre verticale `|` comme dans `[[ébéniste|wikilien vers ébéniste]]`, ce qui va créer un lien vers ł'article intitulé ébéniste tout comme `[[ébéniste]]`, mais c'est le texte « wikilien vers ébéniste » qui va figurer sur le lien dans le rendu visuel de l'article.

**3.3.2.2 Liens vers une page de catégorie** Une catégorie fait référence, dans le cas de Wikipédia, à une page intitulée par le nom de la catégorie et qui rassemble tous les liens d'article classés manuellement par les rédacteurs dans cette catégorie.

Pour catégoriser un article, nous y placerons un marqueur de catégorie, c'est-à-dire, un lien vers la page d'une catégorie prédéfinie ou une catégorie à définir[3]. La syntaxe pour déclarer une catégorie est illustrée par l'exemple suivant : `[[Catégorie:Métier du bois]]` est un marqueur qu'on trouve dans l'article d'*ébéniste* aussi bien dans celui de *menuisier*, qui a pour effet de classer l'article dans la catégorie de *métier du bois* .

Par conséquent, dans la page intitulée *métier du bois* de Wikipédia, nous retrouvons la liste des liens qui renvoient à tous les articles classés en « Métier du bois » . Par ailleurs, il est possible et courant de placer plusieurs marqueurs de catégorie dans un même article pour exprimer une multiple catégorisation.

3. Dans le cas où la catégorie n'existe pas encore, la page qui lui est associée sera créée automatiquement au moment de la déclaration du marqueur de catégorie WIKIPÉDIA 2017b.

#### 3.3.2.3 Liens vers une image

Pour placer un lien vers une image, nous suivons la syntaxe ci-dessous :

```
[[Image:Nom de l'image|thumb|left|alt=Texte alternatif|Légende]]
```

Puisque l'image ne fait pas partie de textes de définitions que nous étudions, nous allons supprimer tous les liens d'image dans la phase de nettoyage.

#### 3.3.2.4 Liens de redirection

Le lien de redirection dans une page fait pointer cette page directement vers une autre page. La page redirigée est par conséquent invisible comme page puisque la redirection se déclenche automatiquement au moment de sa consultation. En pratique, cette fonctionnalité permet d'attribuer à un contenu textuel plusieurs entrées, et ainsi d'améliorer son accessibilité dans la recherche par titres.

Pour déclarer un lien de redirection dans une page, nous utilisons la syntaxe suivante `#REDIRECTION [[Article vers lequel le lien pointe]]`. Elle peut être analysée comme la composition d'un marqueur de redirection `#REDIRECTION` et la déclaration d'un lien interne `[[Article vers lequel le lien pointe]]`.

### 3.3.3 Liens externes

Les liens externes sont les hyperliens qui pointent vers une autre page Web à l'extérieur de Wikipédia. Généralement pour déclarer un lien externe, nous mettons l'hyperlien en question dans une paire de crochets carrés `[]` comme `[http://www.wikimedia.fr]` ou `[http://www.wikimedia.fr Association Wikimédia France]` lorsqu'on veut personnaliser l'affichage du lien.

Dans l'actuel recherche, nous limitons l'interrogation des données à la portée de Wikipédia. Autrement dit, le caractère référentiel des liens externes ne sera pas pris en compte, les balises servant à cet effet serons enlevées dans la phase de nettoyage de texte.

### 3.3.4 Éléments de mise en forme

Nous présentons dans ce qui suit des éléments wikicode essentiellement esthétiques, ceux de la mise en forme du texte. L'objectif est de les considérer dans une perspective de reconstitution du texte brut à partir du texte mise en forme.

#### 3.3.4.1 Balises pré-interprétatives

Les balises pré-interprétatives autorisées par la syntaxe de wikicode prennent la forme suivante `<début du bloc> wikicode </fin du bloc>`, elles permettent de spécifier la manière dont on souhaite d'interpréter le texte. Par exemple, la syntaxe `<nowiki> wikicode </nowiki>` permet d'afficher le wikicode tel qu'il est, `<math> latexcode </math>` déclare un texte en syntaxe LaTeX, elle sera affichée comme une formule mathématique. Il existe également des balises qui permettent d'afficher une formule chimique, des lignes mélodiques, ou d'insérer une carte géographique.

#### 3.3.4.2 Modèle

L'utilisation des modèles permet de reproduire un texte préformaté et prédéfini soit par Wikipédia, soit par des rédacteurs (Wikipédia 2017f). Pour placer un modèle qui ne demande pas d'argument, il suffit de mettre le nom du modèle et les délimiter par une paire de doubles crochets comme `{{nom du modèle}}`. Pour les modèles plus complexes, il leur faut préciser des arguments en plus du nom du modèle. En général, les différents arguments suivent le nom du modèle dans la paire des doubles crochets, ils sont séparés par une barre vertical comme `{{Date|22|février|2008}}`.

Compte tenu de la variété des modèles et leurs fonctionnalités dans Wikipédia, il semble difficile de prendre en compte tous les modèles utilisées. Notre choix est de ne considérer que quelques modèles pertinents pour l'étude de définitions, d'ignorer par conséquent, l'information que véhicule d'autres modèles.

L'intitulé *infobox* (Wikipédia 2017c) désigne un type de modèles à champs multiples. Il se présente sous la forme d'un tableau sommaire, permet d'avoir une vue synthétique et récapitulative sur l'article. Il existe, pour le domaine de métier, un *infobox* « Modèle :Infobox Métier » qui est conçu pour la présentation de métier (Wikipédia 2017d). Le modèle d'*infobox* que nous retrouvons dans l'article concernant le menuisier est présenté dans la figure 2.

Dans la section 3.4 à propos de l'extraction de textes, c'est sur la présence d'un *infobox* Métier que les critères de présélection se base pour effectuer l'extraction, ensuite les liens internes voire d'autres marqueurs que nous présentons dans l'actuelle section vont ensuite entrer en jeu et nous aider à vérifier la pertinence des articles retenus, ou découvrir des articles pertinents sans *infobox* Métier.

#### 3.3.4.3 Autres syntaxes de mise en forme

Pour mettre en italique un texte, nous le plaçons dans une paire de doubles apostrophes comme

```
{{Infobox Métier
| appellations =
| image = Travail du bois.jpg
| taille image =
| légende = Un menuisier au travail
| secteur d'activité =
| compétences =
| perspectives =
| formation =
| métiers voisins =[[charpentier]] - [[Ébéniste]]
| code ROME = H2207
| code CNP =
| contraintes =
| horaires =
| pénibilité =
| risques =
}}
```

FIGURE 2 – Modèle :Infobox Métier sur l'exemple de menuisier

```
''texte mis en italique''
```

La syntaxe de mise en gras utilise une paire de triple apostrophes comme

```
'''texte mis en gras'''
```

Une paire de quintuple apostrophes sert à mettre à la fois en gras et en italique comme

```
'''''texte mis en italique et en gras'''''
```

Il existe quatre modèles particuliers permettant de modifier la taille de police, ils prennent la forme d'un modèle à argument unique. Pour augmenter la taille d'un texte, on utilise le modèle suivant

```
{{grossir|texte à grossir}}
```

; au contraire, pour diminuer la taille d'un texte, il consiste à utiliser le modèle suivant

```
{{Petit|texte dont on souhaite diminuer la taille}}
```

Pour la mise en indice et en exposant, les syntaxes suivantes sont respectivement

`{{ind|texte mis en indice}}`, `{{exp|texte mis en exposant}}`

Pour positionner un texte, on utilise principalement des balises positionnelles comme

```
<center>texte à centrer</center>
```

pour centrer un texte. On peut également mette en indice avec des balises positionnelles comme `<sub>texte mis en indice</sub>` pour une mise en indice, `<sup>texte mis en exposant</sup>` pour une mise en exposant.

Les listes peuvent être déclarées grâce aux `*` et `#` (WIKIPÉDIA 2017e). Par exemple,

```
* texte non-numéroté du niveau 1 d'une liste
```

met le texte au niveau 1 d'une liste non-numérotée,

```
** texte non-numéroté du niveau d'une liste
```

met le texte au niveau 2 d'une liste non-numérotée, ainsi de suite.

Pour créer une liste numérotée, on suit la même syntaxe en remplaçant le symbole astérisque `*` par une dièse `#`. Par exemple, pour placer un élément de niveau 2 dans une liste numérotée, la syntaxe[4] ressemble à

```
## texte numéroté du niveau 2 d'une liste.
```

Les commentaires qui ne sont pas à afficher sont délimités par `<!--` et `-->`. Nous mettons les commentaires dans entre les deux balises comme dans

```
<!-- un commentaire invisible à l'affichage. -->.
```

## 3.4 Repérage et extraction de définitions

Nous avons conçu un processus simple qui permet d'extraire des définitions à partir de la copie locale de Wikipédia. Cette méthode pratique consiste en 4 étapes suivantes : 1. Repérage des articles du domaine de métier 2. Nettoyage des éléments de mise en forme dans le texte 3. Repérage du premier paragraphe qui contient le titre de l'article en question 4. Extraction de la première phrase qui contient le titre et validation manuelle de celle-ci.

Le processus tout entier est implémenté en Python 2.7 avec la bibliothèque Python lxml qui permet d'analyser un fichier de format XML. Pour le repérage des articles du domaine, nous avons retenus dans un premier temps tous les articles dans lesquels un

4. La combinaison de la dièse et l'astérisque permet d'exprimer au sein d'une même liste des éléments numérotés et non-numérotés, voir en plus (WIKIPÉDIA 2017e).

modèle infobox intitulé *métier* est présent. Nous avons remarqué que la majorité des articles sélectionnés grâce à cette astuce sont également listées dans la page consacrée à la catégorie de *métier*. Cependant, les articles recensés par la catégorie de métier contient plus d'entités nommés qui désigne des personnalités du domaine de métier. Cela signifie plus de bruit et moins de précision pour le repérage. Grâce au simple repérage de la présence du modèle infobox *métier*, 765 articles ont été ainsi extraits du fichier XML de Wikipédia.

Le nettoyage des éléments de mise en forme consiste à enlever tous les éléments wikicode dans les articles retenus avec l'application de vingtaine de patrons de nettoyage que nous avons formulés en expressions régulières. Les patrons de nettoyage ont été conçus d'après notre étude sur la syntaxe de wikicode et que nous avons présenté dans les sous-sections précédentes.

Dans l'article de Wikipédia, nous avons constaté que pour la plupart du temps, la première proposition qui contient le titre, c'est-à-dire, l'entrée encyclopédique est une définition pour cette entrée. Pour cette raison, les étapes 3 et 4 ont été conçues dans le but d'extraire la première proposition définitoire dans laquelle le titre est mentionné tel qu'il est. 703 propositions sont ainsi validées à l'issue du processus, elles constituent le corpus intégral du domaine de métier.

# 4 Typologie de structures de définition

## 4.1 Introduction

### 4.1.1 Natures diverses des structures de définition

Il s'agit dans cette section de modéliser et formaliser les principales structures de définitions à partir de l'observation des données de notre corpus d'étude. Avant de commencer l'exposé, nous précisons ce que nous entendons par structures de définition du point de vue de la normalisation. Les structures de définitions que nous cherchons à modéliser dans ce mémoire sont des structures linguistiques qui permettent de définir ou d'énoncer une définition. Elles sont de natures diverses, à la fois syntaxiques, sémantiques et lexicales : elles sont syntaxiques puisque c'est en syntaxe que nous commençons à distinguer certaines définitions dont les structures syntaxiques ne sont pas directement comparables ; elles sont également sémantiques car sans rentrer dans l'analyse du sens des termes qui composent une définition, il est impossible de savoir si un genre prochain valide ou une différence spécifique valide est réellement présent dans un énoncé définitoire. Ces structures de définition sont également lexicales car la présence de certains éléments lexicaux est indispensable pour signaler le fait et la fonction de définir dans un certain type de définition, par exemple, des expressions comme *consister en* (activité professionnelle), *avoir pour profession* (activité professionnelle).

En conséquence, la typologie que nous allons présenter sur les principales structures de définitions dans notre corpus d'étude prend en compte ces trois niveaux linguistiques de définition. La perspective dans laquelle nous nous plaçons est celle de l'automatisation de l'analyse de définitions dans le but d'effectuer une normalisation de leurs structures à l'échelle. Or, l'automatisation nécessite une prise en compte de la structure des définitions dans sa totalité.

Pour faciliter la compréhension de nos lecteurs, nous décidons de baser notre exposé sur l'analyse d'exemples réels, issus de notre corpus d'étude. Pour pour mieux visualiser les différences entre ces structures de définitions, nous allons présenter des exemples annotés. Les annotations que nous avons réalisées sur exemples permettent de mettre en avant certaines propriétés syntaxiques, sémantiques ou lexicales liées aux structures de définitions.

### 4.1.2 Annotations syntaxique, sémantique et lexicale

Nous avons ci-dessous, par exemple, une définition simple du métier de *palefrenier* :

*Le* *palefrenier*$_{\text{M}}$ *est un* *employé*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *chargé de* *s' occuper des chevaux*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (8)

Dans la définition, des lexèmes ou des constituants syntaxiques utiles à notre discussion sont suivis par une notation qui se compose d'une lettre majuscule et d'une abréviation [5] des traits syntaxiques placés entre crochets comme $\text{X}_{[\text{NOM}]}$ qui représente *palefrenier*, $\text{Y}_{[\text{NOM}]}$ utilisé pour désigner *chevaux* dans l'exemple. Il s'agit de la partie syntaxique de l'annotation.

L'utilisation des encadrés de couleur associés aux lettres majuscules M, G et A est réservée à l'annotation des propriétés sémantiques des structures de définitions.

Le mot vedette, c'est-à-dire le terme à définir est encadré en vert et suivi de l'indice M. De la même manière, le genre prochain est indiqué par un cadre rouge suivi de l'indice G. La présence d'une différence spécifique, ou d'un trait de définition dénotant activité principale du métier est quant à elle signalée par un cadre bleu suivi de l'indice A. Par soucis de cohérence, l'annotation syntaxique des éléments présents à l'intérieur d'un cadre conserve la couleur de ce cadre.

En dernier lieu, les éléments lexicaux dont le paradigme est relativement restreints par rapport à un type de définition, sont encadrés de filet gris. Par exemple, *chargé de* qui sert du connecteur entre G et A, qui est un groupe lexical caractéristique d'un type de définition selon notre typologie.

### 4.1.3 Types majeurs de structures de définitions

Il nous semble important de préciser le processus par lequel nous sommes amenés à proposer les 5 principaux types de définitions, avant l'analyse même de ceux-ci à travers l'examen des modèles qui leur correspondent. Nous présentons ci-dessous 15 définitions partiellement annotées pour commencer :

*Un* *greffier*$_{\text{M}}$ *est un* *officier* *de* *justice*$_{\text{G}}$. (1)

5. L'ensemble des traits syntaxiques utilisés dans ce mémoire se trouve dans liste de symboles au début de ce document.

*Un* *procureur général*$_{M}$ *est une* *personne* *de* *droit*$_{G}$. (2)

*métreur*$_{M}$ *est un* *métier*$_{G}$ *du* *bâtiment*$_{A}$. (3)

*Le* *frigoriste*$_{M}$ *est un* *technicien*$_{G}$ *du* *froid*$_{A}$. (4)

*Le* *lardonnier*$_{M}$ *est un* *préparateur*$_{G}$ *de* *lardons*$_{A}$. (5)

*Un* *patenôtrier*$_{M}$ *est un* *fabricant*$_{G}$ *des* *chapelets*$_{A}$. (6)

*Le* *pâtissier*$_{M}$ *est un* *artisan*$_{G}$ *spécialisé dans* *la fabrication des pâtisseries*$_{A}$. (7)

*Le* *palefrenier*$_{M}$ *est un* *employé*$_{G}$ *chargé de* *s' occuper des chevaux*$_{A}$. (8)

*Le* *wagonnier*$_{M}$ *est un* *cheminot*$_{G}$ *chargé de* *la manœuvre des wagons*$_{A}$. (9)

*Un* *joquetier*$_{M}$ *est un* *domestique*$_{G}$ *conduisant une voiture en postillon*$_{A}$. (10)

*Un* *corailleur*$_{M}$ *est une* *personne*$_{G}$ *qui* *pêche le corail afin de le revendre*$_{A}$. (11)

*aviculteur*$_{M}$ *est le* *métier* *qui consiste en* *l'élevage d'oiseaux ou de volailles*$_{A}$. (12)

*porteur d'eau*$_{M}$ *est un* *métier / activité* *qui consiste à* *transporter de l'eau*$_{A}$. (13)

*Le* *marchand de tapis*$_{M}$ *a pour profession* *le commerce des tapis*$_{A}$. (14)

*Un* *agriculteur multiplicateur*$_{M}$ *a pour activité* *[...]* *la multiplication des semences*$_{A}$. (15)

Dans les exemples ci-dessus, les définitions (1) - (13) sont des propositions construites avec la copule *être*. En revanche, les 2 dernières définitions ont pour verbe principal *avoir*. Nous effectuons une première distinction : nous appelons les premières *définitions à structure copulative* et les secondes *définitions à structure non copulative*. Parmi les définitions à structure copulative, les définitions (1) - (9) sont des propositions simples, tandis que les définitions (10) - (13) sont des définitions complexes ou assimilables à une proposition complexe comme (10).

Dans un second temps, nous observons les définitions (1) - (6), on constate des différences sur la présence et la modalité du déterminant sur le mot vedette entre (3) et (4), la présence et la modalité du déterminant sur la différence spécifique présumée entre (1) et (4). Les définitions (1) - (6) présentent ainsi des différences, mais elles ont également un élément en commun : leur description est construite avec deux substantifs enchaînés à l'aide de la préposition *de*. Nous proposons donc de les classer dans un premier type intitulé type 1.

Ensuite, nous observons les définitions (7) - (9), on voit que le marqueur de relation qui permet de relier le genre prochain présumé et une précision sur l'activité concernée prend

la forme d'un adjectif suivi d'une préposition régie comme *spécialisé dans*, *chargé de*. L'activité peut s'exprimer aussi bien par un groupe nominal comme *la fabrication des pâtisseries* dans (7) ou *la manœuvre des wagons* dans (9), que par un groupe verbal *s' occuper des chevaux* comme dans (8), nous les classons, dans un premier temps, dans un second type ou le type 2.

Les définitions (10) et (11) ont pour point commun une subordonnée qui emploie un verbe d'action pour caractériser l'activité principale du métier à définir, comme *conduire* pour (10) et *pêcher* pour (11). Nous leur consacrons ainsi le troisième type de définitions ou le type 3.

Les définitions (12) et (13) concernent également des propositions complexes. Cependant l'utilisation des marqueurs de relation comme *qui consiste en* et *qui consiste à* permet de définir le métier par l'activité concernée sans pour autant mentionner un genre prochain comme *technicien* *artisan* *domestique* dans (4) (7) (10), par exemple. Nous leur consacrons ainsi un quatrième type de définitions ou le type 4. Les définitions (14) et (15) emploient l'expression *N avoir N pour N*. Nous les classons dans un cinquième type de définitions ou le type 5.

Par un processus similaire à celui que nous venons de montrer, nous avons relevé ces 5 types et modèles lexico-syntaxiques de définitions, à partir des 100 définitions les plus courtes du corpus intégral. Comme résumé dans la figure 3, ces modèles sont réunis en 3 groupes distinctifs pour être étudiés séparément : le premier groupe est constitué des 2 modèles qui possèdent une construction en proposition simple [6] dont le verbe principal est une copule, le deuxième groupe contient 2 modèle utilisant une construction à proposition complexe dont le verbe principal est aussi une copule. Un autre modèle est placé dans le dernier groupe qui rassemble des modèles de définitions à tête verbale mais non-copulative. Par ailleurs, au sein d'un même modèle qui correspond à un type principal, nous différencions des sous-types qui correspondent au même schéma, mais se distinguent par la contrainte imposée aux traits syntaxiques d'un élément lexical ou par des différences que nous jugeons secondaires.

6. Cette distinction n'exclut pas que les modèles puissent apparaître dans une proposition complexe ou former une définition avec un autre type de modèle présenté ici.

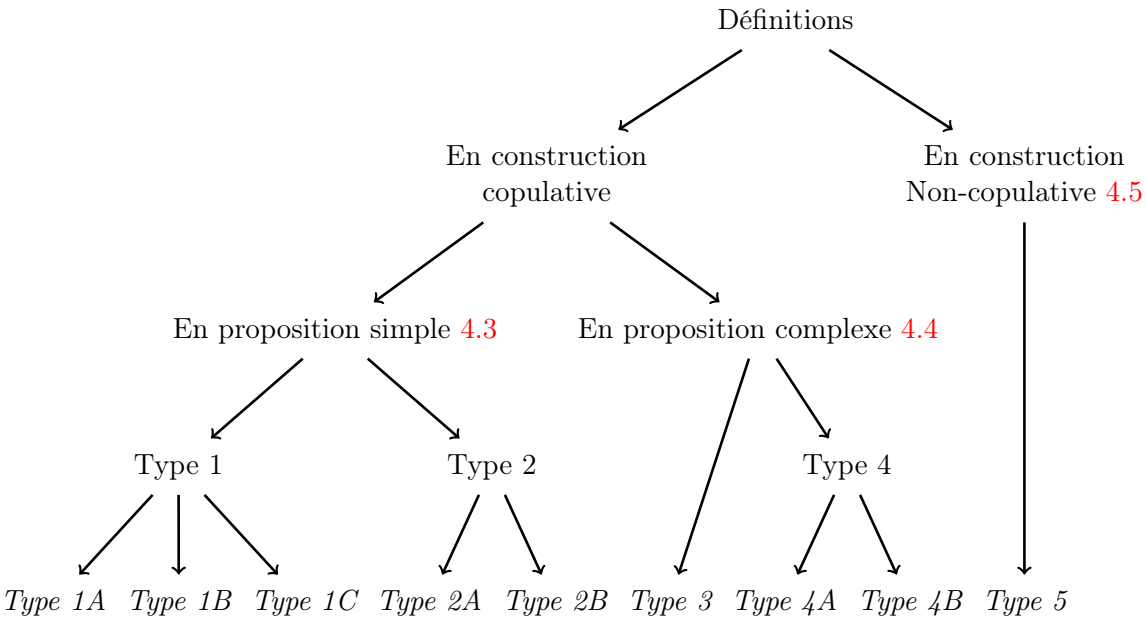


FIGURE 3 – Arbre typologique de modèles de définitions présentés par la section 4

## 4.2 Convention de notation sur les graphes

### 4.2.1 Notation sur la structure syntaxique

Nous employons la structure syntaxique pour représenter un modèle de définition. Une structure syntaxique de surface est un graphe orienté, acyclique et connexe. Elle sert à représenter le niveau syntaxique d'une proposition. Sur le graphe, la place de chaque nœud est occupée par un élément lexical, les arêtes représentent chacune une relation syntaxique entre deux éléments lexicaux. La direction des flèches entre deux éléments va du dominant au dominé. La dominance syntaxique d'un élément sur l'autre signifie que le dominant légitime la présence de son dominé, c'est bien cette relation de dominance ou de dépendance qui permet d'hiérarchiser une proposition et la représenter ensuite sous forme d'un graphe syntaxique.

**4.2.1.1 Graphe de la structure syntaxique de surface** Pour exprimer un élément lexical, nous employons deux types de notation : le premier type consiste à représenter un élément lexical par son lexème. Un lexème est l'ensemble des signifiants dont les signifiés ne se distinguent que par des significations flexionnelles (MEL'ČUK et MILIĆEVIĆ 2014 : 250). Comme convention dans ce mémoire, les lexèmes sont mis en majuscules suivis facultativement par le nom en abréviation des traits syntaxiques, qui sont, quant à

eux, mis en indice. Par ailleurs, des traits syntaxiques en abréviation sont présentés avec leurs noms complets dans la liste des symboles que nous pouvons retrouver au début du présent document. À titre d'exemple, nous représentons une copule conjuguée au présent à l'indicatif, d'après la notation qu'on vient de présenter, comme $\hat{\text{ÊTRE}}_{[\text{IND}][\text{PRÉS}]}$.

Le deuxième type de notation, quant à lui, fournit la possibilité de désigner un élément lexical inconnu. Il s'écrit en une seule lettre alphabétique suivi des indications concernant des traits syntaxiques, également mises en indice. Par exemple, $\text{X}_{[\text{NOM}]}$ représente un élément nominal inconnu, et $\text{Y}_{[\text{GV}]}$ un groupe verbal inconnu. Cette dernière notation sert principalement à exprimer un élément lexical dont on ne connaît que les traits syntaxiques.

Par ailleurs, si un élément lexical est mis en parenthèses, cela signifie que dans la structure, sa présence est facultative ; autrement dit, il peut être omis. Lorsqu'il est absent, la relation syntaxique qui lui est associée sera à ignorer, de même pour tous ses descendants. Si plusieurs lexèmes sont mis en crochets à la place d'un nœud du graphe syntaxique, cela signifie que pour la structure syntaxique, tous ces lexèmes sont considérés comme admissibles.

Plus spécialement pour la définition, nous précédons le mot vedette, le terme à définit, par une dièse # afin de marquer sa position et de fournir un point de repère dans le graphe pour nos lecteurs.

#### 4.2.1.2 Arguments et paramètres du modèle

Nous notons simplement que les arguments avec les éléments lexicaux connus sont les éléments ayant une contribution sémantique importante par rapport à la construction d'une définition au regard de son modèle. Par contre, les paramètres sont pour la plupart des éléments plus syntaxiques que sémantiques. Ils servent à relier des éléments lexicaux connus et arguments d'après les règles syntaxiques.

Afin de faciliter la présentation, nous regroupons les éléments lexicaux inconnus d'un modèle qui commencent par une lettre en majuscule en tuple comme ($\text{X}_{[\text{NOM}]}$, $\text{Y}_{[\text{GV}]}$) que nous appelons les arguments du modèle. De la même manière, on peut regrouper des éléments lexicaux inconnus qui commencent par une lettre grec, comme ($\alpha_{[\text{ART}]}$, $\beta_{[\text{ART}]}$), nous proposons de les appeler les paramètres du modèle.

### 4.2.2 Notation sur la structure sémantique

Nous utilisons les guillemets simples pour dénoter le sens d'un élément lexical. Par exemple, 'métier' signifie le sens du lexème *métier*. Cette notation simplifie nos discussions en sémantique sur des éléments lexicaux et les diverses relations qu'il peut y avoir du point de vue sémantique.

## 4.3 Définitions à structure copulative en proposition simple

### 4.3.1 Type 1

Le modèle de définitions du type 1 concerne 30 définitions sur 100 (30%) du corpus d'étude. Il apparaît 7 fois comme la structure de la proposition principale de définition, 23 fois comme la structure complète de définition. Illustrons maintenant ce type de définitions en reprenant les exemples (1) - (6) :

#### 4.3.1.1 Reconnaissance du type

*Un* *greffier*$_{\text{M}}$ *est un* *officier*$_{\text{X}_{[\text{NOM}]}}$ *de justice*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{G}}$. (1)

*Un* *procureur général*$_{\text{M}}$ *est une* *personne*$_{\text{X}_{[\text{NOM}]}}$ *de droit*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{G}}$. (2)

*métreur*$_{\text{M}}$ *est un* *métier*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *du* *bâtiment*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (3)

*Le* *frigoriste*$_{\text{M}}$ *est un* *technicien*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *du* *froid*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (4)

*Le* *lardonnier*$_{\text{M}}$ *est un* *préparateur*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *de* *lardons*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (5)

*Un* *patenôtrier*$_{\text{M}}$ *est un* *fabricant*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *des* *chapelets*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (6)

Comme nous l'avons mentionné dans l'introduction de la section 4, les critères qui permettent de distinguer le type 1 des autres types de définitions sont d'abord syntaxiques, nous énumérons les points communs des exemples ci-dessus comme critères :

1. L'énoncé en question doit être réalisé en proposition simple
2. La tête syntaxique doit être la couple *être* au présent à l'indicatif
3. Le definiens, la zone qui sert à définir doit suivre le schéma suivant :

$$\texttt{X}_{\texttt{[NOM]}}\ \texttt{\{de, du, des, de la\}}\ \texttt{Y}_{\texttt{[NOM]}}$$

Par ailleurs, pour une reconnaissance plus fine et robuste des définitions du type 1, nous proposons de les modéliser par leur structure syntaxique commune. La structure syntaxique est représentée dans la figure 4 ci-dessous.

Le schéma dans la figure 4 représente un type de proposition simple construite avec une copule au présent à l'indicatif étant la tête syntaxique. Le mot vedette #M, c'est-à-dire, le terme à définir, est connecté avec la copule par la relation subjectale. #M peut posséder un déterminant facultative ($\alpha_{[\text{ART}]}$). L'autre actant syntaxique de la tête

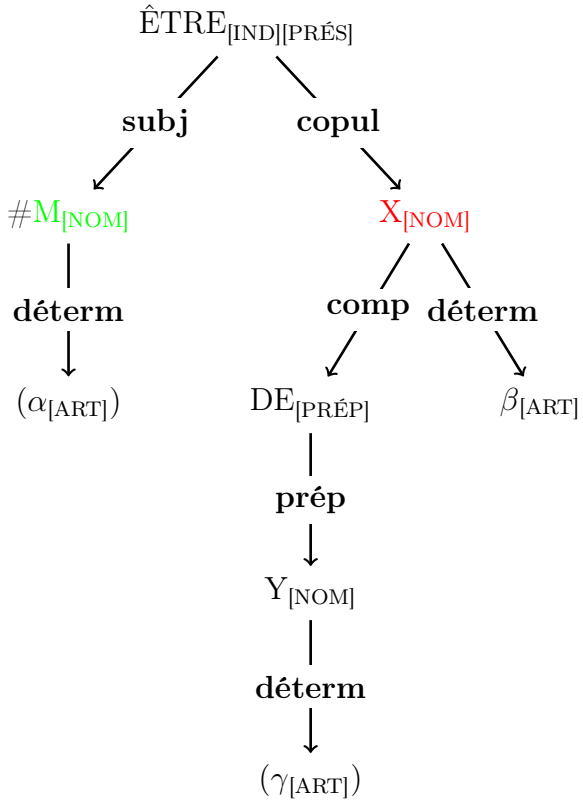


FIGURE 4 – Structures syntaxique correspondant aux définitions du type 1

domine directement un élément nominal inconnu $X_{[NOM]}$ par la relation copulaire. $X_{[NOM]}$ domine à son tour la préposition ($DE_{[PRÉP]}$). Nous soulignons que la préposition DE dans le modèle dépend du $X_{[NOM]}$ par la relation complétive. Autrement dit, il est un complément de $X_{[NOM]}$. Il permet d'introduire un autre groupe nominal inconnu qui est $Y_{[NOM]}$. Pour inclure une définition plus basique du type $\alpha_{[ART]}$ #M $ÊTRE_{[IND][PRÉS]}$ $X_{[NOM]}$ dans notre modèle, nous avons rendu la préposition ($DE_{[PRÉP]}$) optionnel. En son absence, ses dépendants sont également considérés comme non fournis. À son tour, $Y_{[NOM]}$ peut posséder un déterminant facultatif qu'on note par ($\gamma_{[ART]}$).

#### 4.3.1.2 Reconnaissance des sous-types

Dans un second temps, nous établissons les critères permettant de distinguer les trois sous-types au sein des définitions du type 1. Bien que cette distinction de sous-types paraissent secondaire, c'est à ce niveau de distinction, nous sommes en mesure d'identifier et valider le vrai genre prochain et la vraie différence spécifique à partir d'une définition du type 1.

**Critères pour le type 1A** Les définitions (1) et (2) sont reproduites ci-dessous pour accompagner la présentation des critères de reconnaissance pour le type 1A :

*Un* *greffier*$_{\text{M}}$ *est un* *officier*$_{\text{X}_{[\text{NOM}]}}$ *de justice*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{G}}$. (1)

*Un* *procureur général*$_{\text{M}}$ *est une* *personne*$_{\text{X}_{[\text{NOM}]}}$ *de droit*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{G}}$. (2)

Nous consacrons le type 1A aux définitions (1) et (2), les critères qui permettent de les distinguer des définitions (3) - (6) sont :

1. $\text{X}_{[\text{NOM}]}$ dénote une personne ou un métier
2. $\text{X}_{[\text{NOM}]}$ n'est pas un dérivé sémantique qui signifie 'celui qui fait' comme *préparateur* qui signifie 'la personne qui prépare quelque chose' dans (5), *fabricant* voulant dire 'la personne qui fabrique quelque chose' dans (6).
3. $\text{Y}_{[\text{NOM}]}$ ne possède pas de déterminant

**Critères pour le type 1B** Les définitions (3) et (4) sont reproduites ci-dessous pour accompagner la présentation des critères de reconnaissance pour le type 1B :

*métreur*$_{\text{M}}$ *est un* *métier*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *du* *bâtiment*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (3)

*Le* *frigoriste*$_{\text{M}}$ *est un* *technicien*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *du* *froid*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (4)

Nous consacrons le type 1B aux définitions (3) et (4), les critères qui permettent de les distinguer des définitions (1) (2) (5) (6) sont :

1. $\text{X}_{[\text{NOM}]}$ dénote une personne ou un métier
2. $\text{X}_{[\text{NOM}]}$ n'est pas un dérivé sémantique qui signifie 'celui qui fait'. Autrement dit, aucun verbe V ne vérifie $S_1(\text{V}) = \text{X}_{[\text{NOM}]}$
3. $\text{Y}_{[\text{NOM}]}$ possède un déterminant indéfini

**Critères pour le type 1C** Les définitions (5) et (6) sont reproduites ci-dessous pour accompagner la présentation des critères de reconnaissance pour le type 1C :

*Le* *lardonnier*$_{\text{M}}$ *est un* *préparateur*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *de* *lardons*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (5)

*Un* *patenôtrier*$_{\text{M}}$ *est un* *fabricant*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *des* *chapelets*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (6)

Nous consacrons le type 1B aux définitions (5) et (6), les critères qui permettent de les distinguer des définitions (1) - (4) sont :

1. $X_{[NOM]}$ dénote une personne ou un métier
2. $X_{[NOM]}$ est un dérivé sémantique vérifiant $S_1(V) = X_{[NOM]}$ pour un certain verbe V

**Discussion sur la relation sémantique entre '$X_{[NOM]}$' et '$Y_{[NOM]}$'** Il s'agit de distinguer deux relations sémantiques entre '$X_{[NOM]}$' et '$Y_{[NOM]}$' dans le modèle, qui sont deux relations utiles à la transformation des définitions dans un autre type.

D'après la définition (6) tirée du corpus d'étude, un $X_{[NOM]}$ dont la valeur est *fabricant* est un dérivé sémantique nominal actantiel du verbe *fabriquer*. Cette relation peut être exprimée par la fonction lexicale '$S_1$' qui 'signifie celui qui fait' (L'Homme 1998 ; Verlinde et al. 2004) comme $S_1(fabriquer) = fabricant$.

Dans le cas du type 1C où $X_{[NOM]}$ est un dérivé sémantique nominal actantiel, nous observons que c'est '$X_{[NOM]}$' qui domine sémantiquement sur '$Y_{[NOM]}$' par son deuxième actant. Cette relation signifie souvent que '$Y_{[NOM]}$' est un objet direct du verbe qui correspond à '$X_{[NOM]}$'. Dans notre définition (6), '$Y_{[NOM]}$' = 'Chapelets', '$X_{[NOM]}$' = 'fabricant' impliquent que '$X_{[NOM]}$ fabrique les chapelets'. Dans le cas du type 1A et 1B, c'est '$Y_{[NOM]}$' qui effectue une prédication sur $X_{[NOM]}$ en tant que son modifieur. Nous présentons ces deux relations sémantique entre $X_{[NOM]}$ et $Y_{[NOM]}$ dans les figures 5 et 6 en montrant les structures sémantiques qui leur correspondent.

'$X_{[NOM]}$' ← **1** — '$Y_{[NOM]}$'

Figure 5 – Structure sémantique des définitions du type 1A, 1B

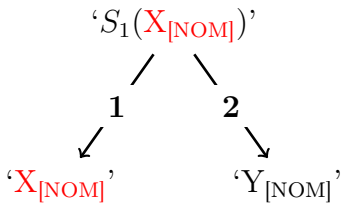


Figure 6 – Structure sémantique des définitions du type 1C

#### 4.3.1.3 Type 1A

*Un* *greffier*$_{M}$ *est un* *officier*$_{X_{[NOM]}}$ *de justice*$_{Y_{[NOM]}}$$_{G}$. (1)

*Un* *procureur général*$_{M}$ *est une* *personne*$_{X_{[NOM]}}$ *de droit*$_{Y_{[NOM]}}$$_{G}$. (2)

**Identification du genre prochain présumé** Dans les définitions du type 1A, la partie $X_{[NOM]}$ *de* $Y_{[NOM]}$ de l'énoncé est choisie comme genre prochain présumé. Ce choix est motivé par une comparaison avec les définitions du type 1B : la présence du déterminant indéfini sur $Y_{[NOM]}$ dans les définitions du type 1B lui donne plus d'autonomie en sémantique que dans les définitions du type 1A où le déterminant de $Y_{[NOM]}$ est absent. Cette différence au niveau de dégrée d'autonomie sémantique de $Y_{[NOM]}$ nous a incité à considérer l'ensemble de $X_{[NOM]}$ *de* $Y_{[NOM]}$ comme un genre prochain intégral pour le type 1A. Au contraire, pour le type 1B, nous considérons l'élément lexical nominal $Y_{[NOM]}$ comme faisant partie de la différence spécifique.

Pour noter cette décision, nous posons ci-dessous que pour le type 1, le genre prochain présumé G est identifié par

$$G = X_{[NOM]} \textit{ de } Y_{[NOM]}$$

**Validation du vrai genre prochain** Maintenant, nous cherchons à savoir dans quelles conditions le genre prochain présumé G est réellement valide. De notre point de vue, il faut d'abord que la connaissance sur le genre prochain présumé d'une définition du type 1 permette de trouver une classe sémantique plus restreinte que celle de métier, le domaine déjà connu et choisi avant la constitution de notre corpus d'étude. L'identification de la classe sémantique à laquelle le mot vedette $X_{[NOM]}$ appartient peut s'appuyer aussi bien sur l'élément $X_{[NOM]}$ que sur $Y_{[NOM]}$ : dans la définition (1), $X_{[NOM]} = \textit{officier}$ désigne une sous-catégorie du métier, $Y_{[NOM]} = \textit{justice}$ contribue à cerner cette sous-catégorie sur la spécialisation en justice ; dans la définition (2), $X_{[NOM]} = \textit{personne}$ ne dit pas plus qu'un *professionnel*, donc il ne constitue pas un vrai genre prochain à lui seul. Nous aurions pu croire que cette définition décrivait des bénévoles travaillant occasionnellement pour la justice, par exemple. Néanmoins, $Y_{[NOM]} = \textit{droit}$ dans $\boxed{\textit{personne}_{X_{[NOM]}} \textit{ de } \textit{droit}_{Y_{[NOM]}}}_{G}$ nous renseigne sur un domaine de spécialité comme le droit. Pour le type 1A, nous considérons, malgré tout, $\boxed{\textit{personne}_{X_{[NOM]}} \textit{ de } \textit{droit}_{Y_{[NOM]}}}_{G}$ comme un vrai genre prochain car toutes les définitions que nous avons obtenues portent soit sur un professionnel soit sur un métier, et qu'avec ce prérequis, la présence du terme *droit* ne fait que rapprocher $\boxed{\textit{personne}_{X_{[NOM]}} \textit{ de } \textit{droit}_{Y_{[NOM]}}}_{G}$ de $\boxed{\textit{procureur général}}_{M}$.

#### 4.3.1.4 Type 1B

*métreur*$_{\text{M}}$ *est un* *métier*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *du* *bâtiment*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (3)

*Le* *frigoriste*$_{\text{M}}$ *est un* *technicien*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *du* *froid*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (4)

**Identification du genre prochain présumé** Contrairement aux définitions du type 1A, comme nous l'avons expliqué dans le paragraphe précédent, le genre prochain présumé est constitué d'un seul élément lexical, à savoir, $\text{X}_{[\text{NOM}]}$.

$$\text{G} = \text{X}_{[\text{NOM}]}$$

**Validation du vrai genre prochain** Les conditions de validation du genre prochain reposent sur l'élément lexical nominal $\text{X}_{[\text{NOM}]}$. Pour valider $\text{X}_{[\text{NOM}]}$ comme un vrai genre prochain, il faut que $\text{X}_{[\text{NOM}]}$ soit effectivement un terme de métier. Cette vérification peut se faire avec l'aide de ressources lexicales.

**Identification de la différence spécifique** Par réduction, nous posons $\text{Y}_{[\text{NOM}]}$ comme une différence spécifique présumée. Notez ce choix par simplicité comme ci-dessous :

$$\text{A} = \text{Y}_{[\text{NOM}]}$$

**Évaluation de la différence spécifique** D'abord, nous soulignons avoir besoin d'une différence spécifique qui décrit explicitement l'activité principale du métier pour pouvoir être validée comme une différence spécifique. Cette décision est motivée par la forme souvent verbale des descriptions, que nous pouvons facilement transformer en utilisant d'autres parties du discours. Ce type de différence spécifique semble avantageux au regard de la normalisation. Nous formulons ci-dessous cette condition :

$\text{A}^*$ = description assez spécifique sur l'activité principale du métier désigné par M

En conséquence, $\text{Y}_{[\text{NOM}]}$ = *froid* dans (3) et $\text{Y}_{[\text{NOM}]}$ = *bâtiment* dans (4) ne sont pas des différences suffisamment spécifiques, donc non-valides.

#### 4.3.1.5 Type 1C

*Le* *lardonnier*$_{\text{M}}$ *est un* *préparateur*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *de* *lardons*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (5)

*Un* *patenôtrier*$_{\text{M}}$ *est un* *fabricant*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *des* *chapelets*$_{\text{Y}_{[\text{NOM}]}}$$_{\text{A}}$. (6)

**Identification du genre prochain présumé** Bien que pour le type 1C, $X_{[NOM]}$ et $Y_{[NOM]}$ entretiennent une relation sémantique actantielle importante, comme décrit dans la figure 6, nous choisissons de considérer $X_{[NOM]}$ comme un genre prochain présumé car nous préférons considérons, par exemple, $X_{[NOM]}$ = *préparateur* comme un genre, et $Y_{[NOM]}$ = *lardons* comme une différence spécifique. Autrement dit, nous soutenons que deux professions qui consistent à préparer quelque chose sont similaires, que la différence sur ce que ces deux professions préparent est une différence spécifique.

$$G = X_{[NOM]}$$

**Validation du vrai genre prochain** La validation du genre prochain pour le type 1C repose sur le sens verbal dont '$X_{[NOM]}$' découle, noté par '$V_0(X_{[NOM]})$'. Pour le domaine de métier, $V_0(X_{[NOM]})$ doit être un verbe d'action qui dénote l'activité principale du métier, ou tout au moins, une activité professionnelle ou susceptible de l'être pour que $X_{[NOM]}$ soit un genre prochain valide.

$$G^* = X_{[NOM]}$$

si $V_0(X_{[NOM]})$ est un verbe qui dénote une activité professionnelle ou susceptible de l'être

**Identification de la différence spécifique** Dans la définition (5), $Y_{[NOM]}$ = *lardons* n'est pas suffisamment précis pour caractériser l'activité du métier de $\boxed{\textit{lardonnier}}_{M}$. Au contraire, $\boxed{\boxed{\textit{préparateur}_{X_{[NOM]}}}_{G}\ \textit{de}\ \textit{lardons}_{Y_{[NOM]}}}_{A}$ ou *préparation de lardons* semble plus approprié pour être choisi comme la différence spécifique. Le constat est le même pour la définition (6). Nous considérons ainsi $X_{[NOM]}$ {de, des, du, de la} $Y_{[NOM]}$ comme la différence spécifique présumée pour les définitions du type 1C.

$$A = X_{[NOM]}\ \{de,\ des,\ du,\ de\ la\}\ Y_{[NOM]}$$

**Évaluation de la différence spécifique** L'évaluation de la différence spécifique pour le type 1C se fait avec le genre prochain présumé. Ce que nous cherchons à évaluer, c'est la proposition verbale sous-entendue par l'ensemble de $X_{[NOM]}$ {de, des, du, de la} $Y_{[NOM]}$. Si cette proposition décrit une activité professionnelle suffisamment précise, nous considérons la différence spécifique $X_{[NOM]}$ {*de, des, du, de la*} $Y_{[NOM]}$ présumée comme une différence spécifique suffisante.

$$A^* = X_{[NOM]}\ \{de,\ des,\ du,\ de\ la\}\ Y_{[NOM]}$$

si $X_{[NOM]}\ \{de,\ des,\ du,\ de\ la\}\ Y_{[NOM]}$ contient une description d'une activité professionnelle ou susceptible de l'être

**4.3.1.6 *être animé* ou *activité* comme genre prochain** Une distinction supplémentaire doit également être faite sur deux sortes de genres prochains. Le premier consiste à dénoter des personnes professionnelles comme *technicien* dans (4), alors que le second désigne plutôt des professions comme métier dans (3). En effet ces deux genres prochains appartiennent à deux classes sémantiques distinctives : *être animé* et *activité*, ce qui signifie que ces termes ne sont pas directement comparables. Nous formalisons cette distinction en notant ci-dessous deux sortes de genre prochain valide marqués par un astérisque signifiant 'valide' :

$G^*_1 = X_{[NOM]}$ si $X_{[NOM]}$ est un véritable terme de métier qui dénote une entité animée
$G^*_2 = X_{[NOM]}$ si $X_{[NOM]}$ est un véritable terme de métier qui dénote un fait

### 4.3.2 Type 2

Ce type de modèle permet de représenter un grand nombre de définitions dans notre corpus d'étude. Nous en avons recensé 24 définitions sur 100 (24%).

*Le* *pâtissier*$_{\text{M}}$ *est un* *artisan*$_{\text{X}_{[\text{NOM}]}}$ G *spécialisé dans* *la fabrication des pâtisseries* $_{\text{Y}_{[\text{GN}]}}$ A. (7)

*Le* *palefrenier*$_{\text{M}}$ *est un* *employé*$_{\text{X}_{[\text{NOM}]}}$ G *chargé de* *s' occuper des chevaux* $_{\text{Y}_{[\text{GV}]}}$ A. (8)

*Le* *wagonnier*$_{\text{M}}$ *est un* *cheminot*$_{\text{X}_{[\text{NOM}]}}$ G *chargé de* *la manœuvre des wagons* $_{\text{Y}_{[\text{GN}]}}$ A. (9)

**4.3.2.1 Reconnaissance du type** La reconnaissance des définitions du type 2 repose sur un modèle lexico-syntaxique. En résumant les points communs des définitions (7) - (9), nous proposons les critères de reconnaissance suivants :

1. L'énoncé doit être réalisé en proposition simple
2. La tête syntaxique doit être la copule *être* au présent à l'indicatif
3. Le definiens suit le schéma syntaxique suivant

$$\texttt{X}_{\texttt{[NOM]}}\ \texttt{R}_{\texttt{[ADJ]}}\ \gamma_{\texttt{[PRÉP]}}\ \texttt{Y}_{\texttt{[GN]}}$$

ou

$$\texttt{X}_{\texttt{[NOM]}}\ \texttt{R}_{\texttt{[ADJ]}}\ \gamma_{\texttt{[PRÉP]}}\ \texttt{Y}_{\texttt{[GV][INFINI]}}$$

4. ($\text{R}_{[\text{ADJ}]}$, $\gamma_{[\text{PRÉP}]}$) doit faire partie des marqueurs de relation pertinents pour le domaine de métier. Par exemple, parmi ceux qui sont les plus fréquents dans notre corpus d'étude, nous pouvons mentionner

$$(\textit{chargé}, \textit{de}), (\textit{habilité}, \textit{à}), (\textit{responsable}, \textit{de}), (\textit{spécialisé}, \textit{dans})$$

**4.3.2.2 Reconnaissance des sous-types** Lorsque Y est un groupe nominal, nous utilisons la notation $\text{Y}_{[\text{GN}]}$ et obtenons le sous-type 2A. En revanche, si Y est un groupe Y nous le notons $\text{Y}_{[\text{GV}][\text{INFINI}]}$, dans ce cas-là, il s'agit du sous-type 2B.

Par la suite, nous présentons le modèle en structure syntaxique. Dans la figure 7, la structure de définitions possède, à la tête syntaxique, une copule $\text{ÊTRE}_{[\text{IND}][\text{PRÉS}]}$ au présent à l'indicatif. La copule gouverne le mot vedette #M par la relation subjectale, et le mot vedette possède un déterminant $\alpha_{[\text{ART}]}$ qui est obligatoirement présent. Un autre

actant de la copule ÊTRE$_{[IND][PRÉS]}$ connecte un groupe nominal inconnu, noté par X$_{[GN]}$. C'est à cet endroit que commence le definiens de la structure de la définition. Dans le definiens, X$_{[NOM]}$ est un élément lexical nominal qui gouverne un élément adjectival inconnu, marqué par R$_{[ADJ]}$. X$_{[NOM]}$ possède un déterminant obligatoire, notée par $\beta_{[PRÉP]}$. Comme nous l'avons mentionné, R$_{[ADJ]}$ par la relation objectale-oblique domine la préposition qui suit, notée par $\gamma_{[PRÉP]}$. $\gamma_{[PRÉP]}$ est donc une préposition régie. Cette préposition $\gamma_{[PRÉP]}$ sert soit à introduire un groupe nominal inconnu, Y$_{[GN]}$ comme dans la figure 7, soit un groupe verbal à l'infinitif comme Y$_{[GV][INFINI]}$ dans la figure 8. La seule différence entre les structures décrites dans les figures 7 et 8 réside dans les traits syntaxiques permis à l'élément lexical Y. Pour une raison de commodité, la structure syntaxique de la figure 7 représentée est appelée la structure syntaxique du type 2A de définitions. De la même façon, on appelle la structure syntaxique du type 2B celle représentée par la figure 8.

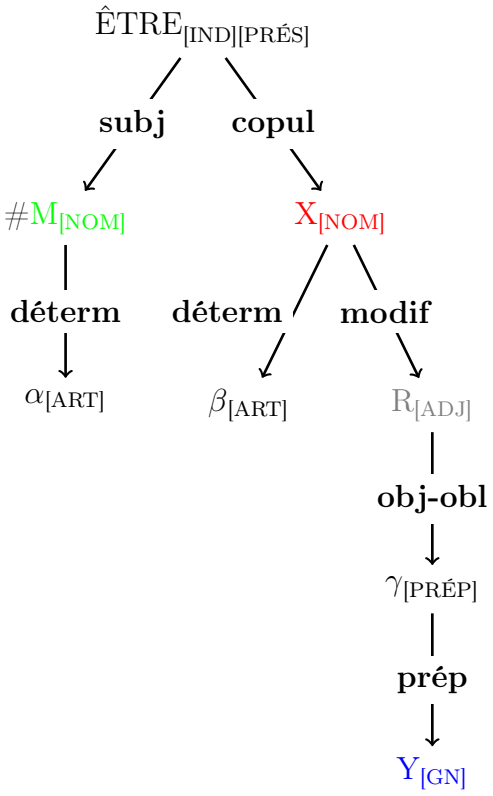


FIGURE 7 – Structure syntaxique correspondant aux définitions du type 2A

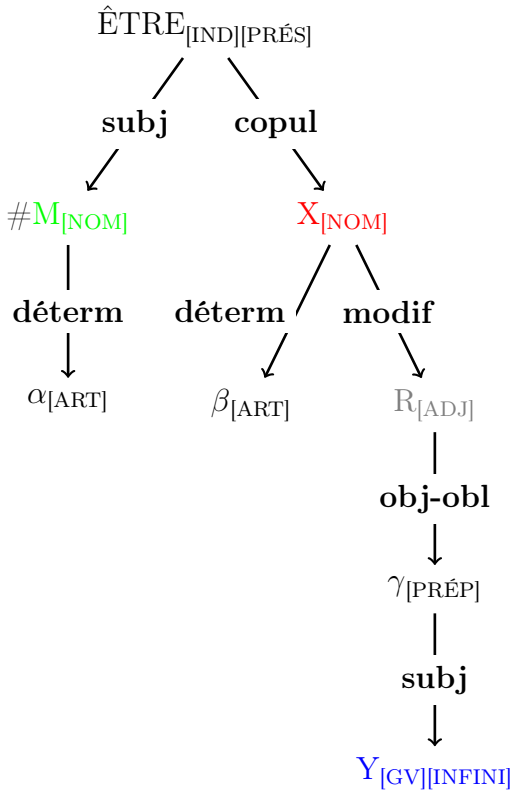


FIGURE 8 – Structure syntaxique correspondant aux définitions du type 2B

**4.3.2.3 Type 2A** Les définitions (7) (9) que nous classons dans le type 2A sont reproduites ci-dessous pour faciliter la lecture :

*Le* *pâtissier*$_{\text{M}}$ *est un* *artisan*$_{\text{X}_{[\text{NOM}]}}$ G *spécialisé dans* *la fabrication des pâtisseries* $_{\text{Y}_{[\text{GN}]}}$ A. (7)

*Le* *wagonnier*$_{\text{M}}$ *est un* *cheminot*$_{\text{X}_{[\text{NOM}]}}$ G *chargé de* *la manœuvre des wagons* $_{\text{Y}_{[\text{GN}]}}$ A. (9)

**Identification du genre prochain** Le genre prochain est un hyperonyme approximatif du mot vedette, par conséquent, il présente les mêmes traits syntaxiques que le mot vedette. Ici, pour le domaine de métier, le genre prochain doit être un substantif.

Dans les définitions (7) et (9), deux constituants nominaux peuvent être intéressants à considérer, à savoir, $\text{X}_{[\text{NOM}]}$ et $\text{Y}_{[\text{GN}]}$. Pour (7), $\text{X}_{[\text{NOM}]}$ = *artisan* est un genre prochain du métier *pâtissier*$_{\text{M}}$ ; alors, $\text{Y}_{[\text{GN}]}$ = *la fabrication des pâtisseries* caractérise l'activité du *pâtissier*. De même, dans la définition (9) pour le métier du *wagonnier*$_{\text{M}}$, $\text{X}_{[\text{NOM}]}$ = *cheminot* est un genre prochain pour le mot vedette, tandis que $\text{Y}_{[\text{GN}]}$ = *la manœuvre des wagons* en caractérise l'activité. On observe que le groupe nominal $\text{Y}_{[\text{GN}]}$ est introduit par un connecteur adjectival comme *spécialisé* (*dans*) ou *chargé* (*de*), et que l'ensemble du connecteur et de ce qui le suit forme un groupe adjectival qui s'ajoute à l'élément nominal $\text{X}_{[\text{NOM}]}$ comme un modifieur de $\text{X}_{[\text{NOM}]}$. Comme les connecteurs, ou marqueurs de relation des définitions du type 2 sont par définition, destinés à introduire une activité professionnel, et que la copule au présent à l'indicatif exprime une identité, c'est à $\text{X}_{[\text{NOM}]}$ de jouer le rôle du genre prochain présumé. Nous posons donc $\text{X}_{[\text{NOM}]}$ comme genre prochain présumé, comme noté ci-dessous :

$$\text{G} = \text{X}_{[\text{NOM}]}$$

**Validation du genre prochain** La validation du vrai genre prochain a besoin des ressources lexicales. Les ressources lexicales permettent d'identifier un terme de métier ou un terme qui dénote une certaine personne professionnelle. Au moment de validation, une attention particulière doit être portée à la distinction des termes qui dénotent des faits et ceux qui dénotent des entités animés, comme nous l'avons expliqué lors de la présentation des définitions du type 1. Ci-dessous, nous notons les genres prochains valides par $\text{G}^*$ auquel nous joignons une indice pour distinguer les vrais genres prochains qui appartiennent à deux classes sémantiques différentes voire disjointes :

$\text{G}^*_1 = \text{X}_{[\text{NOM}]}$ si $\text{X}_{[\text{NOM}]}$ est un véritable terme de métier qui dénote une entité animée

$G^*_2 = X_{[NOM]}$ si $X_{[NOM]}$ est un véritable terme de métier qui dénote un fait

**Identification de la différence spécifique** Les marqueurs de relations du type 2 introduisent $Y_{[GN]}$ comme un trait de définition qui renseigne l'activité du métier désigné par le mot vedette, nous posons ainsi que la différence spécifique présumée est $Y_{[GN]}$, comme noté ci-dessous :

$$A = Y_{[GN]}$$

**Évaluation de la différence spécifique** Comme dans la discussion portant sur l'évaluation de la différence spécifique pour les définitions du type 1, cette validation nécessite également des ressources lexicales. Des ressources pertinentes nous permettent de reconnaître un groupe nominal $Y_{[GN]}$ qui consiste en une description d'une activité professionnelle ou susceptible de l'être.

$$A^* = Y_{[GN]}$$

si $Y_{[GN]}$ recèle une description qui décrit une activité professionnelle ou susceptible de l'être

#### 4.3.2.4 Type 2B

*Le* *palefrenier*$_M$ *est un* *employé*$_{X_{[NOM]}}$$_G$ *chargé de* *s' occuper des chevaux* $_{Y_{[GV]}}$$_A$. (8)

**Identification du genre prochain** Comme les deux sous-types partagent la même structure à l'exception du trait syntaxique d'un élément terminal dans le graphe syntaxique $Y_{[GV]}$ qui fait partie d'un modifieur, l'identification du genre prochain du type 2B reste identique à celle du type 2A. Nous notons ci-dessous :

$$G = X_{[NOM]}$$

**Validation du genre prochain** Identiquement à celle du type 2A pour la même raison, nous notons les conditions de validation ci-dessous :

$G^*_1 = X_{[NOM]}$ si $X_{[NOM]}$ est un véritable terme de métier qui dénote une entité animée
$G^*_2 = X_{[NOM]}$ si $X_{[NOM]}$ est un véritable terme de métier qui dénote un fait

**Identification de la différence spécifique** La différence spécifique est introduite dans le type 2B grâce aux marqueurs de relation, elle prend la forme d'un groupe verbal, nous notons cette remarque ci-dessous :

$$A = Y_{[GV]}$$

**Évaluation de la différence spécifique** Comme pour les définitions du type 2A, la validation d'une vraie différence spécifique repose sur la pertinence de l'information que $Y_{[GN]}$ donne sur une activité professionnelle ou susceptible de l'être, nous notons ce critère ci-dessois :

$$A^* = Y_{[GV]}$$

si $Y_{[GV]}$ recèle une description qui décrit une activité professionnelle ou susceptible de l'être

## 4.4 Définitions à structure copulative en proposition complexe

### 4.4.1 Type 3

Dans notre corpus d'étude, nous avons recensé 22 définitions sur 100 (22%) qui peuvent être classées dans le type 3. Il s'agit des définitions construites en proposition complexe dont la principale possède une tête syntaxique étant la copule *être* au présent à l'indicatif, dont la subordonnée quant à elle est soit composée d'un groupe verbal présent de l'indicatif et du pronom relatif *qui*, soit réalisée avec un participe présent en emploi adominal.

*Un* *joquetier* M *est un* *domestique* $_{X_{[NOM]}}$ G *conduisant une voiture en postillon* $_{Y_{[GV]}}$ A. (10)

*Un* *corailleur* M *est une* *personne* $_{X_{[NOM]}}$ G *qui* *pêche le corail afin de le revendre* $_{Y_{[GV]}}$ A. (11)

**4.4.1.1 Reconnaissance du type** Nous caractérisons ci-dessous les propriétés syntaxiques des définitions (10) et (11), qui servent de critères de reconnaissance du modèle associé au type 3 :

1. L'énoncé doit être réalisé en proposition complexe
2. Le verbe principal de la proposition principale doit être la copule *être* au présent à l'indicatif
3. Le verbe principal de la proposition subordonnée doit être un verbe d'action au présent à l'indicatif
4. Le pronom relatif utilisé est *qui*. Il sert à renvoyer le sujet de la subordonnée à un élément nominal de la principale sur la base de la co-référence
5. L'emploi adnominal du participe présent est considéré comme équivalent à celui du pronom relatif *qui*

**Structure syntaxique** Ensuite, nous présentons le schéma de la structure syntaxique qui correspond aux définitions du type 3. Comme montré dans la figure 9, la tête de la proposition principale est une couple au présent à l'indicatif, il s'agit de l'élément lexical $\text{ÊTRE}_{[\text{IND}][\text{PRÉS}]}$. La copule gouverne le mot vedette M par la relation subjectale, elle domine par la relation copulaire l'élément lexical nominal inconnu $X_{[NOM]}$. Le mot vedette M possède un déterminant obligatoire, son déterminant est noté par $\alpha_{[ART]}$. $X_{[NOM]}$

possède également un déterminant obligatoire, nous le notons par $\beta_{[\text{ART}]}$. Quant à la proposition subordonnée, elle possède une structure simple : la tête doit être un groupe verbal au présent à l'indicatif, noté par $\text{Y}_{[\text{GV}][\text{IND}][\text{PRÉS}]}$. Elle domine un pronom relatif $\text{QUI}_{[\text{PRON}]}$, faisant référence à $\text{X}_{[\text{NOM}]}$ de la proposition principale.

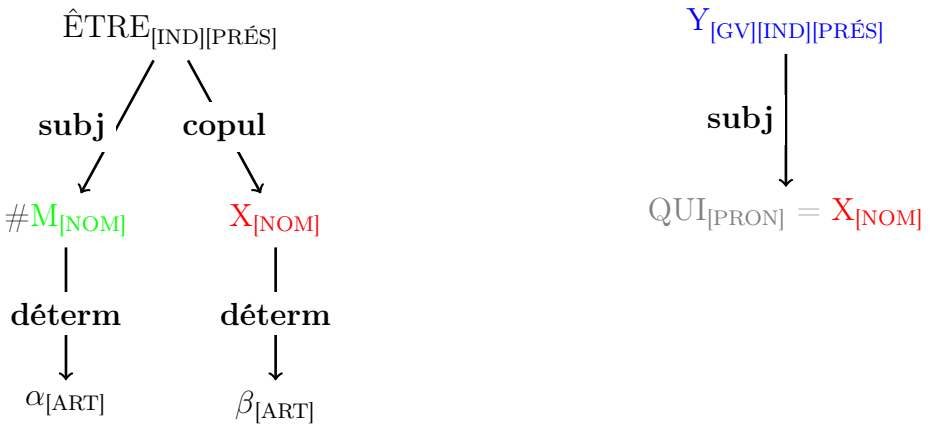


FIGURE 9 – Structure syntaxique correspondant aux définitions du type 3

**Identification du genre prochain** Les définitions (10) et (11) sont reproduites ci-dessous pour faciliter la lecture.

*Un* *joquetier*$_{\text{M}}$ *est un* *domestique*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *conduisant une voiture en postillon* $_{\text{Y}_{[\text{GV}]}}$$_{\text{A}}$.
(10)

*Un* *corailleur*$_{\text{M}}$ *est une* *personne*$_{\text{X}_{[\text{NOM}]}}$$_{\text{G}}$ *qui* *pêche le corail afin de le revendre* $_{\text{Y}_{[\text{GV}]}}$$_{\text{A}}$.
(11)

Comme la copule au présent à l'indicatif exprime une identité sous forme de « M *est un* X *qui* Y » et que les définitions du type 3 ont une subordonnée qui décrit l'activité sous forme verbale comme *conduire une voiture en postillon* dans (10) ou *pêcher le corail afin de le revendre* dans (11) ; par réduction, le genre prochain est $\text{X}_{[\text{NOM}]}$. Cette remarque est notée ci-dessous :

$$\text{G} = \text{X}_{[\text{NOM}]}$$

**Validation du genre prochain** On peut montrer que les conditions de validation sont identiques à celles du type 2. Notons simplement les conditions ci-dessous :

$\text{G}^*{}_1 = \text{X}_{[\text{NOM}]}$ si $\text{X}_{[\text{NOM}]}$ est un véritable terme de métier qui dénote une entité animée
$\text{G}^*{}_2 = \text{X}_{[\text{NOM}]}$ si $\text{X}_{[\text{NOM}]}$ est un véritable terme de métier qui dénote un fait

**Identification de la différence spécifique** Par définition, pour les définitions du type 3, si le pronom relatif *qui* est employé dans la construction, on identifie la différence spécifique présumée au groupe verbal de la subordonnée $Y_{[GV]}$ ; si la subordonnée est réalisée en participe présent, dans ce cas précis, on identifie la différence spécifique à $Y_{[GV]}$ de la proposition réalisée avec le pronom relatif, et qui lui est équivalente.

**Évaluation de la différence spécifique** La différence spécifique des définitions du type 3 est pré-validée puisque dans des critères de repérage des définitions du type 3, le groupe verbal doit dénoter un terme de métier en décrivant son activité principale.

### 4.4.2 Type 4

Puisque chacun des types 4, 5 et 6 ne représente que moins de 5 pourcent des définitions dont dispose notre corpus d'étude, ils sont des types minoritaires. Malgré la faible quantité de données, le type 4 des définitions que nous allons présenter ici fournit une variante importante vis-à-vis des définitions du type 3. Notamment, si le type 3 est privilégié par la définition des personnes de métier d'après notre données, le type 4 qui ressemble au type 3, est souvent associé à la définition des activités de métier.

**4.4.2.1 Reconnaissance du type** Pour commencer, nous reprenons ci-dessous les définitions (12)(13) :

*aviculteur*$_{\text{M}}$ *est le* *métier*$_{\text{X}_{[\text{NOM}]}}$ *qui consiste en* *l'élevage d'oiseaux ou de volailles*$_{\text{Y}_{[\text{GN}]}}$$_{\text{A}}$. (12)

*porteur d'eau*$_{\text{M}}$ *est un* *métier / activité*$_{\text{X}_{[\text{NOM}]}}$ *qui consiste à* *transporter de l'eau*$_{\text{Y}_{[\text{GV}]}}$$_{\text{A}}$. (13)

Comme on peut voir dans les définitions (12)(13), la structure syntaxique des définitions du type 4 paraissent similaires à celle des définitions du type 3, c'est la présence d'un marqueur de relation comme *consister en* ou *consister à* dans la subordonnée du type 4 qui permet de différencier les deux types, nous formulons ainsi les critères de reconnaissance ci-dessous :

1. L'énoncé doit être réalisé en proposition complexe
2. Le verbe principal de la proposition principale doit être la copule *être* au présent à l'indicatif
3. La proposition subordonnée doit contenir un des marqueurs de relations suivants : *consister en*, *consister à*
4. Le pronom relatif utilisé est *qui* servant à renvoyer le sujet de la subordonnée à un élément nominal de la principale sur la base de la co-référence
5. L'emploi adnominal du participe présent est considéré comme équivalent à celui du pronom relatif *qui*

**4.4.2.2 Reconnaissance des sous-types** La distinction des sous-types des définitions du type 4 est syntaxique.

**Critères de distinction entre les types 4A et 4B** Cette distinction ne porte que sur les traits syntaxiques prévus et autorisés pour l'élément lexical Y et sur le choix de la préposition qui permet de l'introduire dans le modèle : une définition du type 4 est classée dans le type 4A si son élément lexical Y est un groupe nominal introduit par la préposition *en*, noté ainsi par $Y_{[GN]}$ ; elle est classée dans le type 4B si Y est un groupe verbal introduit par la préposition *à*, dans ce cas, nous notons Y comme $Y_{[GV]}$.

**Structure syntaxique** Nous présentons ensuite les structures syntaxiques qui correspondent aux définitions du type 4. Au sein des définitions de ce type, nous distinguons deux structures syntaxiques légèrement différentes, nous proposons d'appeler les définitions qui correspondent à la première définitions du type 4A et à celles qui correspondent à la seconde définitions du type 4B. Leurs structures syntaxiques sont représentées respectivement dans les figures 10 et 11.

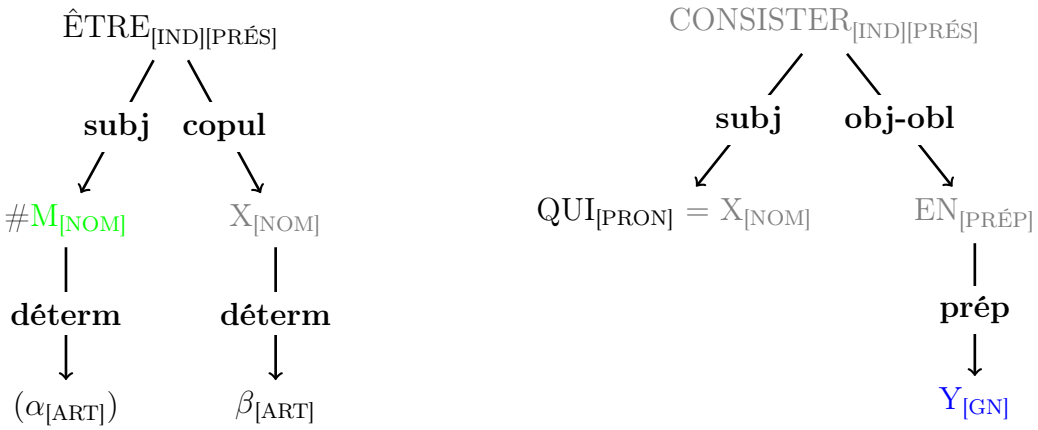


FIGURE 10 – Structures syntaxique correspondant aux définitions du type 4A

**4.4.2.3 Type 4A** La définition (12) du type 4A est reproduite ci-dessous pour faciliter la lecture.

**Identification du genre prochain**

*aviculteur*$_{M}$ *est le* *métier*$_{X_{[NOM]}}$ *qui consiste en* *l'élevage d'oiseaux ou de volailles* $_{Y_{[GN]}}$$_{A}$. (12)

Comme nous pouvons voir dans la définition (12), un vrai genre prochain est absent. D'après nos données, le paradigme permis au choix de l'élément lexical $X_{[NOM]}$ est extrê-

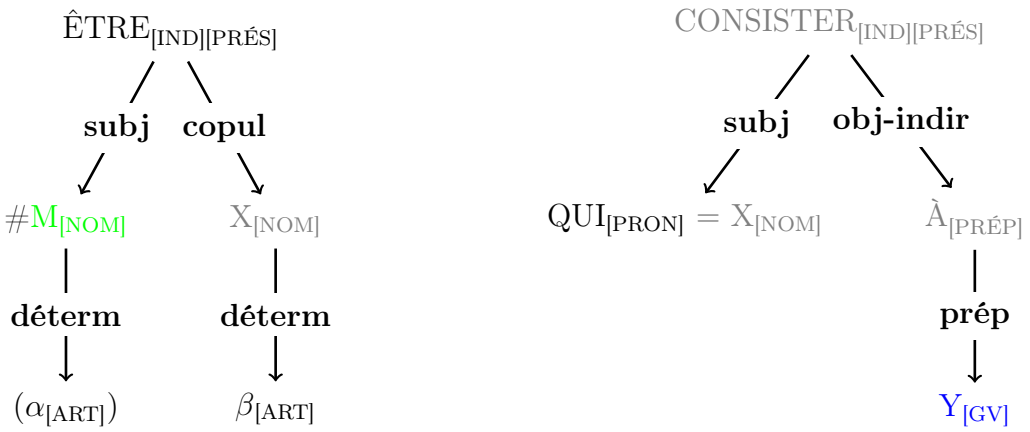


FIGURE 11 – Structures syntaxique correspondant aux définitions du type 4B

ment réduit : les lexèmes attestés sont *métier*, *activité*. Par conséquent, les définitions du type 4 sont employées pour signaler l'activité principale d'un métier sans mentionner aucun genre prochain effectif. Notons ainsi :

$$G^* = G = \varnothing$$

**Identification de la différence spécifique** La différence spécifique présumée se trouve dans la proposition subordonnée comme *l'élevage d'oiseaux ou de volailles* dans (12), sous forme d'un groupe nominal, il s'agit du groupe nominal $Y_{[GN]}$. Notons ainsi ci-dessous :

$$A = Y_{[GN]}$$

**Évaluation de la différence spécifique** La validation de la différence spécifique nécessite des ressources lexicales. Elles serons employées pour aider à la reconnaissance des activités professionnelles ou susceptibles de l'être dans le groupe nominal $Y_{[GN]}$. Les conditions de validation sont identiques à celles pour le type 2A, notons ainsi :

$$A^* = Y_{[GN]}$$

si $Y_{[GN]}$ recèle une description qui décrit une activité professionnelle ou susceptible de l'être

#### 4.4.2.4 Type 4B

**Identification du genre prochain** Le genre prochain est absent pour les mêmes raisons que nous avons évoquées à propos de l'identification du genre prochain pour les définitions du type 4A. Notons ainsi :

$$\mathrm{G}^* = \mathrm{G} = \varnothing$$

**Identification de la différence spécifique** La définition (13) du type 4B est reproduite ci-dessous pour faciliter la lecture.

*porteur d'eau*$_{\mathrm{M}}$ *est un* *métier / activité*$_{\mathrm{X}_{[\mathrm{NOM}]}}$ *qui consiste à* *transporter de l'eau*$_{\mathrm{Y}_{[\mathrm{GV}]}}$$_{\mathrm{A}}$. (13)

**Évaluation de la différence spécifique** L'évaluation voire la validation de la vraie différence spécifique pour les définitions du type 4B est similaire à celle du type 4A. La seule différence est que c'est sur un groupe verbal $\mathrm{Y}_{[\mathrm{GV}]}$ que porte l'évaluation et la validation. Par exemple, dans la définition (13), nous avons $\mathrm{Y}_{[\mathrm{GV}]}$ = *transporter de l'eau*. Puisque 'transporter de l'eau' décrit une activité pouvant faire partie des activités ou missions professionnelles, $\mathrm{Y}_{[\mathrm{GV}]}$ dans (13) est une différence spécifique valide. Notons ainsi :

$$\mathrm{A}^* = \mathrm{Y}_{[\mathrm{GV}]}$$

si $\mathrm{Y}_{[\mathrm{GV}]}$ recèle une description qui décrit une activité professionnelle ou susceptible de l'être

## 4.5 Définitions à structure non copulative

Le seul type de définitions que nous allons présenter dans la section 4.5 a une importance secondaire pour les deux raisons suivantes : premièrement, il arrive que ces types de structures figurent dans un contexte non-définitoire, c'est-à-dire, nous pouvons employer les modèles qui leur correspondent pour exprimer des relations sans énoncer une vraie définition, ce que nous allons montrer dans la section 4.5 ; deuxièmement, il ne s'agit pas de définitions analytiques, c'est-à-dire, elles ne permettent pas à elles seules d'énoncer un genre prochain du mot vedette à définir, mais les différences spécifiques, ce que nous allons également essayer d'expliquer.

### 4.5.1 Type 5

L'importance de ce type de « définitions » réside dans ce qu'il permet d'exprimer une relation générique sous forme d'un triplet, c'est-à-dire, en trois entrées. Les expressions génériques en triplet représentent le format de données, appelé « Resource Description Framework (RDF) », qui se trouve à la base du Web sémantique (Charlet, Laublet et Reynaud 2003). Une expression RDF s'écrit formellement comme (*sujet*, *prédicat*, *objet*). Cette expression exprime une relation de prédication qu'on peut décrire par l'interprétation suivante : un *prédicat* dont le nom peut être identifié par la deuxième entrée prend la première entrée (*sujet*) et la troisième entrée (*objet*) comme ses premier et deuxième arguments. Le *prédicat* sous-entendu dans le triplet fait référence à un vocabulaire conceptuel dont la documentation doit détailler sa signification.

Nous examinons les exemples ci-dessous pour illustrer nos points de vue. Nous observons que le schéma M *avoir* $Y_{[GN]}$ *pour* $X_{[NOM]}$ est à la base de leur construction phrastique en proposition simple. Nous formons le triplet suivant (M, $X_{[NOM]}$, $Y_{[GN]}$) et le considérons comme un triplet RDF en examinant les définitions (16) - (19).

*Le marchand de tapis a pour profession le commerce des tapis.* (16)
*Le puisatier a pour profession le creusement et l' entretien des puits fournissant de l'eau.* (17)
*Le marchand de biens a pour activité le négoce des immeubles /.../* (18)
*Le rédacteur de documentation en informatique a pour fonction la production des textes /.../* (19)

Par exemple, dans la définition (16), nous pouvons relever un triplet comme ci-dessous :

(M, $X_{[NOM]}$, $Y_{[GN]}$) = (*Le marchand de tapis*, *profession*, *le commerce des tapis*)

Les triplets correspondent aux définitions (17)(18)(19) :

(*Le puisatier*, *profession*, *le creusement et l'entretien des puits fournissant de l'eau.*)
(*Le marchand de biens*, *activité*, *le négoce des immeubles [...]*)
(*Le rédacteur de documentation en informatique*, *fonction*, *la production des textes [...]*)

On voit que la deuxième entrée des triples, notée par $X_{[NOM]}$ renseigne la nature de l'information que la troisième $Y_{[GN]}$ entrée apporte sur la première entrée M : par exemple, *profession*, *activité*, *fonction*.

**4.5.1.1 Reconnaissance du type** Les définitions (14)(15) brièvement examinées dans l'introduction de la section 4 sont reproduites ci-dessous :

*Le* *marchand de tapis* M *a pour profession* $_{X_{[NOM]}}$ *le commerce des tapis* $_{Y_{[GN]}}$ A. (14)
*Un* *agriculteur multiplicateur* M *a pour activité* $_{X_{[NOM]}}$ *[...]*
*la multiplication des semences* $_{Y_{[GN]}}$ A. (15)

Comme on peut le voir dans les définitions (14)(15), l'expression 'avoir quelque chose pour quelque chose' est au sein de la construction qui correspond au type 5. D'après les définitions (14)(15), nous proposons ci-dessous une caractérisation du type 5 en 4 points :

1. L'énoncé est réalisé en proposition simple
2. Le verbe principal doit être *avoir* au présent à l'indicatif
3. L'énoncé doit suivre le schéma syntaxique suivant :

   `M avoir Y[GN] pour X[NOM]`

4. $X_{[NOM]}$ doit respecter un paradigme sémantiquement restreint des lexèmes qui dénotent 'profession', 'métier', 'activité' ou 'fonction'.

**Structure syntaxique** Nous présentons ensuite la structure syntaxique qui correspond à ce type d'énoncés. Cette structure syntaxique possède une tête syntaxique verbale, il s'agit du verbe au présent à l'indicatif, que nous dénotons dans la figure 12 par $\text{AVOIR}_{[\text{IND}][\text{PRÉS}]}$. Le schéma syntaxique est représenté dans la figure 12. Dans le schéma, ce verbe principal possède trois actants syntaxiques : le mot vedette M et $\text{Y}_{[\text{GN}]}$ sont directement connectés, tandis que X[NOM] n'est relié qu'indirectement au verbe à l'aide de la préposition $\text{POUR}_{[\text{PRÉP}]}$ régie. Par ailleurs, M contrôle un déterminant obligatoirement présent dans la structure, nommé $\alpha_{[\text{ART}]}$.

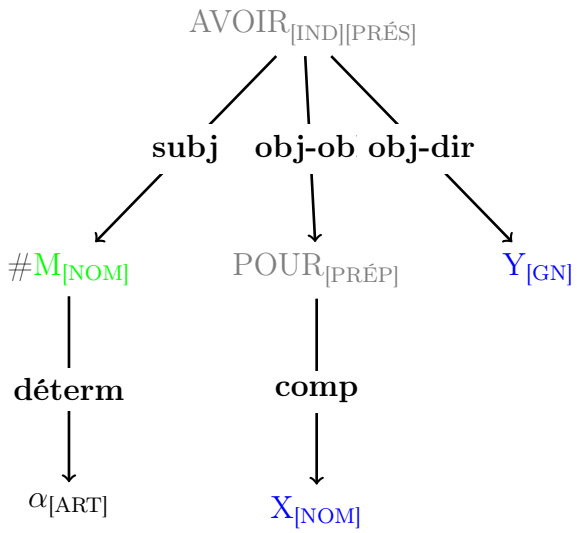


FIGURE 12 – Structures syntaxique correspondant aux énoncés du type 5

**Identification du genre prochain** Dans le modèle qui correspond au type 5 des définitions, le genre prochain n'est pas renseigné. Notons ainsi :

$$\text{G}^* = \text{G} = \varnothing$$

**Identification de la différence spécifique** Nous reproduisons les définitions (14) et (15) ci-dessous pour faciliter la lecture :

*Le* *marchand de tapis* M *a pour profession* $_{\text{X}_{[\text{NOM}]}}$ *le commerce des tapis* $_{\text{Y}_{[\text{GN}]}}$ A. (14)

*Un* *agriculteur multiplicateur* M *a pour activité* $_{\text{X}_{[\text{NOM}]}}$ *[...]*

*la multiplication des semences* $_{\text{Y}_{[\text{GN}]}}$ A. (15)

La différence spécifique présumée des définitions du type 5 correspond à $Y_{[GN]}$ qui dénote des activités professionnels comme *le commerce des tapis* dans (14) ou *la multiplication des semences* dans (15).

**Validation de la différence spécifique** Nous validons $Y_{[GN]}$ comme une vraie différence spécifique si l'activité décrite dans $Y_{[GN]}$ dénote une activité correspondant au domaine de métier.

$$A^* = Y_{[GN]}$$

si $Y_{[GN]}$ recèle une description qui décrit une activité professionnelle ou susceptible de l'être

## 4.6 Récapitulation

Nous résumons dans la table 1 les types de définitions présentées et analysées au cours de la section 4. Pour chaque modèle consacré à un type de définitions, le genre prochain et la différence spécifique présumés sont renseignés. A chaque type de définition, les critères de bonne formation du genre prochain et de la différence spécifique du modèle correspondant sont également indiqués.

| Type de modèle | G | A | $G^*$ | $A^*$ |
|---|---|---|---|---|
| Type 1A | $X_{[NOM]}$ *de* $Y_{[NOM]}$ | ∅ | C1 | ∅ |
| Type 1B | $X_{[NOM]}$ | $Y_{[NOM]}$ | C2 | C3 |
| Type 1C | $X_{[NOM]}$ | $X_{[NOM]}$ *de* $Y_{[NOM]}$ | | |
| Type 2A | $X_{[NOM]}$ | $Y_{[GN]}$ | | |
| Type 2B | $X_{[NOM]}$ | $Y_{[GV][INFINI]}$ | | |
| Type 3 | $X_{[NOM]}$ | $Y_{[GV][IND][PRÉS]}$ | | |
| Type 4A | ∅ | $Y_{[GN]}$ | ∅ | |
| Type 4B | ∅ | $Y_{[GV][INFINI]}$ | | |
| Type 5 | ∅ | $Y_{[GN]}$ | | |
| C1 : $G^*$ = G si $X_{[NOM]}$ ou $Y_{[NOM]}$ dénote un sous-domaine effectif de métier<br>C2 : $G^*$ = G si $X_{[NOM]}$ dénote un sous-domaine effectif de métier<br>C3 : $A^*$ = A si A recèle une description qui décrit une activité professionnelle ou susceptible de l'être | | | | |

TABLE 1 – Tableau de récapitulation pour les définitions présentées et analysées au cours de la section 4

# 5 Grammaire de normalisation

## 5.1 Introduction

La normalisation de structures de définitions que nous proposons ici s'effectue en deux temps. Dans un premier temps, les définitions en attente de normalisation rentrent dans un processus d'analyse. Le processus consiste à détecter les énoncés définitoires reconnaissables avec les modèles que nous avons précédemment établis et présentés dans la section 4, extraire à partir de chaque énoncé reconnu le genre prochain et la différence spécifique, valider les deux composantes essentielles de définition avec l'aide de certaines ressources lexicales, adapter préalablement le trait syntaxique de la différence spécifique aux autres types de définitions avec l'aide des fonctions lexicales et des règles de la conjugaison. La seconde phase consiste à régénérer les définitions dans les autres modèles de définitions à partir des composantes de définition obtenues de la phase précédente, en fonction de la validité du genre prochain et de la différence spécifique.

Le processus d'analyse dans la grammaire de normalisation est guidé par la table récapitulative de la section 4.6. Nous la reproduisons sous forme d'un graphe illustratif dans la figure 13.

Pour exposer la grammaire de normalisation, nous procédons de la manière suivante : nous commençons par présenter un sous-ensemble de définitions que nous avons constitué à partir des exemples analysés lors de la section précédente. Il contient au total 5 énoncés définitoires dont chacun correspond à un type majeur enregistré dans notre typologie. Nous montrons ensuite comment employer la grammaire de normalisation que nous formalisons ici sous la forme de graphe schématique pour transformer ces énoncés dans d'autres types de structure de définition, dans la mesure où la transformation permet de conserver le même contenu définitoire, ou au contraire, la perte d'information dans ce processus mérite une discussion particulière.

## 5.2 Analyse de définitions

### 5.2.1 Exemples choisis pour la démonstration

Voici les 5 exemples de définitions qui vont faire l'objet de la démonstration :

*Le* *lardonnier*$_{\text{M}}$ *est un* *préparateur*$_{\text{G}}$ *de* *lardons*$_{\text{A}}$. (5)

*Le* *pâtissier*$_{\text{M}}$ *est un* *artisan*$_{\text{G}}$ *spécialisé dans* *la fabrication des pâtisseries*$_{\text{A}}$. (7)

*Un* *corailleur*$_{\text{M}}$ *est une* *personne*$_{\text{G}}$ *qui* *pêche le corail afin de le revendre*$_{\text{A}}$. (11)

*porteur d'eau*$_{\text{M}}$ *est un* *métier / activité* *qui consiste à* *transporter de l'eau*$_{\text{A}}$. (13)

*Le* *marchand de tapis*$_{\text{M}}$ *a pour profession* *le commerce des tapis*$_{\text{A}}$. (14)

La définition (5) est classée dans le type 1C, elle contient un dérivé sémantique nominal actantiel *préparateur* qui permet de décrire avec *lardons* l'activité principale du métier de *lardonnier*. La définition (7) se classe dans le type 2A, elle contient un groupe nominal *la fabrication des pâtisseries* qui dénote l'activité principale de *pâtissier*. La définition (11) fait partie des définitions du type 3, elle possède une proposition complexe dont la subordonnée contient un groupe verbal à l'indicatif qui consiste à décrire l'activité de métiers de *corailleur*. La définition (13) concernant le métier de *porteur d'eau* est classée dans le type 4B, elle possède une proposition complexe dont la subordonnée contient l'expression *consister à* et le groupe verbal *transporter de l'eau* introduite par l'expression décrit l'activité du métier en tant qu'un groupe verbal. La définition (14) qui emploie l'expression *avoir pour* est classée dans le type 5, elle renseigne l'activité du métier de *marchand de tapis* par un groupe nominal, à savoir, *le commerce des tapis*.

### 5.2.2 Extraction du genre prochain et de la différence spécifique présumés

Avant d'extraire des composantes essentielles d'une définition, le genre prochain et la différence spécifique, il faut d'abord repérer son type, le type de modèle qui lui correspond.

Pour y parvenir nous effectuons une analyse syntaxique sur la définition, puis comparons son schéma syntaxique à chacun des modèles de définition que nous avons proposés dans la typologie afin d'en déterminer le type ou de signaler un type non-reconnu par la typologie.

Pour les définitions reconnues par la typologie de structures de définition, le modèle syntaxique que nous proposons dans la section 4 permet de distinguer des arguments de

définitions. Cette distinction se fait concrètement par le marquage d'un lexème ou d'un groupe lexical principalement par $X_{[NOM]}$, $Y_{[NOM]}$, $Y_{[GN]}$, $Y_{[GV][INFINI]}$ ou $Y_{[GV][IND][PRÉS]}$. Les définitions annotées de la sorte sont présentées ci-dessous :

T1C : *Le* *lardonnier* M *est un* *préparateur* $X_{[NOM]}$ G *de* *lardons* $Y_{[NOM]}$ A. (5)

T2A : *Le* *pâtissier* M *est un* *artisan* $X_{[NOM]}$ G *spécialisé dans* *la fabrication des pâtisseries* $Y_{[NOM]}$ A. (7)

T3 : *Un* *corailleur* M *est une* *personne* $X_{[NOM]}$ G *qui* *pêche le corail afin de le revendre* $Y_{[GV][IND][PRÉS]}$ A. (11)

T4B : *porteur d'eau* M *est un* *métier / activité* $X_{[NOM]}$ *qui consiste à* *transporter de l'eau* $Y_{[GV][INFINI]}$ A. (13)

T5 : *Le* *marchand de tapis* M *a pour profession* $X_{[NOM]}$ *le commerce des tapis* $Y_{[NOM]}$ A. (14)

Nous allons, par la suite, nous servir du graphe schématique dans la figure 13 pour repérer le genre prochain et la différence spécifique présumés. Dans ce graphe en question, les nœuds placés au milieu désignent chacun un modèle de définitions reconnu par notre typologie, ceux placé en haut et en bas désignent des composantes essentielles de définition avec une précision faite sur leurs traits syntaxiques. Les flèches pointent toujours d'un nœud désignant un modèle vers un nœud représentant soit le genre prochain en haut, soit une des différences spécifiques en bas. Une flèche exprime une opération d'extraction. Par exemple, la flèche qui part du nœud étiqueté *T2A* et pointe vers le genre prochain, signifie que le modèle du type 2A permet d'extraire un genre prochain présumé qui correspond à l'élément lexical du modèle, $X_{[NOM]}$. Ce dernier élément lexical est renseigné, marqué sur la flèche.

C'est ainsi que nous avons extrait de ces 5 exemples le genre prochain et la différence spécifique présumés. Le résultat est présenté dans la table ci-dessous :

### 5.2.3 Validation du genre prochain et de la différence spécifique

**5.2.3.1 Validation du genre prochain** Dans la table d'extraction 2, trois genres prochains présumés sont présents : *préparateur*, *artisan*, *personne*. Or, *préparateur*, *personne* ne sont pas reconnus comme hyperonymes du domaine de métier. Seul artisan est un vrai genre prochain dans nos exemples.

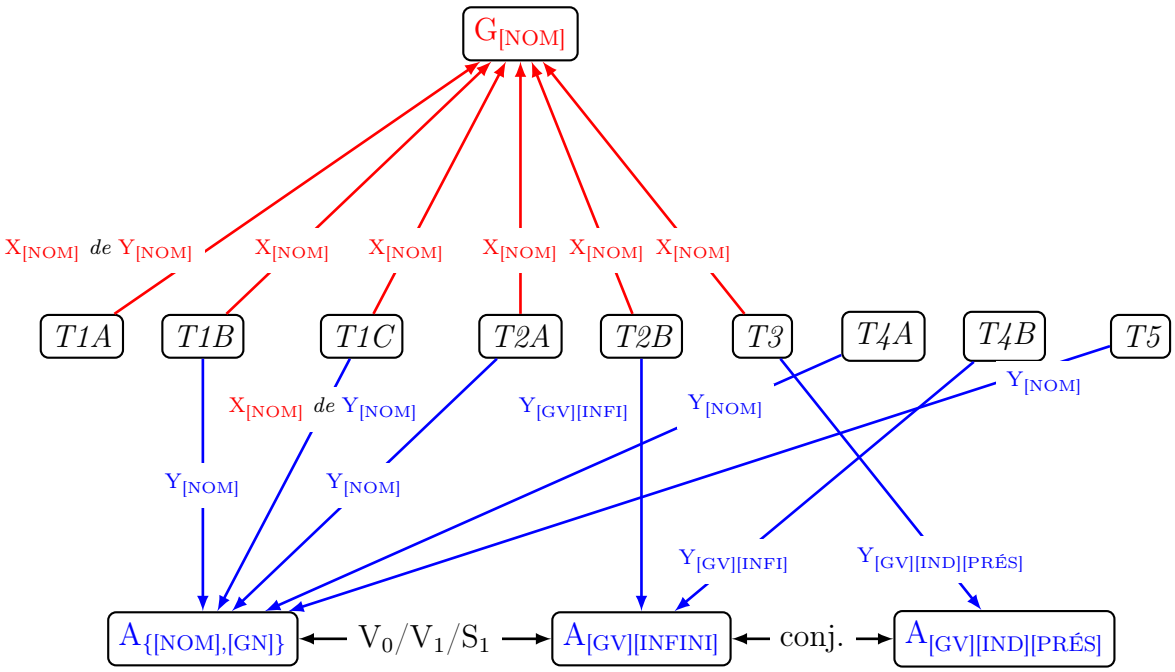


FIGURE 13 – Graphe schématique de la grammaire de normalisation en phase d'analyse

| Id | Type | Mot vedette M | Genre prochain G |
|---|---|---|---|
| 5 | *T1C* | *lardonnier* | *préparateur* |
| 7 | *T2A* | *pâtissier* | *artisan* |
| 11 | *T3* | *corailleur* | *personne* |
| 13 | *T4B* | *porteur d'eau* | ∅ |
| 14 | *T5* | *marchand de tapis* | ∅ |

| Id | Différence spécifique A |
|---|---|
| 5 | $A^*_{\{[NOM],[GN]\}}$ = *préparateur de lardons* |
| 7 | $A^*_{\{[NOM],[GN]\}}$ = *la fabrication des pâtisseries* |
| 11 | $A^*_{[GV][IND][PRÉS]}$ = *pêche le corail afin de le revendre* |
| 13 | $A^*_{[GV][INFINI]}$ = *transporter de l'eau* |
| 14 | $A^*_{\{[NOM],[GN]\}}$ = *le commerce des tapis* |

TABLE 2 – Résultat de la validation du genre prochain et de la différence spécifique

#### 5.2.3.2 Validation de la différence spécifique

Par contre, les différences spécifiques désignent des activités précises dans tous les exemples. Elles sont assimilables au groupe verbal constitué d'un verbe d'action et son patient. Dans la ligne de la table 2 qui correspond à la définition (11), la différence spécifique contient une autre information *afin de le revendre*, un constituant adverbial qui renseigne le but fixé pour l'activité.

Le résultat de la validation du genre prochain et de la différence spécifique pour chacun de nos exemples est résumé dans la table 3.

| Id | Type | Mot vedette M | Vrai genre prochain $G^*$ |
|---|---|---|---|
| 5 | *T1C* | *lardonnier* | ~~*préparateur*~~ |
| 7 | *T2A* | *pâtissier* | *artisan* |
| 11 | *T3* | *corailleur* | ~~*personne*~~ |
| 13 | *T4B* | *porteur d'eau* | ∅ |
| 14 | *T5* | *marchand de tapis* | ∅ |

| Id | Différence suffisamment spécifique $A^*$ |
|---|---|
| 5 | $A^*_{\{[NOM],[GN]\}}$ = *la préparation de lardons* |
| 7 | $A^*_{\{[NOM],[GN]\}}$ = *la fabrication des pâtisseries* |
| 11 | $A^*_{[GV][IND][PRÉS]}$ = *pêche le corail afin de le revendre* |
| 13 | $A^*_{[GV][INFINI]}$ = *transporter de l'eau* |
| 14 | $A^*_{\{[NOM],[GN]\}}$ = *le commerce des tapis* |

TABLE 3 – Résultat de la validation du genre prochain et de la différence spécifique

### 5.2.4 Pré-adaptation de la différence spécifique

Considérons une transformation d'une définition d'un certain type dans un autre type. Lorsque sur la base du modèle du premier type, les arguments lexicaux que nous avons extraits possèdent des traits syntaxiques incompatibles avec les traits des arguments que le modèle du second type demande pour créer une définition, l'adaptation des traits syntaxiques est un passage nécessaire. Puisque le genre prochain est un hyperonyme du mot vedette, celui-ci est censé partager les mêmes traits syntaxiques. Par conséquent, l'adaptation du genre prochain n'est pas nécessaire. En revanche, la différence spécifique que nous pouvons obtenir, par exemple, à partir de la définition (11) prend la forme d'un groupe verbal à l'infinitif, alors que pour employer le modèle du type 5, la différence spécifique doit être réalisée sous forme d'un groupe nominal. Dans ce cas-là, on emprunte une nominalisation pour en produire un équivalent nominal.

L'adaptation des traits syntaxiques peut s'effectuer de deux manières : la première

approche consiste à pré-générer tous les équivalents possibles pour chaque différence spécifique, nous l'appelons la pré-adaptation de la différence spécifique. Cette approche est envisageable lorsque les types de destination vers lesquels on convertit des définitions ne sont pas en nombre trop conséquent. Dans ce cas-là, le coût en terme de quantité de données et de temps de traitement reste acceptable. Une autre approche consiste à générer les équivalents à la demande et à la volée. C'est-à-dire que c'est au moment de décider de transformer une définition dans un autre type, que nous analysons la compatibilité des traits syntaxiques entre la structure de définition du départ et la structure de définition de destination. Ainsi, l'adaptation procède en fonction des traits syntaxiques requis par la structure de définition de destination. Cette seconde approche semble plus flexible puisqu'entre deux types de structure, la conversion n'est pas toujours possible et pertinente à moins que la perte d'informations définitoires soit acceptable. Autrement dit, il est possible que certains équivalents d'une différence spécifique ne sont pas utilisés ou utilisables par la normalisation.

En ce qui nous concerne, nous optons pour la pré-adaptation de la différence spécifique afin de nous concentrer sur la présentation de la grammaire de normalisation, en laissant la discussion sur l'implémentation de cette grammaire à des études ultérieures.

Dans le graphe de la figure 13, les nœuds qui correspondent à trois différentes réalisations de la différence spécifique $A_{\{[\text{NOM}],[\text{GN}]\}}$, $A_{[\text{GV}][\text{IND}][\text{PRÉS}]}$, $A_{[\text{GV}][\text{INFINI}]}$ sont reliés entre deux par des doubles flèches en couleur noire. Ces flèches-ci représentent des ressources que nous empruntons pour effectuer la pré-adaptation de la différence spécifique afin de passer d'une forme syntaxique à une autre.

**5.2.4.1 Adaptation basée sur les fonctions lexicales** Le passage d'une forme nominale $A_{\{[\text{NOM}],[\text{GN}]\}}$ à une forme verbal $A_{[\text{GV}][\text{INFINI}]}$ est une verbalisation tandis qu'une nominalisation consiste en une conversion de $A_{[\text{GV}][\text{INFINI}]}$ en $A_{\{[\text{NOM}],[\text{GN}]\}}$.

Deux verbalisations sont distinguées, dénotées respectivement par les fonctions lexicales $V_0$ et $V_1$. $V_0$ représente une fonction de verbalisation qui donne la forme verbale d'un nom qui dénote une action ou un fait, comme $V_0(\textit{commerce}) = \textit{vendre}$ ; $V_1$, quant à elle, dénote une fonction lexicale qui donne la forme verbale d'un nom désignant une action ou un fait par celui qui la / le pratique, par exemple, $V_1(\textit{préparateur}) = \textit{préparer}$.

La nominalisation des activités peut être représentée par une fonction, il s'agit de la fonction $S_0$. Elle permet de donner la forme nominale d'un verbe qu'elle reçoit. Par

exemple, $S_0(transporter) = transportation$.

**5.2.4.2 Adaptation de la modalité du verbe** Dans la figure 13, le passage entre $A_{[GV][INFINI]}$ et $A_{[GV][IND][PRÉS]}$ représente un changement de mode verbal. Des ressources sur la conjugaison sont largement disponibles pour ce type de transformation.

Le résultat de la ré-adaptation syntaxique de la différence spécifique pour les 5 exemples, grâce aux notions de fonctions lexicales et celle de la conjugaison est montré dans la table 4. Les équivalents verbaux à l'indicatif ne sont pas présentés car leur conversion vers le mode infinitif paraît évident.

| Id | Type | M | $G^*$ |
|---|---|---|---|
| 5 | *T1C* | *lardonnier* | ~~*préparateur*~~ |
| 7 | *T2A* | *pâtissier* | *artisan* |
| 11 | *T3* | *corailleur* | ~~*personne*~~ |
| 13 | *T4B* | *porteur d'eau* | ∅ |
| 14 | *T5* | *marchand de tapis* | ∅ |

| Id | $A^*_{\{[NOM],[GN]\}}$ | $A^*_{[GV][INFINI]}$ |
|---|---|---|
| 5 | *(la) préparation de lardons* | $V_1$ = *préparer des lardons* |
| 7 | *(la) fabrication des pâtisseries* | $V_0$ = *fabriquer les pâtisseries* |
| 11 | $S_0$ = *(la) pêche du corail afin de le revendre* | *pêcher le corail afin de le revendre* |
| 13 | $S_0$ = *(la) transportation de l'eau* | *transporter de l'eau* |
| 14 | *(le) commerce des tapis* | $V_0$ =*vendre les tapis* |

TABLE 4 – Le résultat de pré-adaptation de la différence spécifique accompagné de la présentation du genre prochain

## 5.3 Synthèse de définitions

La seconde phase de la normalisation consiste à régénérer, à partir des composantes validées et présentées dans la table 4, la définition dans tous les types reconnus par la typologie dans la mesure du possible et de la pertinence. Pour ce faire, nous reproduisons la table 4 ci-dessous.

| Id | Type | M | $G^*$ |
|---|---|---|---|
| 5 | *T1C* | *lardonnier* | ~~*préparateur*~~ |
| 7 | *T2A* | *pâtissier* | *artisan* |
| 11 | *T3* | *corailleur* | ~~*personne*~~ |
| 13 | *T4B* | *porteur d'eau* | ∅ |
| 14 | *T5* | *marchand de tapis* | ∅ |

| Id | $A^*_{\{[NOM],[GN]\}}$ | $A^*_{[GV][INFINI]}$ |
|---|---|---|
| 5 | *(la) préparation de lardons* | *préparer des lardons* |
| 7 | *(la) fabrication des pâtisseries* | *fabriquer les pâtisseries* |
| 11 | *(la) pêche du corail afin de le revendre* | *pêcher le corail afin de le revendre* |
| 13 | *(la) transportation de l'eau* | *transporter de l'eau* |
| 14 | *(le) commerce des tapis* | *vendre les tapis* |

TABLE 5 – Le genre prochain et la différence spécifique pré-adaptée et validés

### 5.3.1 Modèles de définitions du point de vue de la synthèse

En ce qui concerne la synthèse de définitions, le modèle du type 1 ne fait pas l'objet de génération puisque les types 1A, 1B ne permettent pas d'exprimer la différence spécifique, 1C ne le peut que partiellement à travers le dérivé sémantique nominal actantiel. Nous proposons ci-dessous les modèles pour génération de définitions, à partir du genre prochain et de la différence spécifique. Les modèles présentés concernent les types 2A, 2B, 3, 4A, 4B, 5 :

T2A(M, $G^*$, $A^*_{[GN]}$) : $\alpha_{[ART]}$ M *est* $\beta_{[ART]}$ $G^*$ {*spécialisé*(*e*) dans, *chargé*(*e*) *de*} $A^*_{[GN]}$

T2B(M,$G^*$, $A^*_{[GV][INFINI]}$) : $\alpha_{[ART]}$ M *est* $\beta_{[ART]}$ $G^*$ {*chargé*(*e*), *responsable*} *de* $A^*_{[GV][INFINI]}$

T3(M, $G^*$, $A^*_{[GV][IND][PRÉS]}$) : $\alpha_{[ART]}$ M *est* $\beta_{[ART]}$ $G^*$ *qui* $A^*_{[GV][IND][PRÉS]}$

T4A(M, ∅, $A^*_{[GN]}$) : ($\alpha_{[ART]}$) M *est un métier qui consiste en* $A^*_{[GN]}$

T4B(M, ∅, $A^*_{[GV][INFINI]}$) : ($\alpha_{[ART]}$) M *est un métier qui consiste à* $A^*_{[GV][INFINI]}$

T5(M, ∅, $A^*_{[GN]}$) : $\alpha_{[ART]}$ M *a pour* {*profession*, *activité*, *fonction*} $A^*_{[GN]}$

### 5.3.2 Génération de définitions

Nous montrons comment employer les modèles destinés à la génération de définitions proposés dans la section 5.3.1 pour produire des paraphrases définitoires à partir des composantes de définitions présentées dans la table 5 pour les 5 exemples. Nous pouvons constater qu'à part la définition (7), les autres exemples ne possèdent finalement pas de vrai genre prochain. En absence du genre prochain, seul les modèles T4A(M, ∅, $A_{[GN]}$), T4B(M, ∅, $A_{[GV][IND][PRÉS]}$) et T5(M, ∅, $A_{[GN]}$) sont accessibles.

Pour produire une nouvelle définitions d'un certain type, nous reprenons le modèle qui lui correspond. Ensuite, nous remplaçons ses arguments formels par leurs valeurs effectives, présentées dans la table 5 pour chaque exemple, ces valeurs sont issues de la précédente extraction et évaluation. Les paraphrases obtenues ainsi pour chacun des exemples sont listées ci-dessous.

Pour la définition (5) du métier de *lardonnier* du type 1C, nous avons comme paraphrases :

T4A = Lardonnier est *un métier qui consiste en* la préparation des lardons
T4B = Lardonnier *est un métier qui consiste à* préparer des lardons
T5 = Le lardonnier *a pour profession* la préparation des lardons

Pour la définition (7) du métier de *pâtissier* du type 2A, nous avons accès à 5 modèles :

T2B = Le pâtissier *est* un artisan *responsable de* la fabrication des pâtisseries
T3 = Le pâtissier *est* un artisan *qui* fabrique les pâtisseries
T4A = Pâtissier est *un métier qui consiste en* la fabrication des pâtisseries
T4B = Pâtissier *est un métier qui consiste à* fabriquer les pâtisseries
T5 = Le pâtissier *a pour profession* la fabrication des lardons

Pour la définition (11) du métier de *corailleur* du type 3, nous pouvons régénérer la définition dans les types T4A, T4B, T5 :

T4A = Corailleur est *un métier qui consiste en* la pêche du corail afin de le revendre
T4B = Corailleur *est un métier qui consiste à* pêcher le corail afin de le revendre
T5 = Le corailleur *a pour profession* la pêche du corail afin de le revendre

Pour la définition (13) du métier de *porteur d'eau* du type 4B, nous pouvons régénérer la définition dans les types T4A, T5 :

T4A = Porteur d'eau est *un métier qui consiste en* la transportation de l'eau

T5 = Le porteur d'eau *a pour profession* la transportation de l'eau

Pour la définition (14) du métier de *Marchand de tapis* du type 5, nous pouvons régénérer la définition dans les types T4A, T4B :

T4A = Marchand de tapis est *un métier qui consiste en* la commerce des tapis

T4B = Marchand de tapis *est un métier qui consiste à* vendre les tapis

# 6 Expérimentation

Dans cette section, nous allons présenter une chaîne de traitement que nous avons conçue pour la normalisation des structures de définition. Le résultat d'une expérimentation que nous avons menée avec ce système, sur un corpus d'évaluation, sera également fourni et commenté.

## 6.1 Implémentation

La chaîne de traitement est représentée sous forme d'un diagramme de blocs dans la figure 14. Un bloc rectangulaire correspond à un processus, tandis qu'un bloc en forme d'ellipse désigne des données qui peuvent être soit les données d'entrée ou de sortie, soit des ressources que le système utilise.

Notre chaîne de traitement est constituée de 5 processus principaux, que nous désignons dans le schéma par les intitulés suivants : *détection*, *extraction*, *validation*, *conversion*, *génération*. Chacun des blocs dispose d'une entrée principale, des entrées de ressources et une sortie.

Le traitement commence par une détection du type correspondant à la définition en attente de normalisation, la détection s'appuie sur les données contenant la description des modèles issues de notre typologie de définitions. Les modèles des définition sont décrits en expressions régulières pour permettre une conception rapide du repérage.

Lorsque la structure de définition est reconnue, le processus envoie le type identifié au processus d'extraction. Le bloc intitulé *extraction* consiste à extraire le genre prochain et la différence spécifique présumés en utilisant le modèle de définitions qui correspond au type déterminé par le bloc précédent.

Le genre prochain et la différence présumés sont ensuite transmis au bloc de validation. Dans la phase de validation, des ressources lexicales sont employées pour évaluer la pertinence de ces deux composantes essentielles de définition par rapport au domaine que nous avons choisi à titre d'exemple, celui de métier. Le genre prochain et la différence spécifique validés sont ensuite envoyés au bloc suivant pour la conversion. Comme protocole de communication, un genre prochain non-validé s'exprime par l'envoi d'un genre prochain validé mais nul[7] à l'étape suivante, de même pour la différence spécifique.

7. Il s'agit d'une chaîne de caractère vide

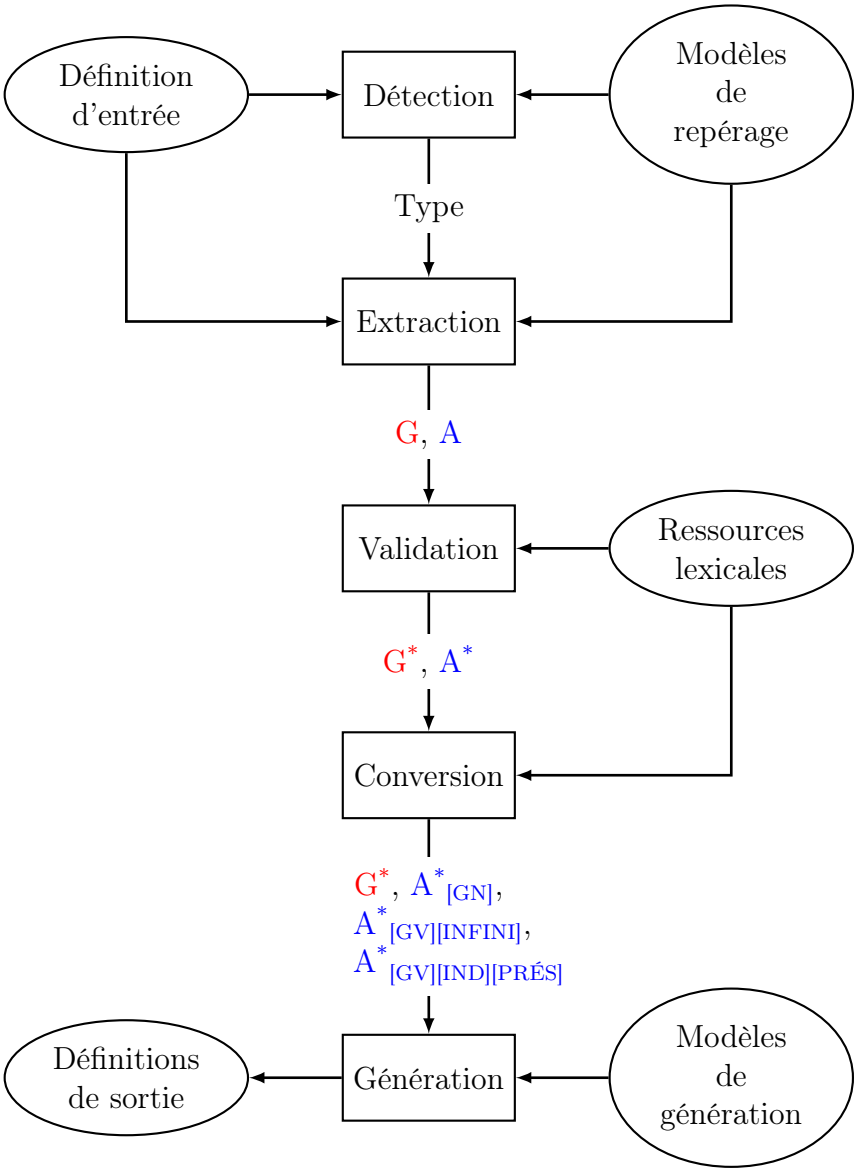


FIGURE 14 – Chaîne de traitement pour la normalisation de structures de définition

Le processus de conversion consiste à produire des formes alternatives de la différence spécifique : les formes verbales de modalité différente pour la même différence spécifique nominale, les différentes formes nominales pour la même différence spécifique verbale, etc. Cette conversion s'appuie sur des ressources lexicales que nous avons conçues pour l'expérimentation. Ces ressources contiennent des fonctions lexicales et règles de conjugaison.

Le bloc intitulé *génération* reçoit le genre prochain validé et les différentes formes de la différence spécifique validé issues du précédent traitement. Le processus de génération consiste à synthétiser des définitions en essayant de remplir les arguments de chaque modèle de génération [8] par les composantes validées de définitions qu'il reçoit.

### 6.1.1 Un type supplémentaire pour la génération de définitions

Un sixième type a été implémenté comme modèle de génération, il permet de signaler la différence spécifique comme une information portant sur le métier et le genre prochain comme une information sur la personne qui l'exerce. Le type intitulé T6 est réparti en 2 sous-types intitulés T6A et T6B. Nous formulons leurs modèles de génération ci-dessous :

T6A(M, $G^*$, $A^*_{[GV][INFINI]}$) : ($\alpha_{[ART]}$) M *est un(e)* $G^*$ dont le métier est de $A^*_{[GV][INFINI]}$
T6B(M, $G^*$, $A^*_{[GN]}$) : ($\alpha_{[ART]}$) M *est un(e)* $G^*$ dont le métier consiste en $A^*_{[GN]}$

## 6.2 Corpus d'évaluation

Nous rappelons que le corpus d'étude contient les 100 définitions les plus courtes du corpus intégral. Un corpus d'évaluation est créé par la suite en sélectionnant dans le reste du corpus intégral les 50 définitions les plus courtes. Nous allons ensuite présenter le résultat obtenu en appliquant le système de normalisation sur le corpus d'évaluation.

## 6.3 Résultat d'expérimentation

### 6.3.1 Détection du type de structure de définition

En ce qui concerne la détection du type de définition, le taux de rappel est à 82% : 41 définitions ont été correctement repérées, 9 définitions ont, par contre, été déclarées non-identifiées. Comme nous pouvons voir dans la table 6, le type 3 représente une grande

8. Les modèles de génération sont des propositions définitoires avec des cases manquantes que nous avons proposés à un sous-ensemble des types étudiés dans la typologie, ils sont listés dans la section 5.3.1.

majorité de définitions, 21 définitions sur 50 (41%). Les définitions du type 2A et 2B viennent en deuxième position avec un taux d'occurrence de 26%. Arrive en troisième position, les définitions du type 1 qui réunit trois sous-types qui ensemble représentent 8% de définitions dans le corpus d'évaluation.

| Type | Occurrences | Taux |
|---|---|---|
| T3 | 21 | 42% |
| T2A | 10 | 20% |
| T2B | 3 | 6% |
| T1C | 2 | 4% |
| T4B | 2 | 4% |
| T1A | 1 | 2% |
| T1B | 1 | 2% |
| T5 | 1 | 2% |
| Non-identifié | 9 | 18% |

TABLE 6 – Distribution des types de définition reconnus ou non-identifiées

### 6.3.2 Extraits des exemples générés

Nous sélectionnons 4 définitions traitées par le système de normalisation pour illustrer le résultat de la sortie de la chaîne de traitement. Pour ces 4 exemples, le résultat de la validation du genre prochain et de la différence spécifique est présenté dans la table 7. Nous pouvons constater que seul l'exemple avec l'identifiant 37 du type T4B possède un faux genre prochain et que toutes les définitions ont une différence spécifique jugée suffisamment complète pour caractériser une activité de métier.

| Id | Type | un vrai G ? | A suffisamment spécifique ? |
|---|---|---|---|
| 21 | T2B | o | o |
| 28 | T1B | o | o |
| 37 | T4B | x | o |
| 47 | T3 | o | o |

TABLE 7 – Extrait du résultat de l'évaluation sur la pertinence des composantes essentiels de définition

L'exemple ci-dessous avec identifiant 21 portant sur le métier de *garde-barrières* est originellement classé dans le type intitulé T2B, sa forme originale est placée dans la première ligne. Dans les lignes qui suivent, se trouvent les paraphrases générées par le système dans les types suivants : T2A, T3, T4A, T4B, T5, T6A, T6B.

*=== Id 21 ===*

M = *garde-barrière*

*T2B : Un M est un agent des chemins de fer chargé de garder les barrières d'un passage à niveau.*

*T2A : Le M est un agent des chemins de fer spécialisé dans le contrôle des barrières d'un passage à niveau*

*T2B : Le M est un agent des chemins de fer chargé de garder les barrières d'un passage à niveau*

*T3 : Le M est un agent des chemins de fer qui garde les barrières d'un passage à niveau*

*T4A : Le M est un métier qui consiste en le contrôle des barrières d'un passage à niveau*

*T4B : Le M est un métier qui consiste à garder les barrières d'un passage à niveau*

*T5 : Le M a pour profession le contrôle des barrières d'un passage à niveau*

*T6A : Le M est une agent dont le métier est de garder les barrières d'un passage à niveau*

*T6B : Le M est une agent dont le métier consiste en le contrôle des barrières d'un passage à niveau*

L'exemple ci-dessous avec identifiant 28 portant sur le métier de *cordier* est originellement classé dans le type intitulé T1C, sa forme originale est placée dans la première ligne. Dans les lignes qui suivent, se trouvent les paraphrases générées par le système dans les types suivants : T2A, T2B, T3, T4A, T4B, T5, T6A, T6B.

*=== Id 28 ===*

M = *cordier*

*T1C : Un* M *est un ouvrier fabricant de la corde , des cordages ou des câbles .*

*T2A : Le* M *est un ouvrier spécialisé dans la fabricaiton de la corde, des cordages ou des câbles*

*T2B : Le* M *est un ouvrier chargé de fabriquer la corde, les cordages ou les câbles*

*T3 : Le* M *est un ouvrier qui fabrique la corde, les cordages ou les câbles*

*T4A : Le* M *est un métier qui consiste en la fabricaiton de la corde, des cordages ou des câbles*

*T4B : Le* M *est un métier qui consiste à fabriquer la corde, les cordages ou les câbles*

*T5 : Le* M *a pour profession la fabricaiton de la corde, des cordages ou des câbles*

*T6A : Le* M *est une personne dont le métier est de fabriquer la corde, les cordages ou les câbles*

*T6B : Le* M *est une personne dont le métier consiste en la fabricaiton de la corde, des cordages ou des câbles*

L'exemple ci-dessous avec identifiant 47 portant sur le métier de *marchand d'allumettes* est originellement classé dans le type intitulé T4B, sa forme originale est placée dans la première ligne. Dans les lignes qui suivent, se trouvent les paraphrases générées par le système dans les types suivants :T4A, T5.

*=== Id 37 ===*

M = *marchand d'allumettes*

*T4B :* M *est un ancien métier consistant à vendre dans la rue des allumettes.*

*T4A : Le* M *consiste en la commerce dans la rue des allumettes*

*T5 : Le* M *a pour profession la vente dans la rue des allumettes*

L'exemple ci-dessous avec identifiant 47 portant sur le métier de *compositeur ou une compositrice de musique* est originellement classé dans le type intitulé T3, sa forme originale est placée dans la première ligne. Dans les lignes qui suivent, se trouvent les paraphrases générées par le système dans les types suivants : T2A, T2B, T4A, T4B, T5, T6A, T6B.

*=== Id 47 ===*

M = *compositeur ou une compositrice de musique*

*T3 : Un* M *est une personne musicienne qui élabore de la musique.*

*T2A : Le* M *est une personne musicienne spécialisé dans l'élaboration de la musique*

*T2B : Le* M *est une personne musicienne chargé d'élaborer de la musique*

*T4A : Le* M *est un métier qui consiste en l'élaboration de la musique*

*T4B : Le* M *est un métier qui consiste à élaborer de la musique*

*T5 : Le* M *a pour profession l'élaboration de la musique*

*T6A : Le* M *est une personne dont le métier est d'élaborer de la musique*

*T6B : Le* M *est une personne dont le métier consiste en l'élaboration de la musique*

# 7 Conclusion

Nous avons commencé par souligner la problématique de la faible comparabilité des définitions non-formalisées en la montrant en examinant des extraits de définition sur la Wikipédia française. Pour pouvoir étudier de plus près cette question, un corpus de définitions portant sur le domaine de métier a été constitué à partir de Wikipédia, il fournit les données principales à notre étude sur les structures de définition.

5 types majeurs de structure de définition ont été ensuite proposés et analysés, ce qui permet de donner une typologie sur les définitions encyclopédiques. Bien que la typologie que nous avons proposée ne soit pas exhaustive, elle permet de prendre en compte une majorité de définitions courtes dans le corpus.

Une grammaire de normalisation est ensuite conçue et formalisée, elle permet de guider l'extraction des composantes essentielles de définition comme le genre prochain et la différence spécifique présumés, elle fournit les critères d'évaluation pour valider ou réfuter le genre prochain et la différence spécifique présumés, elle permet à la fin, de reformuler la définition en employant les modèles de définition pour la génération.

La reformulation consiste précisément à synthétiser des définitions de destination dans de différents types à partir des composantes extraites et évaluées d'une définition de source. Pour automatiser les différents traitements qui correspondent à cette grammaire de normalisation, ou de reformulation, nous proposons une chaîne de traitement. Cette chaîne schématise l'implémentation de notre système de normalisation.

Le résultat que nous avons obtenu à la sortie du système de normalisation montre que notre approche permet de générer automatiquement des paraphrases définitoires pour chaque définition reconnue par notre typologie. Faisant bénéficier ainsi à chaque définition un groupe de paraphrases qui correspondent à d'autres type reconnus, notre système contribue à un travail de la formalisation automatique des définitions visant à améliorer la comparabilité des définitions.

# Références


Blanc, Olivier (2006). « Algorithmes d'analyse syntaxique par grammaires lexicalisées : optimisation et traitement de l'ambiguité ». Thèse de doct. Université Paris-Est.

Bourigault, Didier et Nathalie Aussenac-Gilles (2003). « Construction d'ontologies à partir de textes ». In : *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues*, p. 27–50.

Brun, Caroline et Caroline Hagege (2003). « Normalization and paraphrasing using symbolic methods ». In : *Proceedings of the second international workshop on Paraphrasing-Volume 16*. Association for Computational Linguistics, p. 41–48.

Brun, Caroline, Bernard Jacquemin et Frédérique Segond (2005). « Exploitation de dictionnaires\'{e} lectroniques pour la d\'{e} sambigu\"{i} sation s\'{e} mantique lexicale ». In : *arXiv preprint cs/0506049*.

Charlet, Jean, Philippe Laublet et Chantal Reynaud (2003). *Le web sémantique*. Cépaduès-Ed.

Hagège, Caroline et Claude Roux (2003). « Entre syntaxe et sémantique : Normalisation de la sortie de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes ». In : *TALN, Batz-sur-mer, France*.

Kamel, Mouna et Nathalie Aussenac-Gilles (2009). « Construction automatique d'ontologies à partir de spécifications de bases de données ». In : *IC 2009 : 20es Journées Francophones d'Ingénierie des Connaissances «Connaissance et communautés en ligne»*, p. 85–96.

L'Homme, Marie-Claude (1998). « Fonctions lexicales pour représenter les relations sémantiques entre termes ». In : *Revue. Volume* 1.1.

L'homme, Marie-Claude (2008). « Ressources lexicales, terminologiques et ontologiques : une analyse comparative dans le domaine de l'informatique ». In : *Revue française de linguistique appliquée* 13.1, p. 97–118.

Mel'čuk, I.A. et J. Milićević (2014). *Introduction à la linguistique*. HR.Hors Collec vol. 1. Hermann. isbn : 9782705680589. url : https://books.google.fr/books?id=oNc3mwEACAAJ.

Polguère, Alain (2003). *Lexicologie et sémantique lexicale : notions fondamentales*. Pum.

Polguère, Alain (2011). « Classification sémantique des lexies fondée sur le paraphrasage ». In : *Cahiers de lexicologie* 98, p. 197–211.

Tchechmedjiev, Andon (2012). « État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances ». In : *JEP-TALN-RECITAL 2012*, p. 295.

Verlinde, Serge et al. (2004). « Les schémas actanciels dans le dictionnaire : point de convergence entre la morphologie et la sémantique lexicale ». In : *G. Williams et S. Vessier (éds), Proceedings of the Eleventh EURALEX International Congress, EURALEX*, p. 427–436.

*Wikimedia Downloads* (2017). URL : https://dumps.wikimedia.org/.

Wikipédia (2017a). *Aide :Syntaxe (wikicode)*. URL : https://fr.wikipedia.org/wiki/Aide:Syntaxe_(wikicode).

— (2017b). *Wiki :Catégorisation*. URL : https://fr.wikipedia.org/wiki/Aide:Cat%C3%A9gorisation.

— (2017c). *Wiki :Infobox*. URL : https://fr.wikipedia.org/wiki/Aide:Infobox.

— (2017d). *Wiki :InfoboxMetier*. URL : https://fr.wikipedia.org/wiki/Mod%C3%A8le:Infobox_M%C3%A9tier.

— (2017e). *Wiki :Listes*. URL : https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Ordonnancement_des_listes.

— (2017f). *Wiki :Model*. URL : https://fr.wikipedia.org/wiki/Aide:Mod%C3%A8le.